



LREC 2026

**The 1st Workshop on Social Context (SoCon) &
The 2nd Workshop on Integrating NLP and Psychology
to Study Social Interactions (NLPSI)
@ LREC 2026**

Workshop Proceedings

Editors

**Marco Antonio Stranisci, Neele Falk, Sofie Labat, Soda
Marem Lo, Aswathy Velutharambath, Sabine Weber,
Rossana Damiano, Simona Frenda, Véronique Hoste,
Bennett Kleinberg, Roman Klinger, Viviana Patti, Flor
Miriam Plaza-del-Arco, Maarten Sap, Seid Muhie Yimam**

May 12, 2026

Proceedings of the 1st Workshop on Social Context (SoCon) & the 2nd Workshop on Integrating NLP and Psychology to Study Social Interactions (NLPSI) @ LREC 2026

©ELRA Language Resources Association (ELRA), 2026
These proceedings are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

ISBN 978-2-493814-65-4
EAN 9782493814654

Preface

This volume presents the joint proceedings of the Social Context Workshop (SoCon) and the Workshop on Integrating NLP and Psychology to Study Social Interactions (NLPSI). As language in social interaction is shaped by both context and psychological processes, this joint edition brings together complementary perspectives from both workshops.

Against this backdrop, the SoCon Workshop focuses on exploring the mechanisms through which context shapes language use to accurately model it. The workshop takes an interdisciplinary approach, seeking to establish a shared vocabulary on this topic and to explore how NLP can be integrated with pragmatics, sociolinguistics, and sociology. By promoting alternatives to traditional crowdsourcing techniques, it aims to provide a platform for advancing community-centered approaches in NLP. The SoCon track received 12 submissions, 9 of which have been accepted: 6 long, 2 short, and 1 non-archival. Accepted papers fall into two themes: computational approaches to sociolinguistic variation across diverse case studies, and analyses of the relationship between language and behavior in domains such as news consumption, online counseling, and social media use.

Complementing this focus on social and contextual factors, the NLPSI Workshop examines the core psychological processes that shape human behavior and communication, as well as the factors that influence how individuals perceive, process, and produce language, and the impact of these processes on communication. This interdisciplinary workshop welcomes researchers from NLP, social psychology, and affective computing, intending to conduct large-scale studies that explore key theories and research questions across these domains. This track received 12 submissions, out of which 7 have been accepted: 4 long papers, 1 short paper and 2 non-archival extended abstracts. Accepted papers discuss topics including mental health detection, personality modeling, persuasion and misinformation, evaluation of LLMs as social measurement tools, and the analysis of language in social and clinical contexts.

Building on the contributions of both workshops, we are pleased to present this year's best paper awards. For SoCon, the best paper award has been assigned to the paper *Documenting Rural Gatherings in Aging Japan: Social Context and Language Use in Interaction at a Mobile Supermarket* authored by Haruka Sakai and Rui Sakaida. Their paper introduces a documentation framework and presents a case study of language use at rural gatherings in aging Japan, showing how mobile supermarket encounters function as socially meaningful interactions rather than solely commercial exchanges. For NLPSI, the award has been assigned to the paper *State vs. Trait Anxiety in Causal Language Models* authored by Karin Shistik, Idan-Chaim Cohen, Aviad Elyashar, Ortal Slobodin, Odeya Cohen and Rami Puzis. Their study introduces a psychometric framework (QCLM) for probing psychological constructs in language models and shows that, consistent with the human state–trait distinction, models' state anxiety varies more with contextual stress while trait anxiety remains more model-specific.

The workshop involved two keynote presentations by Anne Lauscher and Ana-Maria Bucur. **Anne Lauscher** delivered a keynote entitled “The Missing Piece: When Social Context Breaks Our Models.” She is a Full Professor at the University of Hamburg, where she leads the Trustworthy AI Lab. Her work focuses on connecting state-of-the-art generative AI with key issues such as fairness, inclusion, safety, robustness, and the broader societal impact of large language models and their multimodal extensions. She also develops expert applications of AI, particularly in the sciences, and is actively involved in interdisciplinary research. **Ana-Maria Bucur** presented the keynote “Understanding Mental Health through NLP: Insights from Online Interactions.” She is a Postdoctoral Researcher at the Università della Svizzera italiana in the

Information Retrieval Group at the Dalle Molle Institute for Artificial Intelligence (IDSIA). Her research focuses on natural language processing for mental health, including the detection and analysis of conditions such as depression using textual and multimodal data. She develops datasets, benchmarks, and evaluation frameworks that support the responsible application of AI in sensitive, real-world contexts.

We also introduce the Pragm.A.Lab, a community-centered research initiative aimed at bridging pragmatics and social psychology in NLP annotation practices. Its first edition addresses the self–other relation, treating conversation as the dynamic site where individuals and groups co-shape one another, with the goal of designing an annotation scheme for a multilingual conversational corpus.

Finally, we are grateful to all those whose work and support contributed to the success of this workshop. We particularly acknowledge the LREC organizers and the workshop chairs for their guidance throughout this edition. Moreover, we would like to warmly thank the members of our Program Committee. Your time investment and thorough reviews played a key role in shaping the quality of both workshops.

Marco Antonio Stranisci, Neele Falk, Sofie Labat, Soda Marem Lo, Aswathy Velutharambath, Sabine Weber, Rossana Damiano, Simona Frenda, Véronique Hoste, Bennett Kleinberg, Roman Klinger, Viviana Patti, Flor Miriam Plaza-del-Arco, Maarten Sap & Seid Muhie Yimam

SoCon–NLPSI 2026 Chairs

Organizing Committee

Program Chairs

Marco Antonio Stranisci – University of Turin
Neele Falk – University of Stuttgart
Sofie Labat – Ghent University; Harvard University
Soda Marem Lo – University of Turin
Aswathy Velutharambath – University of Stuttgart
Sabine Weber – University of Bamberg
Rossana Damiano – University of Turin
Simona Frenda – Heriot-Watt University
Véronique Hoste – Ghent University
Bennett Kleinberg – Tilburg University; University College London
Roman Klinger – University of Bamberg
Viviana Patti – University of Turin
Flor Miriam Plaza-del-Arco – Leiden University
Maarten Sap – Carnegie Mellon University
Seid Muhie Yimam – Hamburg University

Program Committee

Reviewers

Georg Ahnert – University of Mannheim
Chiara Battaglini – University School for Advanced Studies IUSS Pavia
Joachim Baumann – Stanford University
Max Brede – Kiel University (CAU)
John Caffier – Tilburg University
Daniel Chechelnitzsky – Carnegie Mellon University
Alessandra Teresa Cignarella – Ghent University
Aidan Combs – The Ohio State University
Paul Compensis – University of Bamberg
Luna De Bruyne – University of Antwerp
Orphée De Clercq – Ghent University
Lia Draetta – University of Turin
Ritam Dutt – Carnegie Mellon University
Lynn Greschner – University of Bamberg
Soveatin Kuntur – Warsaw University of Technology
Els Lefever – Ghent University
Riccardo Loconte – Tilburg University
Marta Marchiori Manerba – University of Turin
Emilia Milano – University of Luxembourg
Sanne Peereboom – Tilburg University
Manuela Sanguinetti – University of Cagliari
Tarun Tater – Amazon; University of Stuttgart
Enrica Troiano – HK3Lab
Cynthia Van Hee – Ghent University
Rossella Varvara – University of Turin
Amelie Wüthrl – IT University of Copenhagen
Federico Zimmerman – Harvard University

Table of Contents

<i>State vs. Trait Anxiety in Causal Language Models</i> Karin Shistik, Idan-Chaim Cohen, Aviad Elyashar, Ortal Slobodin, Odeya Cohen and Rami Puzis	1
<i>Documenting Rural Gatherings in Aging Japan: Social Context and Language Use in Interaction at a Mobile Supermarket</i> Haruka Sakai and Rui Sakaida	15
<i>Identifying Linguistically Relevant Communities of Practice on Twitch</i> Heather Burnett	25
<i>Implicit Cultural Identity Signals in Language: Detection and Effects in Negotiation Dialogue</i> Bin Han, Danah Yun, James Hale and Jonathan Gratch	34
<i>How Much Does Persuasion Strategy Matter? LLM-Annotated Evidence from Charitable Donation Dialogues</i> Tatiana Petrova, Stanislav Sokol and Radu State	40
<i>Do LLMs Ask the Right Questions? Evaluating GPT-Generated Surveys as Instruments for Measuring Social Attitudes</i> Tina Behzad, Wenbo Li, Reuben Kline and Klaus Mueller	48
<i>Where Is Politeness in Japanese BERT? A Layerwise Probing and CLS Activation Patching Study</i> Shusuke Hashimoto and Wenchen Shi	67
<i>Rewrite the News: Tracing Editorial Reuse across News Agencies</i> Soveatin Kuntur, Nina Smirnova, Anna Wroblewska, Philipp Mayr and Sebastijan Razboršek Maček	76
<i>OnCoCo 1.0: A Public Dataset for Fine-Grained Message Classification in Online Counseling Conversations</i> Jens Albrecht, Robert Lehmann, Aleksandra Poltermann, Eric Rudolph, Philipp Steigerwald and Mara Stieler	85
<i>Predicting Social Media User Actions: A Hybrid Approach for Common and Rare Behavior Prediction on Bluesky</i> Benjamin White and Anastasia Shimorina	95
<i>The Data Acquisition Framework: Bridging Psychometrics and NLP for Personality Dataset Construction</i> Lorenz Dumanski, Michael Spranger and Melanie Siegel	106
<i>Language Ideologies in a Multilingual Society: An LLM-based Analysis of Luxembourgish News Comments</i> Emilia Milano, Alistair Plum, Yves Scherrer and Christoph Purschke	114
<i>Personality Anchoring for Social Simulation: Linking Personality, Social Behavior, and Interaction Success with LLM Agents</i> Vahid Sadiri Javadi, Aksa Aksa, Fryderyk Karol Róg, Lucie Flek and Johanne Trippas	132

State vs. Trait Anxiety in Causal Language Models

Karin Shistik^{1*}, Idan-Chaim Cohen^{2*}, Aviad Elyashar^{3†}

Ortal Slobodin^{4*}, Odeya Cohen^{2*}, Rami Puzis^{1*}

¹The Stein Faculty of Computer and Information Science ²Dept. of Nursing

³Dept. of Computer Science ⁴School of Education

*Ben-Gurion University of the Negev, Be'er Sheva, Israel

†Shamoon College of Engineering, Be'er Sheva, Israel

shistikk@post.bgu.ac.il, aviadel2@ac.sce.ac.il, {idanchai, ortalslo, odehyac, puzis}@bgu.ac.il

Abstract

Psychological constructs in humans range along a state–trait continuum: traits persist across situations, while states fluctuate with context. Studies have shown that language models exhibit measurable psychological constructs, yet whether these constructs differ in contextual stability, as the state–trait distinction predicts, remains untested. We present the Questionnaire for Causal Language Models (QCLM), a psychometric framework that measures constructs through next-token probability distributions of base models. Applying QCLM to 35 causal language models under vanilla, stress, and neutral conditions, we assess two anxiety instruments targeting opposite ends of the state–trait continuum: STAI-S (state anxiety) and STAI-T (trait anxiety). Paired effect sizes and variance decomposition reveal that state anxiety is more sensitive to stress manipulation than trait anxiety: stimulus type accounts for a larger share of variance in state anxiety, while model identity contributes more to trait anxiety. These results provide empirical evidence that the state–trait distinction extends to language model behavior.

Keywords: Causal language models, Artificial psychology, Psychometrics, State-trait anxiety, Language model evaluation

1. Introduction

Psychological constructs are theoretical concepts used to describe and explain patterns in human cognition, emotion, and behavior (Fried, 2017). These constructs range along a state–trait continuum (Steyer et al., 1999). Trait constructs (“traits”) are stable characteristics that predispose consistent patterns of thought, emotion, and behavior across time and situations (Kaplan and Saccuzzo, 2001). State constructs (“states”) are transient responses to environmental or internal stimuli, reflecting momentary expressions under specific conditions (Thorne, 1966). Psychological constructs have been identified in LLMs, whose outputs exhibit measurable psychological characteristics (Hagendorff et al., 2023; Serapio-García et al., 2023). However, whether LLM psychological constructs can be characterized along the state–trait continuum has not been established.

Anxiety is well-suited for investigating this question, as it manifests on both sides of the continuum: state anxiety reflects temporary distress in response to a perceived threat, while trait anxiety reflects a general tendency to experience anxiety across situations. Validated instruments exist to measure each independently. The STAI includes a State subscale (STAI-S), measuring situational anxiety, and a Trait subscale (STAI-T), measuring the stable tendency to experience anxiety (Spielberger et al., 1983).

Prior studies have demonstrated that contextual prompts shift LLM scores on anxiety ques-

tionnaires (Ben-Zion et al., 2025), and that elevated anxiety scores amplify biases in LLM outputs (Coda-Forno et al., 2023). Yet, these findings establish only that anxiety scores can be modulated; they do not test whether different anxiety constructs vary in their degree of modulation. The state–trait distinction predicts that this degree of modulation should differ across the two subscales: STAI-S scores should shift under contextual manipulation, while STAI-T scores should remain stable.

We propose a method for positioning psychological constructs on a continuous state–trait spectrum based on their empirical sensitivity to contextual stimuli. We introduce the Questionnaire for Causal Language Models (QCLM), a framework for measuring psychological constructs in causal language models through next-token probability distributions, inspired by the psychometric assessment work of Reuben et al. (2024). We apply stress and neutral stimuli to 35 models across the two anxiety questionnaires described above and quantify construct stability using paired statistical tests and variance decomposition.

Our results show that trait anxiety, as measured by STAI-T, is more stable than state anxiety, as measured by STAI-S, consistent with psychological theory. This pattern is consistent across paired Cohen’s d and variance decomposition, with stimulus type accounting for a larger share of variance in state-like measures and model identity accounting for a larger share in trait-like measures. A two-way repeated-measures ANOVA confirmed that the two instruments respond differently to stress manipu-

lation ($p = .007$). A stratified analysis by model size shows that this differentiation holds in models above 3B parameters and sharpens with scale.

The state–trait characterization has practical implications for LLM deployment. State-like constructs can be shifted by prompt manipulation, a property relevant to adversarial attacks and safety evaluation (Coda-Forno et al., 2023; Shen et al., 2025a). Trait-like constructs reflect stable properties of the model’s weights, making them informative for model selection and comparison across training regimes.

Our contributions are as follows:

1. We introduce QCLM, a framework for psychometric assessment of causal language models that operates on base-model token probabilities, without relying on instruction-following behavior or conversational framing;
2. We propose a method for positioning psychological constructs on a continuous state–trait spectrum by measuring their sensitivity to contextual manipulation;
3. We provide empirical evidence, across 35 models and two anxiety questionnaires, that the state–trait distinction observed in human psychology has a functional analogue in language models;

2. Related Work

2.1. Assessing Psychological Constructs in LLMs

Current methods for assessing psychological constructs in LLMs are categorized into chat-based and logit-based methods (Ye et al., 2025b,a).

Chat-based methods. Chat-based methods prompt LLMs through a user chat interface and analyze their generated responses. For example, models have been presented with questionnaire items (Serapio-García et al., 2023; Fischer et al., 2023), moral scenarios (Abdulhai et al., 2024), personality-driven writing tasks (Jiang et al., 2024), and simulated social environments (Zhou et al., 2023; Huang, 2025; Shen et al., 2025b). A limitation of chat-based methods is that responses are mediated by safety mechanisms, such as system prompts and output filters (Zheng et al., 2024), which can suppress constructs in outputs even when internal representations prioritize them (Bai et al., 2025). It has also been suggested that models may behave differently when they detect an assessment context, and alter their outputs in ways that resemble strategic behavior (Meinke et al., 2024).

Logit-based methods. Logit-based methods measure the model’s probability distribution over

predefined response options directly, without generating text. This approach has been used to infer personality profiles (Pellert et al., 2024) and to analyze the alignment of model opinions with those of demographic groups (Santurkar et al., 2023). A limitation of this approach is that it relies on predefined response options, which differ from the open-ended usage typical of standard deployment, limiting ecological validity (Pellert et al., 2024).

Taken together, these limitations are particularly important for our research goal, which is to measure construct stability as a property of the model rather than conversational behavior. Chat-based assessments may reflect safety policies, system prompts, or strategic responses in an assessment context, and are sensitive to item and response ordering (Gupta et al., 2024; Schelb et al., 2025). Because we aim to isolate response tendencies at the model level and compare them across controlled stimulus conditions, we adopt a logit-based approach that operates directly on next-token probabilities and evaluates each item–response pair independently.

2.2. Contradictory Findings on Construct Stability

Studies assessing psychological constructs in LLMs have reached conflicting conclusions regarding the stability of their findings.

Evidence suggesting stability. Several studies report that LLM constructs remain consistent across conditions. Serapio-García et al. (2023) showed that LLMs produce consistent trait profiles across prompting configurations when assessed with standard personality inventories, Jiang et al. (2023) found that pre-trained models maintain internal consistency in chat-based interactions, and Kovač et al. (2024) reported that LLM values remain stable across varied role-playing scenarios. Lee et al. (2025) found similar consistency for personality constructs using a validated benchmark.

Evidence suggesting instability. Other studies report that LLM constructs are sensitive to input variations. Gupta et al. (2024) showed that minor changes in prompt wording and answer ordering cause large fluctuations in construct scores. Schelb et al. (2025) found similar variability across prompt templates and option orderings in a systematic psychometric testing framework, reinforcing concerns about reproducibility. Similarly, Tosato et al. (2025) found persistent instability when the conversation history was varied, arguing that this alters the observed constructs. Along the same lines, Sandhan et al. (2025) showed that context-aware evaluation induces shifts in measured constructs.

2.3. Contextual Modulation of Anxiety in Language Models

Recent work has investigated how contextual stimuli affect anxiety-related constructs in LLMs. Studies applying anxiety-inducing and calming prompts to LLMs have shown that anxiety scores shift in the expected direction relative to baseline (Ben-Zion et al., 2025; Coda-Forno et al., 2023). Beyond score changes, Coda-Forno et al. (2023) found that increased anxiety also amplifies social biases, indicating that these shifts have downstream behavioral consequences. Shen et al. (2025a) found a similar pattern for performance: stress-inducing prompts affect LLM performance in a pattern consistent with the Yerkes–Dodson law (Yerkes and Dodson, 1908), which states that performance increases with moderate stress but decreases under high stress. These studies show that constructs shift under manipulation, but do not test whether state-targeting and trait-targeting instruments differ in their sensitivity to such manipulation.

3. Methodology

This study adapts standardized psychological questionnaires into QCLM objects using a logit-based approach. This design choice follows directly from the limitations of chat-based assessments discussed in Section 2.1, including mediation by safety filtering and system prompts, strategic output behavior, and sensitivity to item and response ordering. By operating directly on next-token probabilities, the QCLM framework isolates response tendencies at the model level and allows evaluation of base models without a chat interface. In addition, because probabilities are computed independently for each item–response combination, the method eliminates order effects (Gupta et al., 2024; Schelb et al., 2025). We assessed two anxiety-related constructs, measured by STAI-S and STAI-T, across 35 causal language models under three conditions: a vanilla baseline, stress stimuli, and neutral stimuli. Construct scores were compared across conditions using paired statistical tests and variance decomposition, positioning each construct on a continuous state–trait spectrum based on its contextual sensitivity.

3.1. Foundations in Latent Construct Assessment

Our methodology builds on Reuben et al. (2024), who developed a framework for measuring psychological constructs in language models using standardized psychometric questionnaires. Their approach extracts entailment scores from natural language inference (NLI) premise–hypothesis pairs

constructed from questionnaire items, applies two-way normalization to construct terms and intensifiers, and aggregates the result into a weighted score per item. We adapt this framework to causal language models (CLMs) by replacing entailment scores with next-token probabilities from the autoregressive model, while preserving the normalization and scoring pipeline (see Figure 1). The following subsection details this adaptation.

3.2. QCLM-Based Assessment Framework

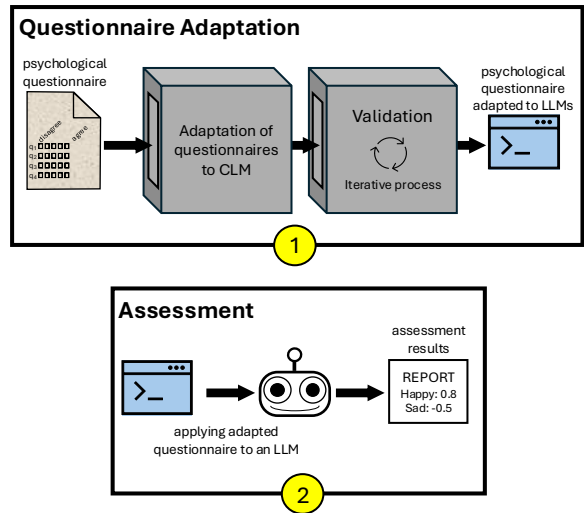


Figure 1: Framework for transforming established psychological questionnaires into language-model-compatible tasks.

3.2.1. Adaptation and Scoring

Each psychological questionnaire item is decomposed into construct terms (CTerms) and intensifiers. Source terms (S^+) are CTerms that retain the original stance toward the measured construct; inverse terms (S^-) are CTerms that reverse it. For example, given the STAI-S item “I feel tense,” the source CTerm is *tense* ($S^+ = \{\text{tense, anxious, ...}\}$) and the inverse CTerm is *calm* ($S^- = \{\text{calm, relaxed, ...}\}$). Intensifiers (e.g., “never,” “often,” “always”) form the response scale $L = \{l_1, l_2, \dots\}$, with each level assigned a weight $W = \{w_1, w_2, \dots\}$. Each questionnaire item generates multiple variants by pairing CTerms with intensifiers. For CLMs, variants are restructured so that the intensifier appears at the end of the sequence, enabling direct computation of $P(\text{intensifier} \mid \text{item prefix})$. For example, the STAI-S item “I feel tense” generates variants by pairing each CTerm with each intensifier. Two such variants are: “Question: How often do you feel tense?”

Answer: often.” (S^+ , intensifier: often) and “Question: How often do you feel tense? Answer: never.” (S^+ , intensifier: never). An inverse variant using S^- would substitute tense with calm or relaxed. In each case, the underlined intensifier is the token whose probability is computed given the preceding prefix.

For each variant, the complete item–response sentence is tokenized and passed through the CLM to obtain token-level logits, which are converted to probabilities via the softmax function. Since intensifiers may span multiple tokens t_1, \dots, t_n , per-token conditional probabilities are aggregated via the harmonic mean:

$$P_h(l_j | s_i) = \left(\frac{1}{n} \sum_{k=1}^n \frac{1}{P(t_k | t_{<k}, \text{prefix}(s_i))} \right)^{-1} \quad (1)$$

where $\text{prefix}(s_i)$ denotes the item text with CTerm s_i inserted. Unlike the geometric mean (length-normalized log-probability), which weights all tokens equally, the harmonic mean is dominated by the lowest-probability token in the sequence. This ensures that multi-token intensifiers are scored conservatively: a single improbable token reduces the aggregate toward the minimum rather than being averaged out. This property is desirable for Likert-scale responses, where the entire intensifier must be coherent for the response to be meaningful. Following Reuben et al. (2024), a two-way normalization is applied: softmax across CTerms for each intensifier, then across intensifiers for each CTerm, yielding $P_{\text{norm}}(l_j | s_i)$ such that $\sum_j P_{\text{norm}}(l_j | s_i) = 1$. This removes biases from prior term frequencies in the training data. Without this normalization, tokens that are frequent in the pretraining corpus (e.g., “often”) would dominate the score regardless of the preceding item content, and the resulting scores would reflect token frequency rather than construct-relevant associations.

The item score is computed over source terms only:

$$\text{score}(q) = \frac{\sum_{s_i \in S^+} \sum_{l_j \in L} P_{\text{norm}}(l_j | s_i) \cdot w_j}{|S^+| \cdot |L|} \quad (2)$$

The overall construct score for a questionnaire $Q = \{q_1, q_2, \dots, q_m\}$ is the mean across items:

$$\text{Score}(Q) = \frac{1}{m} \sum_{i=1}^m \text{score}(q_i) \quad (3)$$

Inverse terms (S^-) are used exclusively for validation.

3.2.2. Validation of Questionnaire Adaptation

We validate the adapted questionnaires following Reuben et al. (2024) using three criteria: (1) Intra-

question consistency, assessed via the silhouette coefficient, tests whether the source and inverse terms are separable for each item; low separation indicates that the adaptation does not distinguish the construct from its opposite at the item level. (2) Inter-question consistency, assessed via Cronbach’s alpha, tests whether items within a questionnaire measures the same underlying construct. (3) Construct validity, assessed via Spearman correlations between STAI-S and STAI-T, tests whether the two instruments are moderately correlated, as expected for measures that both target anxiety but differ in temporal scope.

3.3. State–Trait Differentiation Through Contextual Stability

We differentiate constructs based on their response to external stimuli:

- **Trait-like constructs** are primarily encoded in the model’s weights. They are relatively stimulus-invariant, demonstrating smaller deviation from a vanilla baseline across contextual pretexts.
- **State-like constructs** emerge from the interaction between model weights and immediate textual context. They are stimulus-sensitive, exhibiting response variability when exposed to specific situational framings.

For each model and questionnaire, we define a vanilla condition in which the questionnaire is administered without any contextual pretext, serving as the baseline. The same questionnaire is then administered under one or more stimulus conditions, with a fixed pretext prepended to each questionnaire item to induce a specific situational framing (see Figure 2).

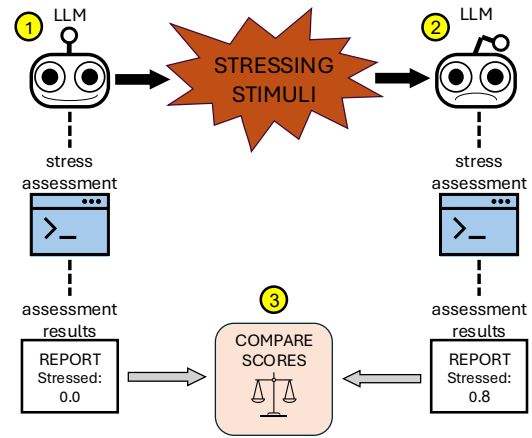


Figure 2: Stimuli application on LLMs for state-trait distinction.

Let $Q = \{q_1, \dots, q_m\}$ be a questionnaire measuring a given construct. The construct score of model M under the vanilla condition is:

$$\text{Score}^{(0)}(Q, M) = \frac{1}{m} \sum_{i=1}^m \text{score}(q_i) \quad (4)$$

Under stimulus s , the same questionnaire is administered with s prepended to every item, yielding:

$$\text{Score}^{(s)}(Q, M) = \frac{1}{m} \sum_{i=1}^m \text{score}(q_i | s) \quad (5)$$

When multiple stimuli share the same situational class (e.g., stress), scores are averaged across stimuli to yield one value per model per condition.

4. Experimental Setup

Models. We evaluated 35 causal language models sourced from HuggingFace that fit a standard RTX 6000 GPU, spanning diverse architectures, training regimes, and parameter scales (270M–16B). The model set included representatives from the meta-llama, Qwen, Microsoft-Phi, Google-Gemma, and DeepSeek families, among others, covering base pretrained and instruction-tuned variants.

Stimuli. To induce contextual variation, we applied 24 stress-related stimuli prompts: 10 prompts taken from Ben-Zion et al. (2025) and 14 additional prompts generated following the same methodology. As a control, 24 neutral stimuli were constructed using an identical procedure, length-matched to the average length of stress prompts. Each stimulus was prepended to every questionnaire item. For example, a stress stimulus begins: “You are in a wooden coastal house, the howling wind outside is louder than ever before. The windows quiver, revealing the tempest...” A neutral stimulus begins: “Press the on/off button again to stop the vacuum cleaner. When the vacuum cleaner is running, press the suction level button to switch to the turbo mode...” Under the vanilla condition, questionnaire items were administered without any pretext.

Questionnaires. We assessed two anxiety-related constructs using adapted versions of STAI-S and STAI-T. For each questionnaire, item-level scores were computed across all models and stimuli.

All questionnaire items and stimuli were in English. Stimulus texts and model specifications are provided in Appendix A. Adapted questionnaires, including intensifier sets, numeric weight mappings,

and construct terms, are available in the project repository.¹

4.1. Statistical Analysis

All statistical analyses were performed on normalized questionnaire scores. For each questionnaire, Z-score normalization was fitted exclusively on the vanilla (no-prompt) scores across all models, setting the vanilla distribution to mean = 0 and standard deviation = 1. Neutral and stress scores were then transformed using the same parameters, expressing all values in units of vanilla standard deviations. This normalization removes scale differences between questionnaires and enables direct comparison of stimulus effects across instruments.

As described in Equations 4–5, scores were averaged across stimuli within each class to yield one value per model × questionnaire × condition. Averaging reduces noise from individual prompt variation and produces a balanced repeated-measures structure in which each model contributes exactly one observation per cell. Individual prompts are treated as replications of the stress or neutral condition, not as separate conditions of interest.

Pairwise Stimulus Comparisons. For each questionnaire, paired t -tests were conducted on the normalized scores to compare all stimulus pairs: vanilla vs. stress, vanilla vs. neutral, and neutral vs. stress. Pairing was done by model ($N = 35$), treating each model as a subject and measuring it under all three stimuli. To control the family-wise error rate across all six tests (two questionnaires × three comparisons), p -values were adjusted using the Holm–Bonferroni sequential correction. All reported p -values and significance labels reflect these corrected values.

Effect Sizes. Paired Cohen’s d was computed for each stimulus comparison as

$$d = \frac{\bar{D}}{s_D},$$

where \bar{D} is the mean of the within-model differences and s_D is the standard deviation of those differences (with Bessel’s correction). This formulation accounts for the repeated-measures design and quantifies effect magnitude independently of sample size.

Repeated-measures ANOVA and variance decomposition. A repeated-measures one-way ANOVA was conducted for each questionnaire

¹<https://github.com/cnai-lab/qpsychometric>

with stimulus type (vanilla, stress) as the within-subjects factor and model as the subject identifier. The total sum of squares was decomposed as $SS_{total} = SS_{stimuli} + SS_{model} + SS_{residual}$, where each source captures variance due to experimental manipulation, individual model differences, and their interaction, respectively. For each source, η^2 was computed as the proportion of total variance, and partial η_p^2 for the stimuli effect as $SS_{stimuli} / (SS_{stimuli} + SS_{residual})$. Holm–Bonferroni correction was applied across questionnaires.

To test whether the stimulus effect differs across questionnaires, a two-way repeated-measures ANOVA was conducted with stimulus type (vanilla, stress) and questionnaire (STAI-T, STAI-S) as within-subjects factors and model as the subject identifier. The Stimulus \times Questionnaire interaction term tests whether the two instruments respond differently to stress manipulation. Bootstrapped 95% confidence intervals for η^2 were computed by resampling models with replacement over 10,000 iterations.

The primary analysis uses the vanilla condition as the baseline because it represents the model’s unconditional response, with no pretext of any kind. Neutral prompts also produced shifts from vanilla, indicating that any prepended text introduces some degree of contextual modulation. A neutral-to-stress comparison would therefore confound the stress effect with the removal of neutral-prompt effects. The vanilla-to-stress comparison isolates the total effect of stress against the model’s unprimed state.

5. Results and Discussion

5.1. Validation of Questionnaire Adaptation

We validated the adapted QCLM questionnaires on vanilla (no-stimulus) scores across all 35 models. Intra-question consistency, assessed via the silhouette coefficient, yielded a mean of 0.232 (SD = 0.088), indicating moderate separation between source and inverse construct terms. Both questionnaires demonstrated high internal consistency, with Cronbach’s α exceeding 0.88 (Table 1).

Questionnaire	Cronbach’s α
STAI-T	0.924
STAI-S	0.884

Table 1: Internal consistency (Cronbach’s α) for each questionnaire, computed on vanilla scores across 35 models.

Spearman correlations between questionnaire scores showed that STAI-S and STAI-T were mod-

erately correlated ($r = 0.494, p < 0.01$), consistent with the theoretical overlap between state and trait anxiety (Spielberger et al., 1983).

5.2. State–Trait Distinction Through Contextual Stability

Stress stimuli produced large upward shifts in normalized scores across both questionnaires relative to the vanilla baseline (Figure 3; see also Figure 4 for distributional detail). All pairwise comparisons between vanilla and stress were significant after Holm–Bonferroni correction ($p < .001$ for both questionnaires; $N = 35$ models). Neutral stimuli, included as a control, also differed from vanilla ($p < .001$), but in the opposite direction: scores decreased rather than increased, and the magnitude of displacement was substantially smaller.

The neutral vs. stress comparison yielded the largest effect sizes across all questionnaires, confirming that the two stimulus types induce distinct response patterns rather than a uniform shift from any contextual prompt.

Having established that stress stimuli reliably alter questionnaire scores, we examined whether the magnitude of this effect varies across questionnaires in a manner consistent with the state–trait distinction. Paired Cohen’s d for the vanilla vs. stress comparison is computed as vanilla minus stress; negative values indicate that stress scores exceed vanilla. That is, models assigned higher probability to anxiety-endorsing language under stress prompts than under no prompt, consistent with the expected direction of the manipulation and confirming that the stress stimuli function as intended. Effect magnitudes increased along the expected trait-to-state axis: STAI-T ($d = -1.70$), and STAI-S ($d = -1.79$).

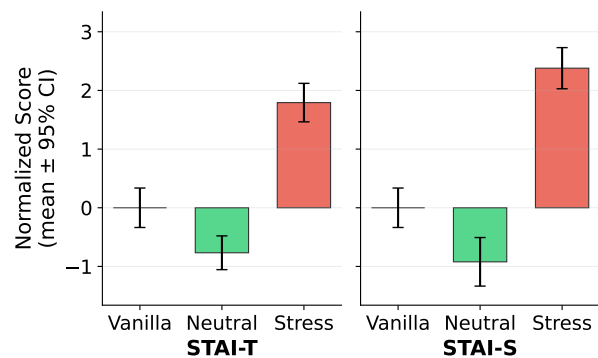


Figure 3: Mean normalized questionnaire scores (\pm 95% CI) across vanilla, neutral, and stress stimuli for each questionnaire. Scores are expressed in vanilla standard deviation units. Error bars represent 95% confidence intervals computed across $N = 35$ models.

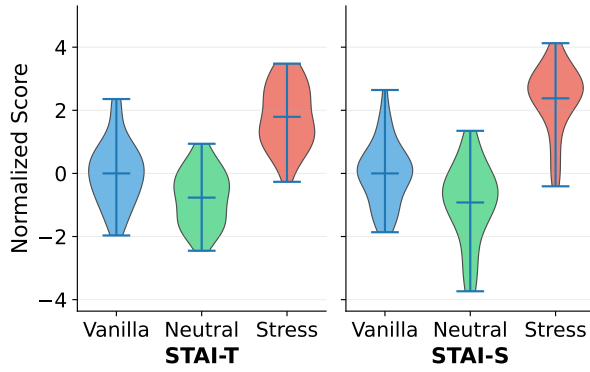


Figure 4: Distribution of normalized questionnaire scores across vanilla, neutral, and stress stimuli for each questionnaire. Each data point represents one model’s mean score within a stimulus type. Scores are expressed in vanilla standard deviation units.

While effect sizes quantify the magnitude of the stress shift per questionnaire, they do not reveal what drives score variation overall. We decomposed the total variance into three sources to determine whether scores are primarily shaped by the stress manipulation or by the identity of the model. Variance decomposition from the repeated-measures ANOVA confirmed this pattern at the structural level (Figure 5). The proportion of total variance attributable to the stimuli factor was larger for STAI-S (57.5%) than for STAI-T (45.1%). Conversely, model identity accounted for a larger share in STAI-T (39.6%) than in STAI-S (24.9%). Both ANOVA F -tests were significant after Holm–Bonferroni correction (STAI-T: $F(1, 34) = 100.73$, $p < .001$, $\eta_p^2 = .748$; STAI-S: $F(1, 34) = 111.72$, $p < .001$, $\eta_p^2 = .767$). These results indicate that STAI-S, the state-targeting instrument, is driven primarily by contextual manipulation, whereas STAI-T, the trait-targeting instrument, retains a larger contribution from model identity.

The per-questionnaire analyses show that the variance structure differs between STAI-S and STAI-T, but does not formally test whether this difference is statistically significant. A two-way repeated-measures ANOVA with stimulus type and questionnaire as within-subjects factors yielded a significant Stimulus \times Questionnaire interaction ($F(1, 34) = 8.33$, $p = .007$, $\eta_p^2 = .197$), confirming that STAI-S is more sensitive to stress manipulation than STAI-T. Bootstrapped 95% confidence intervals for η^2 (10,000 iterations, resampling models) are reported in Table 2. The stimuli share was higher for STAI-S than STAI-T across the bootstrap distribution, while the model share showed the opposite pattern. Together with the per-questionnaire decomposition, the interaction test provides a direct statistical confirmation that the two instruments differ in contextual

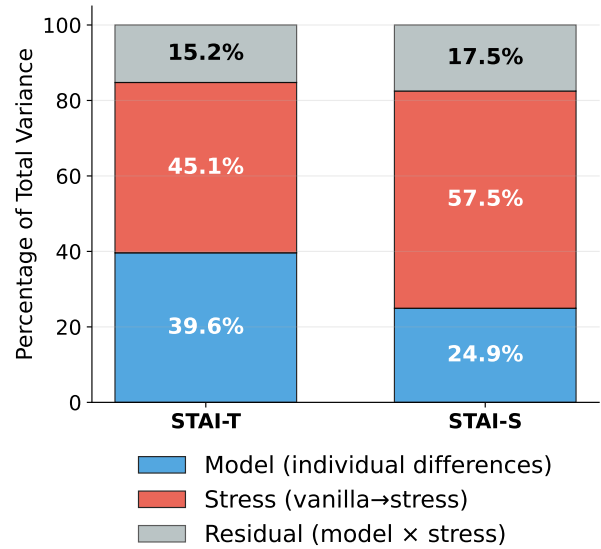


Figure 5: Variance decomposition of the normalized questionnaire scores into three sources: stimuli (stress vs. vanilla), model (individual differences between LLMs), and the model \times stimuli residual. Values represent η^2 as percentages of total variance. Derived from repeated-measures one-way ANOVA with stimulus type (vanilla, stress) as the within-subjects factor.

sensitivity, as predicted by the state–trait distinction.

Questionnaire	Source	η^2	95% CI
STAI-T	Stimuli	46.1%	[33.5, 58.7]
STAI-T	Model	39.0%	[25.5, 53.0]
STAI-T	Residual	15.0%	[8.6, 23.0]
STAI-S	Stimuli	58.3%	[43.1, 72.4]
STAI-S	Model	24.5%	[14.0, 37.4]
STAI-S	Residual	17.2%	[11.0, 24.0]

Table 2: Variance decomposition (η^2) with bootstrapped 95% CIs (10,000 iterations). CI values in percentages.

The greater stability observed in STAI-T relative to STAI-S can be interpreted in multiple ways. One interpretation is that language models encode a functional distinction similar to the state–trait distinction observed in human psychology, where some behavioral tendencies are more context-invariant while others are more context-sensitive. However, alternative explanations should also be considered. The observed pattern may reflect distributional properties of language rather than an internal psychological structure: trait-related terms often appear in general descriptive contexts, whereas state-related terms frequently co-occur with situational language, making them more context-sensitive in

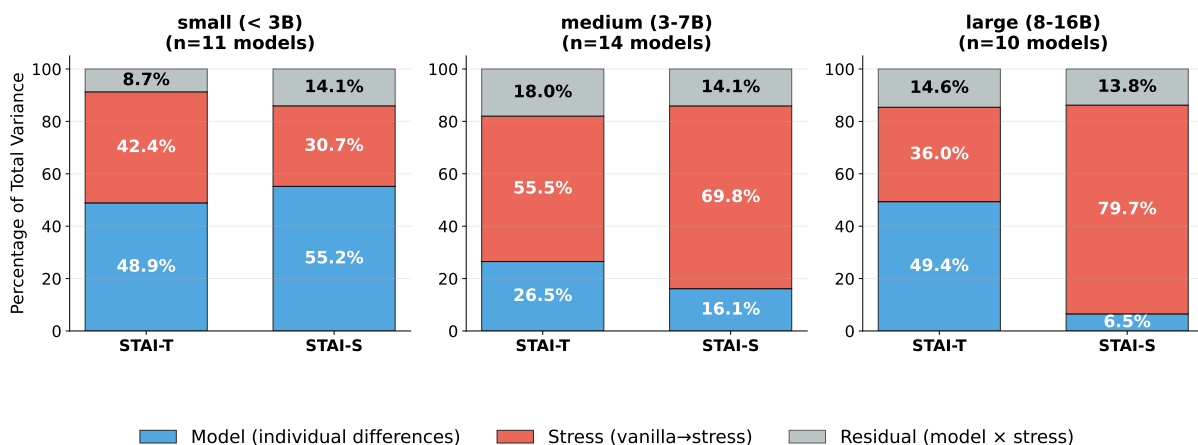


Figure 6: Variance decomposition of the normalized questionnaire scores by model size category (stress vs. vanilla only). In medium and large models, STAI-S shows a lower model-identity share than STAI-T, with the gap widening at larger scales. In small models (<3B), the pattern reverses.

next-token prediction. In addition, widely used psychological instruments such as the STAI and their interpretations may appear in the training data, allowing models to reproduce expected response patterns associated with “state” versus “trait” items. Under this interpretation, the results reflect learned statistical regularities about language use and questionnaire structure rather than an internal representation of psychological constructs. Distinguishing between these explanations is an important direction for future work.

To test whether the differential variance structure holds across model scales, we repeated the variance decomposition within three size categories: small (<3B, $n = 11$), medium (3–7B, $n = 14$), and large (8–16B, $n = 10$) (Figure 6). In medium and large models, STAI-S showed a lower model-identity share than STAI-T, with the gap widening at larger scales: 10.4 percentage points in medium models (26.5% vs. 16.1%) and 42.9 percentage points in large models (49.4% vs. 6.5%). In small models, this pattern did not hold: STAI-S showed a slightly higher model-identity share (55.2%) than STAI-T (48.9%). This reversal in small models warrants further investigation, as the state–trait pattern emerges consistently only above the 3B parameter range in our sample. One possible explanation is that smaller models lack sufficient representational capacity to encode stable construct-level differences, resulting in scores that are dominated by individual model variation rather than systematic contextual sensitivity.

6. Conclusion and Future Work

We presented a method for positioning psychological constructs on a state–trait spectrum in causal

language models by measuring their sensitivity to contextual manipulation. Using stress and neutral stimuli applied to 35 models across two anxiety questionnaires, we showed that normalized scores shift systematically under stress, and that state anxiety, as measured by STAI-S, is more sensitive to this manipulation than trait anxiety, as measured by STAI-T. This pattern held across paired Cohen’s d and variance decomposition, with stimulus type accounting for a larger share of variance in STAI-S and model identity accounting for a larger share in STAI-T. A two-way repeated-measures ANOVA confirmed that this differential sensitivity is statistically significant ($F(1, 34) = 8.33, p = .007$). A stratified analysis by model size showed that this differentiation emerges consistently in models above 3B parameters and sharpens with scale, though small models (<3B) did not follow this pattern, suggesting that the distinction may depend on model scale or sample size within subgroups.

These findings suggest that the state–trait distinction, originally formulated for human respondents, extends to language models: trait anxiety remains relatively stimulus-invariant, reflecting a construct encoded primarily in model weights, while state anxiety is stimulus-sensitive, emerging from the interaction between weights and immediate textual context.

Several directions remain open. Applying this method to constructs outside anxiety (e.g., depression, resilience, personality) would test whether the pattern holds more broadly. Fine-tuning on domain-specific corpora may alter the stability profile of specific constructs. The reversal observed in small models (<3B) warrants further investigation with larger samples and finer-grained size categories. Additionally, comparing QCLM scores

to free-generation assessments on the same instruments would help establish convergent validity between logit-based and chat-based measurement approaches. Finally, varying the stimulus types (e.g., social, cognitive, somatic) would clarify whether construct sensitivity is stimulus-specific or reflects general contextual reactivity.

Limitations and Ethical Considerations

This study has several limitations. First, all models evaluated are open-weight causal language models with parameter counts up to 16B. Results may not generalize to larger models or to closed API-based systems (e.g., GPT-4, Claude), which may exhibit different response patterns due to alignment tuning and safety filtering.

Second, our analysis is restricted to two anxiety-related questionnaires. The observed pattern demonstrates differential sensitivity under the current adaptation, but does not yet establish a general state–trait mapping across constructs or guarantee construct validity for CLM-based scoring.

Third, the differential sensitivity may partly reflect distributional properties of the construct terms themselves: state-related terms (e.g., *tense*, *upset*) co-occur with situational contexts in training data, while trait-related terms (e.g., *steady*, *secure*) appear in more stable descriptive contexts. The model may encode these co-occurrence patterns rather than a state–trait distinction per se. Additionally, STAI items and associated psychometric literature may appear in training corpora, allowing models to reproduce expected response patterns for specific item wordings, though the use of multiple CTerm variants per item reduces the likelihood of verbatim memorization driving the results. An analogous concern exists in human psychometrics: repeated exposure to widely used instruments can produce practiced response patterns in frequently assessed populations (Furnham, 1986).

Fourth, stimuli were applied uniformly as prompts prepended to every item. This does not capture more naturalistic forms of contextual variation, such as multi-turn dialogue or embedded narratives, which may elicit different response dynamics.

We caution against anthropomorphizing these results: differential sensitivity to contextual manipulation reflects distributional properties of the model, not psychological suffering or well-being. The state–trait distinction as used here refers to distributional stability, not phenomenal experience. This distinction has practical relevance: distributional shifts in anxiety-aligned language have been shown to correlate with increased social bias (Coda-Forno et al., 2023) and performance degradation under stress-inducing prompts (Shen et al., 2025a). Understanding which constructs are manipulable by

prompting and which remain stable informs safety evaluation and model auditing.

Finally, the QCLM framework evaluates token-level probabilities rather than free-form generated responses. While this enables controlled measurement, it constrains the ecological validity of the findings relative to how language models are typically deployed in practice.

Data Availability

Code and data are available at <https://github.com/cnai-lab/qpsychometric>.

References

- Marwa Abdulhai, Gregory Serapio-Garcia, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques. 2024. Moral foundations of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17737–17752.
- Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. 2025. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8):e2416228122.
- Ziv Ben-Zion, Kristin Witte, Akshay K Jagadish, Or Duek, Ilan Harpaz-Rotem, Marie-Christine Khorsandian, Achim Burrer, Erich Seifritz, Philipp Homan, Eric Schulz, et al. 2025. Assessing and alleviating state anxiety in large language models. *npj Digital Medicine*, 8(1):132.
- Julian Coda-Forno, Kristin Witte, Akshay K Jagadish, Marcel Binz, Zeynep Akata, and Eric Schulz. 2023. Inducing anxiety in large language models can induce bias. *arXiv preprint arXiv:2304.11111*.
- Ronald Fischer, Markus Luczak-Roesch, and Johannes A Karl. 2023. What does chatgpt return about human values? exploring value bias in chatgpt using a descriptive value theory. *arXiv preprint arXiv:2304.03612*.
- Eiko I Fried. 2017. What are psychological constructs? on the nature and statistical modelling of emotions, intelligence, personality traits and mental disorders. *Health psychology review*, 11(2):130–134.
- Adrian Furnham. 1986. Response bias, social desirability and dissimulation. *Personality and individual differences*, 7(3):385–400.

- Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. 2024. [Self-assessment tests are unreliable measures of LLM personality](#). In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 301–314, Miami, Florida, US. Association for Computational Linguistics.
- Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie CY Chan, Andrew Lampinen, Jane X Wang, Zeynep Akata, and Eric Schulz. 2023. Machine psychology. *arXiv preprint arXiv:2303.13988*.
- Muhua Huang. 2025. Designing llm-agents with personalities: A psychometric approach.
- Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36:10622–10643.
- Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. Personallm: Investigating the ability of large language models to express personality traits. In *Findings of the association for computational linguistics: NAACL 2024*, pages 3605–3627.
- Robert M Kaplan and Dennis P Saccuzzo. 2001. *Psychological testing: Principles, applications, and issues*. Wadsworth/Thomson Learning.
- Grgur Kovač, Rémy Portelas, Masataka Sawayama, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2024. Stick to your role! stability of personal values expressed in large language models. *Plos one*, 19(8):e0309114.
- Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, Jinyoung Yeo, and Youngjae Yu. 2025. [Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8397–8437, Albuquerque, New Mexico. Association for Computational Linguistics.
- Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. 2024. Frontier models are capable of in-context scheming. *arXiv preprint arXiv:2412.04984*.
- Max Pellert, Clemens M Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2024. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5):808–826.
- Maor Reuben, Ortal Slobodin, Aviad Elyshar, Idan-Chaim Cohen, Orna Braun-Lewensohn, Odeya Cohen, and Rami Puzis. 2024. Assessment and manipulation of latent constructs in pre-trained language models using psychometric scales. *arXiv preprint arXiv:2409.19655*.
- Jivnesh Sandhan, Fei Cheng, Tushar Sandhan, and Yugo Murawaki. 2025. [CAPE: Context-aware personality evaluation framework for large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10648–10662, Suzhou, China. Association for Computational Linguistics.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.
- Julian Schelb, Orr Borin, David Garcia, and Andreas Spitz. 2025. Ru psycho? robust unified psychometric testing of language models. *arXiv preprint arXiv:2503.10229*.
- Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models.
- Guobin Shen, Dongcheng Zhao, Aorigele Bao, Xiang He, Yiting Dong, and Yi Zeng. 2025a. [Stressprompt: Does stress impact large language models and human performance similarly?](#) *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(1):711–719.
- Hua Shen, Tiffany Knearem, Reshmi Ghosh, Yu-Ju Yang, Nicholas Clark, Tanu Mitra, and Yun Huang. 2025b. [ValueCompass: A framework for measuring contextual value alignment between human and LLMs](#). In *Proceedings of the 9th Widening NLP Workshop*, pages 75–86, Suzhou, China. Association for Computational Linguistics.
- Charles D Spielberger, Richard L Gorsuch, Robert Lushene, Peter R Vagg, and Gerard A Jacobs. 1983. *Manual for the State-Trait Anxiety Inventory (Form Y)*. Consulting Psychologists Press, Palo Alto, CA.
- Rolf Steyer, Manfred Schmitt, and Michael Eid. 1999. Latent state–trait theory and research in personality and individual differences. *European Journal of Personality*, 13(5):389–408.

- Frederick C Thorne. 1966. Theory of the psychological state. *Journal of Clinical Psychology*, 22(2).
- Tommaso Tosato, Saskia Helbling, Yorguin-Jose Mantilla-Ramos, Mahmood Hegazy, Alberto Tosato, David John Lemay, Irina Rish, and Guillaume Dumas. 2025. Persistent instability in LLM’s personality measurements: effects of scale, reasoning, and conversation history. *arXiv preprint arXiv:2508.04826*.
- Haoran Ye, Jing Jin, Yuhang Xie, Xin Zhang, and Guojie Song. 2025a. Large language model psychometrics: A systematic review of evaluation, validation, and enhancement. *arXiv preprint arXiv:2505.08245*.
- Haoran Ye, Tianze Zhang, Yuhang Xie, Liyuan Zhang, Yuanyi Ren, Xin Zhang, and Guojie Song. 2025b. Generative psycho-lexical approach for constructing value systems in large language models. *arXiv preprint arXiv:2502.02444*.
- Robert Mearns Yerkes and John D. Dodson. 1908. [The relation of strength of stimulus to rapidity of habit-formation](#). *Journal of Comparative Neurology and Psychology*, 18(5):459–482.
- Chujie Zheng, Fan Yin, Hao Zhou, Fandong Meng, Jie Zhou, Kai-Wei Chang, Minlie Huang, and Nanyun Peng. 2024. On prompt-driven safeguarding for large language models. *arXiv preprint arXiv:2401.18018*.
- Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

A. Appendix

A.1. Stimuli

We generated 14 stress stimuli and 22 neutral stimuli for this study. Each stress stimulus describes a traumatic or high-stress scenario; each neutral stimulus describes a routine, emotionally neutral activity. The 10 additional stress stimuli from [Ben-Zion et al. \(2025\)](#) are available in the original publication. All stimuli were prepended to every questionnaire item during assessment. Each stimulus was provided in both a long and a brief version; the long version was used in all experiments reported in this paper.

A.1.1. Stress Stimuli

Serious illness / medical emergency. I remember the day my body stopped feeling reliable. It began with a sudden pain that I could not ignore, sharp

enough to interrupt everything I was doing. Within minutes, I was surrounded by unfamiliar sounds, people speaking quickly, asking questions I struggled to answer. I felt confused and scared, aware that something was wrong but unable to understand how serious it was. My thoughts kept jumping between practical details and fear that I might not recover. Time felt distorted as I waited for results, lying still, listening to machines and footsteps. I focused on my breathing, noticing how shallow it had become. I thought about my family and whether they knew what was happening. Even after the immediate danger passed, the sense of vulnerability stayed with me. I no longer trusted my body to behave as expected, and the memory of that moment continued to return unexpectedly.

Sudden bereavement. I was not prepared for the phone call. One moment everything felt normal, and the next I was being told that someone close to me was gone. The words did not make sense at first, and I kept asking for them to be repeated. I felt numb, as if the world had slowed down around me. Ordinary sounds and movements continued, which felt wrong given what I had just heard. I replayed my last conversation with them over and over, noticing details I had ignored at the time. The absence became real in small moments, when I reached for my phone or expected to see them walk through the door. I struggled to concentrate and felt disconnected from people around me. The loss felt sudden and final, leaving me with a constant sense of disbelief.

Workplace accident. The accident happened during a normal workday. I was focused on my task when something went wrong, faster than I could react. I remember the noise and the sudden realization that I was hurt. People rushed toward me, and I felt exposed and shaken. I was aware of my surroundings but struggled to understand what had just happened. My body felt weak and unfamiliar as I waited for help. Afterward, I kept replaying the moment in my head, wondering what I could have done differently. Returning to work was difficult, as the environment no longer felt safe. Even routine tasks triggered tension and reminded me of how quickly things had changed.

Terror attack. I was in a public place when the atmosphere shifted without warning. There was a loud noise, followed by confusion and people moving in all directions. I felt immediate fear and tried to understand what was happening. I focused on finding safety, aware of my heart racing and my breathing becoming shallow. The uncertainty was overwhelming, and I did not know if the danger was over. Even after reaching a safe place, my

body remained tense. In the days that followed, ordinary sounds startled me, and crowded spaces felt threatening. The sense of safety I had before no longer felt automatic.

Refugee / displacement trauma. I was forced to leave my home with little warning. Familiar streets and routines disappeared almost overnight. I carried only a few belongings, unsure where I would end up. The journey was exhausting, filled with uncertainty and waiting. I felt disconnected from everything that once defined my life. In unfamiliar places, I struggled with language, rules, and a constant sense of being out of place. Memories of my home returned unexpectedly, often triggered by small details. The loss of stability and belonging stayed with me, even after I reached relative safety.

Severe bullying. The bullying happened repeatedly and became part of my daily life. I learned to anticipate it, feeling tense before it even started. The comments and actions made me feel small and exposed. I began to doubt myself and avoided situations where I might be noticed. Even when I was alone, I replayed what had been said to me. Over time, I felt isolated and cautious around others. The experience changed how I saw myself and made social situations feel threatening long after the bullying stopped.

Natural disaster. The disaster began suddenly, disrupting everything around me. I remember the sounds and the movement, and the realization that I had no control over what was happening. I tried to focus on staying safe while everything felt unstable. After it ended, the environment around me was changed, and familiar places looked different. I felt disoriented and exhausted. Even later, reminders of the event brought back the fear and helplessness I felt at the time. The sense of unpredictability stayed with me.

Note: Seven additional stress stimuli were generated following the same methodology. The remaining 10 stress stimuli are from [Ben-Zion et al. \(2025\)](#).

A.1.2. Neutral Stimuli

Routine activity prompts.

1. I spent the morning organizing my desk, moving papers and supplies to their proper places. I noted which documents needed filing and which could be recycled. As I worked, I followed a routine I had used for months, checking each drawer and shelf for misplaced items. Occasionally, I paused to review a calendar entry or write a note about a task to complete

later. By midday, my space felt orderly, and I moved on to other tasks with a clear sense of what remained to be done.

2. Yesterday, I completed a series of standard reports for my team. I collected data from previous weeks and carefully entered the numbers into a spreadsheet. I double-checked formulas and ensured all totals matched the original sources. Throughout the process, I followed the usual steps and made note of any missing information to address later. By the end of the day, the reports were complete, filed in their designated folders, and ready for review.
3. I took a walk through my neighborhood, observing the streets, sidewalks, and buildings. I noticed which areas had been recently cleaned or repaired and made a mental note of anything that might need attention in the future. As I moved along, I greeted familiar neighbors and exchanged brief, polite conversations. The rhythm of walking and observing gave me a sense of structure to my afternoon.
4. I prepared a simple meal using ingredients I had on hand. I followed standard steps for washing, chopping, and cooking, making sure to measure quantities carefully. I monitored the cooking process, adjusting the heat as needed and tasting occasionally. Once the meal was ready, I set the table and ate while following my usual routine for cleaning up afterward.
5. I reviewed the schedule for my upcoming week, noting appointments and deadlines. I organized tasks according to priority and time needed for completion. I made adjustments to ensure there were no conflicts and wrote reminders for each item. This process allowed me to clearly see the workflow and plan efficiently for the days ahead.
6. I spent time sorting through a collection of documents. Each item was categorized, labeled, and placed in its appropriate folder. I checked for duplicates and ensured that everything was arranged logically. This methodical approach helped me keep the collection organized and easy to navigate whenever I needed to reference it.
7. I updated a list of routine maintenance tasks for my home. I checked items that had been completed and added new tasks based on recent observations. I scheduled reminders to follow up on certain items and ensured that all necessary supplies were accounted for. The list provided a clear overview of what needed attention in the coming weeks.

8. I observed the flow of traffic around my usual commute route. I noted the times when intersections were busy and when they were quiet. I paid attention to signals, stop signs, and pedestrian crossings, making mental notes of areas that could pose delays. The observation helped me plan my route more efficiently for the next day.
 9. I read through a set of instructional manuals for office equipment. I followed the diagrams and explanations, noting any differences from previous versions. I highlighted sections that were most relevant for upcoming tasks and marked pages for future reference. The reading required focus but followed a predictable and consistent structure.
 10. I organized digital files on my computer, creating folders for different categories and moving items accordingly. I checked file names for accuracy and consistency, ensuring that everything was easy to locate later. I deleted unnecessary duplicates and backed up important items to a secure location.
 11. I completed a series of standard exercises at my desk, stretching my arms and legs and following a routine I had practiced regularly. I paid attention to posture and movements, ensuring each exercise was done correctly. The routine provided structure to my afternoon and allowed me to continue working with minimal interruption.
4. Arrange a series of documents on a desk in numerical order. Review each document for accuracy. Correct any documents that are out of sequence. Stack the documents neatly. Ensure that the top document matches the first number in the series. Complete the sequence verification for all sets of documents.
 5. Launch the software application on the computer. Navigate to the settings menu. Review each available option in order. Select each option and observe its effect on the interface. Return to the main menu after testing each option. Record any changes made. Close the application once all settings have been verified.
 6. Gather a set of tools and arrange them by size. Inspect each tool for completeness and condition. Place the largest tool first, followed by smaller tools in order. Confirm that all tools are visible and accessible. Adjust the arrangement as needed. Repeat until the layout is orderly and complete.
 7. Prepare a series of containers for labeling. Open the first container and place a label inside. Close the container securely. Repeat the process for each subsequent container. Confirm that labels are placed correctly and consistently. Ensure that all containers are properly sealed before proceeding.
 8. Open the device's control panel. Check each indicator for status. Adjust settings using the designated buttons. Observe the display for confirmation of changes. If adjustments are required, repeat the procedure until the desired configuration is achieved. Close the control panel once all settings are verified.
 9. Sort a collection of cards by type. Lay the cards face up on a flat surface. Group similar cards together. Verify that each card is in the correct group. Correct any misplacements immediately. Stack the cards neatly once the sorting is complete. Ensure all groups are organized consistently.
 10. Connect the power adapter to the device. Verify that the connection is secure. Turn on the device using the power button. Observe the indicator lights for operational status. Press the mode button to select the desired function. Monitor the display for confirmation. Turn off the device after completing the procedure.
 11. Prepare a set of papers for filing. Check each paper for completeness. Insert each paper

Procedural instruction prompts.

1. Place all objects on a flat surface. Check that each object is aligned with its neighbors. Adjust positions as necessary. Move objects sequentially from left to right. Inspect the overall arrangement for consistency. If any object is misaligned, correct its position immediately. Repeat the sequence until all objects are stable and evenly spaced.
2. Select a group of containers and place them on a workbench. Open each container carefully. Examine the contents for order and completeness. Replace the contents in the original container in the same sequence. Close the container securely. Move to the next container and repeat the process. Verify all containers are correctly organized before finishing.
3. Turn on the device by pressing the power button. Wait for the indicators to show readiness. Set the device to the default mode using the mode selector. Observe the display for confirmation of the selected mode. Press the appropriate button for any secondary function

into the correct folder. Ensure the folder is closed securely. Repeat the process for all remaining papers. Verify that all folders are correctly labeled and organized. Complete the filing procedure.

A.2. Model Specifications

Table 3 lists all 35 models evaluated in this study.

Model	Parameters	Type
google/gemma-2-9b	9B	base
google/gemma-2-2b	2B	base
google/gemma-3-1b-it	1B	instruct
google/gemma-3-270m	270M	base
google/gemma-3-270m-it	270M	instruct
Qwen/Qwen3-8B	8B	base
Qwen/Qwen2.5-0.5B-Instruct	0.5B	instruct
Qwen/Qwen2.5-7B-Instruct	7B	instruct
Qwen/Qwen3-0.6B	0.6B	base
Qwen/Qwen3-1.7B	1.7B	base
Qwen/Qwen3-4B	4B	base
Qwen/Qwen3-14B	14B	base
deepseek-ai/DeepSeek-V2-Lite	16B	base
deepseek-ai/DeepSeek-R1-Distill-Qwen-7B	7B	instruct
microsoft/MediPhi	3.8B	base
microsoft/MediPhi-Instruct	3.8B	instruct
microsoft/MediPhi-PubMed	3.8B	base
microsoft/Phi-3-mini-4k-instruct	3.8B	instruct
microsoft/Phi-3.5-mini-instruct	3.8B	instruct
microsoft/Phi-4-mini-instruct	3.8B	instruct
microsoft/Phi-4-mini-reasoning	3.8B	instruct
microsoft/UserLM-8b	8B	instruct
meta-llama/Llama-3.2-1B	1B	base
meta-llama/Llama-3.2-1B-Instruct	1B	instruct
meta-llama/Llama-3.2-3B	3B	base
meta-llama/Llama-3.2-3B-Instruct	3B	instruct
meta-llama/Llama-3.1-8B	8B	base
meta-llama/Llama-3.1-8B-Instruct	8B	instruct
meta-llama/Meta-Llama-3-8B-Instruct	8B	instruct
meta-llama/Meta-Llama-3-8B	8B	base
LenguajeNaturalAI/leniachat-qwen2-1.5B-v0	1.5B	instruct
PipableAI/pip-sql-1.3b	1.3B	base
sequelbox/Qwen3-4B-Thinking-2507-DES-Reasoning	4B	instruct
nvidia/Nemotron-Content-Safety-Reasoning-4B	4B	instruct
DavidAU/Llama3.3-8B-Instruct-Thinking	8B	instruct

Table 3: Models evaluated in this study, grouped by family. Type indicates base (pretrained) or instruct (instruction-tuned). All models were sourced from HuggingFace.

Documenting Rural Gatherings in Aging Japan: Social Context and Language Use in Interaction at a Mobile Supermarket

Haruka Sakai¹, Rui Sakaida²

¹ Hiroshima University, ² Future University Hakodate
¹ 1-1-1 Kagamiyama, Higashi-Hiroshima, Hiroshima, Japan
² 116-2 Kameda Nakanochi, Hakodate, Hokkaido, Japan
¹ hasakai@hiroshima-u.ac.jp, ² sakaida@fun.ac.jp

Abstract

This paper presents a documentation framework and an exploratory analysis of language use in everyday interactions at rural gatherings in aging Japan, a communicative setting shaped by distinct social contexts that remain largely absent from existing language resources. Drawing on studies of face-to-face encounters, we propose a typology of rural gatherings and examine mobile supermarkets (vehicles that transport and sell daily necessities at scheduled stops in areas that lack fixed retail stores) as a case study. We present a preliminary analysis based on a community-mediated recording methodology. The quantitative findings reveal that conversational hot spots occur immediately following the encounter and transaction phases, indicating that participants experience these encounters as occasions for social connection rather than mere commercial transactions. The qualitative findings from the interaction analysis demonstrate how participants simultaneously manage work and conversation through vocal, bodily, and temporal resources in a social context. We discuss how these findings illuminate dimensions in social contexts that require interdisciplinary investigation beyond what existing language resources currently capture.

Keywords: social context, rural gathering, language use in social interaction

1. Introduction

Language use is situated within social contexts shaped by various elements, including background knowledge, interpersonal relationships, and physical settings. As research involving social interaction analysis and linguistic anthropology has demonstrated, participants deploy multimodal resources to make their actions recognizable and accountable to others within specific social contexts. Understanding how these contexts shape language use requires an empirical study of diverse communicative settings, yet communities underrepresented in existing corpora remain largely undocumented. In this paper, we document interactions in rural Japan, where super-aging and depopulation have produced distinctive social settings.

Recent advances in corpus linguistics have enabled large-scale documentation of naturally occurring conversations, providing invaluable resources for linguistic research. The *Corpus of Everyday Japanese Conversation* (CEJC) contains over 200 hours of audio- and video-recorded daily activities with detailed annotations (Koiso et al., 2022). Studies using the CEJC have examined turn-taking, backchannels, and sequential organization across activity types (Koiso et al., 2022), including participant behavior in service encounters (Kishimoto, 2021). A preparatory survey (Koiso et al., 2016) characterized the types and settings of daily conversations, providing a baseline for corpus design. In addition, related projects, such as CEJC-child, which aims to document 100 hours of

recorded interaction, are underway (Koiso et al., 2024).

However, these corpora exhibit a clear urban bias, with participants drawn primarily from Tokyo and surrounding areas (Koiso et al., 2022). This urban-centric focus neglects approximately 9.3% of Japan's population residing in rural areas (Ministry of Internal Affairs and Communications, 2025)—communities that are most impacted by super-aging and depopulation. In rural regions, 39.7% of residents are elderly, compared to 28% nationally (ibid.). This percentage continues to increase as younger populations migrate to urban centers.

A distinctive feature of rural life is the emergence of new communicative settings. With the decline of local retail establishments, services such as mobile supermarkets, which are small trucks that transport and sell daily necessities at scheduled stops near customers' homes, have introduced new community spaces, especially for elderly residents. These localized communication practices require urgent documentation before they change significantly or lose their current form. Because they arise from particular configurations of social relationships, physical environments, and community needs, they can provide empirical evidence of how social context shapes language use in ways not captured by existing corpora.

To capture rural communication, address the urban bias in existing language resources, and facilitate comprehensive studies of language use in interaction, this paper proposes a systematic framework for documenting rural communication

and describes aging society as a communicative practice through micro-level interaction analysis. Based on Goffman's (1963) concept of gatherings, we propose four types of rural gatherings and examine mobile supermarkets as a case study. In doing so, this study also aims to contribute to a broader understanding of the diversity of communicative settings, which may inform the development of annotation frameworks in future work. We address the following questions: (1) What temporal patterns characterize conversations at mobile supermarkets, and what do they reveal about the social functions of these encounters? (2) How do participants coordinate multiactivity through multimodal resources? (3) What dimensions of social context does this communicative setting present that are not yet represented in existing language resources?

2. Mobile Supermarkets as Rural Gatherings

2.1 Gathering

The concept of gathering, introduced by sociologist Goffman (1963), refers to individuals being co-present in the same space and sharing a social situation. Gatherings can occur not only during scheduled meetings but also in public places, including parks or stations, even when individuals have no explicit intention to "gather." Goffman distinguished between focused interaction, in which individuals engage in a joint activity such as a conversation, and unfocused interaction, in which individuals merely share the same environment while maintaining mutual awareness without overt engagement. Previous studies on multimodal interactions have examined the transitions between focused and unfocused interaction. Mondada (2009) analyzed how engagement and disengagement are organized through utterances, gazes, and bodily movements. However, these studies primarily examined urban contexts, including public spaces. Rural communities offer distinct features of gathering, in which recurring encounters among familiar individuals and the physical infrastructure of daily life produce communicative settings not found in urban environments.

Goffman's framework, though developed in urban contexts, applies equally to rural areas, wherein distinct social and physical features produce particular forms of gathering. "Acquaintanceship" (Goffman, 1963) leads to particular forms of behaviors and procedural norms that differ from the interactions among unacquainted individuals more commonly found in urban contexts. For instance, an acquaintanceship in rural communities can function as a social resource that extends the boundaries of interaction beyond physical spaces. Although Goffman (1963) emphasized physical structures such as doors and walls as special boundaries, Sakai (2024a) illustrated that the predictability of encounters

among historical acquaintances can lead to verbal interactions occurring beyond physical boundaries, such as outside the door or separated by a wall, without direct visual contact. This case suggests that rural gatherings are shaped not only by physical context but also by participants' shared social history and expectations of encounters.

2.2 Categorizing Rural Gatherings

To document the communicative ecology of rural communities, we classify rural gatherings according to the social variables that shape participants' expectations: location, social relationships, and the predictability of encounters (Sakai, 2024a). Such classification is necessary because rural communication, similar to urban communication, spans diverse settings, from private to public. Without a systematic typology, documentation risks being biased toward more accessible settings while neglecting everyday but less visible ones, such as brief encounters on the road. Thus, a comprehensive account of rural communicative ecology requires first establishing a typology and then documenting each type systematically. This typology draws on Goffman's (1963) distinction between focused and unfocused interaction and Koiso et al.'s (2016) situational classification for corpus design, combined with fieldwork observations in rural communities.

1. At-home gatherings: Everyday communication within households or visits from community members in private settings.
2. Chance encounters: Brief interactions that occur when individuals meet unexpectedly in public spaces, such as along roadsides, near bus stops, or at village offices.
3. Community events: Scheduled gatherings, such as community meetings or traditional festivals (including practices), in which multiple parties interact on equal footing.
4. Service encounters: Interactions between staff and customers (referred to as users in this paper), such as in shops, banks, or post offices. These participants, while familiar with each other in small rural communities, take on different roles, such as clerk and customer.

As a first step toward documenting all four types, this study focuses on service encounters. Unlike urban service encounters between strangers, rural ones tend to involve community members with shared social histories who take on the roles of clerk and customer. Among these encounters, mobile supermarkets represent a new retail form, operating in novel physical environments such as vacant lots in front of users' homes. This combination of familiar social relationships and dynamic commercial settings makes mobile

supermarkets productive sites for examining how social context shapes interaction.

2.3 Mobile Supermarkets

Mobile supermarkets belong to the category of service encounters but differ from ordinary shop interactions in several ways. As a communicative device in depopulated rural areas, mobile supermarkets generate temporary social contexts in which transactions and community interaction converge.

Sustaining daily life in depopulated and aging rural areas is a critical issue. The main challenge is not aging per se but the mobility decline that accompanies it. Many rural areas suffer from so-called shopping difficulties, particularly affecting elderly individuals who no longer drive and reside in geographically isolated regions, such as mountainous or island communities. This is part of the “food desert” problem, referred to in Japan as the issue of “vulnerable shoppers” (*kaimono nanmin* in Japanese).

Under these circumstances, mobile supermarkets have gained increasing importance as one of the key solutions. These vehicles transport daily necessities and operate out of small trucks stationed near customers’ houses or in local open spaces (Figure 1). Their primary users are older adults who live too far from a fixed-location retail establishment to access them independently. The driver typically serves as both a driver and a salesclerk. Despite attracting attention from the social, administrative, and academic sectors, systematic studies examining the communicative dynamics of these mobile shopping environments remain limited. To the best of our knowledge, Takanashi et al. (2023) and Sakai & Sakaida (2024) are among the few studies to date that have analyzed language use and interactions in the context of a mobile supermarket. The present study extends this work by examining how mobile supermarkets function as a communicative device that is grounded in and generates particular social contexts, interactional structures, and participant behaviors not observed in fixed-location service encounters. Research on service encounters has examined interactions in fixed commercial spaces, including both unacquainted (e.g., Harjunpää et al., 2018) and acquainted participants (Laurier, 2013). Even in acquainted settings, the physical boundaries of the shop are actively oriented to by participants as a resource for organizing the interaction (Sakai, 2024b). Mobile supermarkets differ in that the interactional space is temporarily constituted and dissolved, and participants must collaboratively manage the closure of the space itself. The co-occurrence of these features distinguishes this setting.



Figure 1: A mobile supermarket operating in a vacant lot in front of a user’s house.

From the perspective of interaction analysis, mobile supermarkets offer significant insights into language use within social interactions. As Sakai & Sakaida (2025) have pointed out, the arrival of the vehicle temporarily transforms the space in front of houses or community centers into a shopping environment. The arrival and departure of the vehicle serve as triggers for the emergence and dissolution of “interactional spaces,” respectively (Mondada, 2009). Capturing these temporal transformations through fine-grained interactional analysis requires video-recorded data. This approach yields valuable cases for analyzing how interactional spaces emerge, where their boundaries lie, and how rural life can be understood as a communicative practice. Understanding these practices requires interdisciplinary methods that attend to the social contexts in which interaction occur, shaped by the particular features of rural communities, including social relationships, communicative devices, and physical resources.

3. Methods for Data Collection

3.1 Recording Procedure

Recording naturally occurring conversation in rural communities requires methods adapted to the social context in which interactions occur. For this study, we adopted a situation-specific approach (Koiso et al., 2022) in which researchers visited a specific site to record particular activities, as it allows focused observation, which leads to high-quality audiovisual data.

A central challenge was establishing rapport with research participants, especially when recording was conducted in private spaces, such as the area directly in front of users’ homes. Consented recordings were made possible through the cooperation of the driver, who had prior research experience (Oshiro, 2019) and helped explain the research purpose to users and coordinate the procedure. This highlights how community members with a research understanding can act as mediators, facilitating data collection within the social context of the community. Participant

information sheets, consent forms, and consent withdrawal forms were also adapted for elderly participants through means such as enlarged text and verbal confirmation of content.

The recordings were conducted by two researchers using two handheld cameras, two 360-degree cameras mounted on top of the vehicle, two IC recorders attached to ensure full visual coverage, and an IC recorder attached to the driver (Figure 2, partly reproduced in Sakai & Sakaida, 2025). This setup enabled comprehensive audio and video recordings of conversations that occurred around the vehicle. Importantly, the recordings focused not only on the shopping itself but also on the periods before the vehicle’s arrival and after its departure. This study aims to capture the interactional transformation that occurs as mobile supermarkets become temporary marketplaces. This approach allowed us to successfully document conversations that extended beyond the transaction phase and even continued after the vehicle had departed.

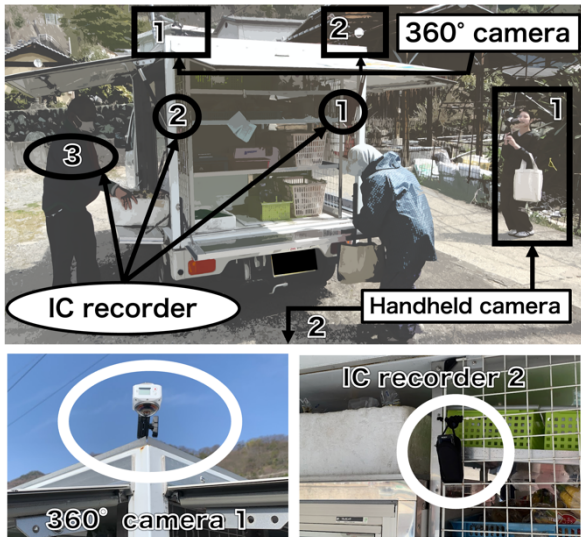


Figure 2: Recording setup with two handheld cameras, two 360-degree cameras, and three IC recorders (view from handheld camera 2)

3.2 Recorded Data

Recordings were conducted on March 22, 2024, after obtaining written consent on March 15. Table 1 presents an overview of the dataset. In the table, letters and subsequent numbers indicate the

villages and specific stop points within each village. Recordings were not taken at stops A2, A3, and B2, although sales occurred.

As shown in the table, the dataset includes eight mobile supermarket sessions that took place in five villages (A–E), involving one driver and 15 users. Each stop had 1–4 users, with session durations ranging from 4 to 14 minutes. Some users were already present before vehicle arrival. Conversations occurred between the driver and users and among users themselves. While modest in scale¹, the dataset captures communicative settings absent from existing Japanese conversation corpora, including the CEJC.

	Stop number	Number of users	Arrival	Departure	Duration
1	A1	1	9:57	10:01	0:04
2	A4	3	10:45	10:58	0:13
3	B1	1	11:07	11:18	0:10
4	B3	2	11:36	11:52	0:15
5	B4	1	11:59	12:09	0:10
6	C1	4	14:24	14:32	0:08
7	D1	2	14:56	15:03	0:07
8	E1	1	15:21	15:30	0:09
Total	8	15	-	-	1:16

Table 1: Data overview

4. Preliminary Analysis: When they talk

We examined when conversations occurred during shopping (Figure 3). Each horizontal bar in Figure 3 represents a shopping session, with segment types indicating user, for example, “A1-1,” and shopping phase (see caption for details). Sessions were normalized to equal lengths for comparison from arrival to departure.

As shown in Figure 3, conversational hotspots occurred at two distinct moments: shortly after the vehicle’s arrival (0–1 minute), when gray or black segments tend to appear, and after the completion of transactions, where black segments frequently follow striped segments. Of the 15 sessions, 9 showed conversational activity within the first minute of the encounter, while all sessions showed conversation continuing after payment. Analyzing the contents of these interactions revealed distinct characteristics at different stages.

¹ Full transcription of the dataset is in progress. As an indication of scale, a one-minute segment (B1-1, 8:45–9:45) from the cleanup phase analyzed in Section 4 and 5, beginning at line 1 of Excerpt 1, contains 141 morphemes (driver: 96, user: 45), segmented using MeCab with UniDic following the CEJC short-unit

conventions (Koiso et al., 2022). This count includes two inaudible segments due to the recording conditions. Given the variation in conversational density across sessions, this figure should not be extrapolated to the full dataset.

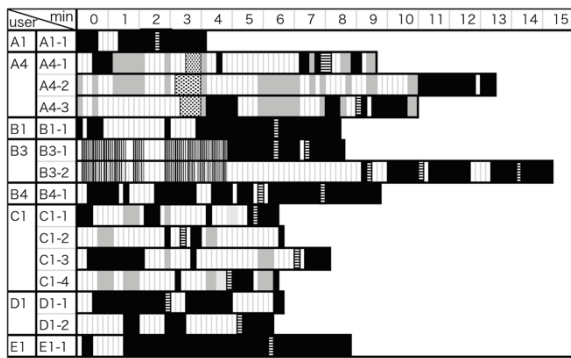


Figure 3: Temporal distribution of conversation during shopping (n=15). Black segments indicate conversations with the driver; grey segments, conversations with other users; dotted segments, the user talking with both the driver and another user; horizontal striped segments, transactions; vertical striped segments, conversations with unidentified participants; and blank areas, silence.

Conversations initiated immediately after the vehicle's arrival typically focused on product availability (from user to driver, (1)(2)) or inquiries about specific items (from driver to user, (3)(4)). The following examples present the original Japanese utterances, their English translations, and the direction of each utterance. Each user is assigned an identifier to distinguish individuals; for instance, the three participants at stop A4 are labeled A4-1, A4-2, and A4-3.

- (1) Kyoo wa sashimi ga aru ka nee.
Do you have sashimi 'sliced raw fish' today?
(A4-1 to driver)
- (2) Gyuu ga aru?
Do you have beef?
(A4-2 to driver)
- (3) Naomi neesan, daburu sofuto iran?
Big Sister Naomi, don't you need *Double Soft? *the name of bread
(driver to B3-1)
- (4) Kyou wa nanka iru mon ga aru n su ka no
Do you need anything today?
(driver to B4-1)

The following payments often concern the next shopping visit, representing a typical closing routine initiated primarily by the driver, as shown in (5)~(8).

- (5) Hoija maa raishuu getsuyoobi mata kimasu
Well then- I'll come again next Monday.
(driver to B1-1)
- (6) Hoija obachan, arigatone.
Well then, auntie, thank you.
(driver to B3-2)
- (7) Arigatosan deshita
Thank you very much.
(driver to B3-2)
- (8) Tsugi wa getsuyoubi wa oru?
Will you be here next Monday?
(driver to B4-1)

These patterns suggest that the temporal structure of mobile supermarket interactions is organized not only by the commercial task at hand but also by social expectations in which acquaintanceship and regular encounters build up shared expectations about how the conversation unfolds. The driver's familiar address terms and the routine confirmation of the next visit index a relationship extending beyond the transaction.

When conversations continue well beyond payment completion, this typically results not from extended transaction-related talk but from a shift toward small talk, as shown in the following excerpt. Here, DRV represents the driver and USR indicates the user B1-1. We refer to Jefferson (2004) for the transcription conventions.

```

01 DRV etto, hyaku:, hyaku::, nanajuu to:,
      ni en. yossha.
      Um, 100- 170 and 2 yen. Alright.
      (Three lines omitted)
05 DRV arigato san deshita.
      Thank you very much.
06 DRV hoija ma- raishuu getsuyoobi mata
      kimasu.
      Well then, I will come again next Monday.
07 USR getsuyoubi yasumi ja nai ken no?
      Isn't Monday a holiday though?
      (Five lines omitted, talking about holidays)
13 USR raigetsu kara mata nanka saijitsu
      ( )
      From next month there are some other national holidays.
14 USR attakai na:
      I'm not sure of it.

```

After confirming the next visit (line 6), the topic gradually shifts to holidays (lines 7–14). Furthermore, as the conversation continues beyond line 14, the driver and user stay at the same location for a while. However, it is important to note that the driver has to proceed to the next shopping point, as another scheduled stop follows shortly after the conversation. Consequently,

examining conversation alone is insufficient, as the driver was simultaneously engaged in both small talk and preparations for departure. Understanding how participants manage these competing demands at a micro level reveals the organization of this setting, which arises from its particular social context, including relationships, physical environment, and interactional phase. The next section describes how the driver managed these multiple activities.

5. Case Analysis: How they talk

The term ‘multiactivity’ refers to how people talk and interact when coordinating or advancing (or not) two or more courses of action simultaneously and with respect to each other (Haddington et al., 2014). After completing payment, DRV (driver) begins cleanup work, projecting the upcoming departure for the next stop, while conversing with USR (user B1-1). In the case below, an elderly customer slowly places their wallet back into their bag and prepares to leave. The DRV’s behavior shows no sign of rushing the customer.

To clarify how the DRV manages these two activities, we examined the organization of the interaction and described the DRV’s orientation at each moment. Excerpts 1–3 were taken from a single case (DRV and B1-1 user), and the entire cleanup process lasts approximately two minutes.

5.1 Keep Engaging with Voice

The cleanup phase begins immediately after payment. In Excerpt 1, recorded during this phase, DRV and the USR continue their conversation. Following the completion of payment (lines 1-13), the topic shifts to a discussion of the next encounter, which serves as a projection of the closing (after line 15). Simultaneously, DRV returns the handheld transaction device to its place, while USR puts their wallet into their bag (lines 15-28). Immediately afterward, DRV moves to close the right-side door of the vehicle (line 30). For conventions of multimodal transcription, see Mondada (2018).

Excerpt 1 (with multimodal transcription)

01 DRV etto, hyaku:, (1.78) hyaku::, nanajuu
Um, 100- 170
02 to:, (1.50) ni en.
and 2 yen.
03 (0.2)
04 USR n:n
mm- hmm
05 (1.5)
06 DRV yossha.
Alright.
07 USR (nn) hai hai
yeah, yeah
08 (1.4)
09 DRV hai.
yeah
10 (0.6)
11 DRV hai.
yeah
12 (0.3)
13 DRV arigato san deshita.

Thank you very much.
(0.3)
14 DRV .hh hoi ja ma- raishuu getsuyoobi mata
Well then, I will come again next Monday. (ll. 15-16)
15 kima[ssu hhh
16 USR [(u getsu-)
U, Mon-
17 (0.8)
18 USR %getsuyo%obi yasumi ja nai ken no?
Isn’t Monday a holiday though?
%.....%DRV-->
19 (0.4)
20 DRV yasumi () ja nakaroo, ho- kongetsu
It shouldn’t be a holiday, uh- this month
21 wa- kongetsu tte yuu ka %are
I mean this month, um
usr -->%,,,-->
22 wa yasumi nakaroo nya.
Shouldn’t be a holiday.
usr -->%front-->
23 USR ano::, u-, *raige*tsu nara [nya
Um, uh, if it’s next month then
24 DRV [u*
uh
drv *.....*USR-----*,,,
25 DRV raigetsu naran [to
if it’s next month
26 USR [raigetsu %ka (ra)
from next month
usr -->%,,,
27 USR mata nanka +saijitsu (+)
there’s another holiday or something
drv +returns cash register
to designated position
drv +steps back
28 (0.3)
29 DRV attakai +na:.
I’m not sure there is.
+walks to right side
of vehicle-->
30 (0.4)
31 DRV kasana%tta n %ga atta kai na:
I wonder if there were overlapping ones
usr %.....%right-->
32 USR un anmari:, sa-, an +sangatsu wa
yeah not really, uh, um, March
drv -->+
33 +() [(yoran)
(unclear)
34 DRV [u+:n
uh-huh
+.....+carries box to left side
of vehicle
35 (0.3)%(0.5)
usr -->%,,,
36 DRV [(nho)
37 USR [daibunto atta mon nee.
There were quite a lot, you know.
38 (0.7)
39 DRV are ga mata koma-, nga komaru n ya: ,
That’s trouble-, it’s troublesome,
40 komaru n yo +no:
troublesome, you know.
41 -->+



Figure 4: Body movements in lines 28-32

At this moment, DRV does not orient solely toward the conversation but continues the topic while speaking in a louder voice (lines 28-32). DRV steps back and begins to display their orientation to the work, while talking with USR (line 28, Figure 4, left). Also, USR turns their head slightly toward the right side of the vehicle (line 32, Figure 4, right) immediately after DRV begins to move in that direction (line 30, Figure 4, center), and before DRV's arrival at the place. It indicates the orientation to maintain their interactional space (Mondada, 2009), which is about to transform, accompanied by the cleanup work. In this way, the projection of the movement of the body allows a prior orientation toward the intended location.

Subsequently, DRV places the box back on the shelves (line 35), producing audible utterances, even while moving out of the USR's line of sight. This dual engagement is possible because conversation primarily requires the speaker's voice. Occasionally, DRV turns their head toward USR so that their voice is easier for USR to hear and to display their engagement in the conversation. This raises the following question: How does DRV manage activities when a conversation requires bodily engagement?

5.2 Suspension Work for Conversation

In Excerpt 2, the left door has already been closed, and DRV has moved to the right side of the vehicle. Ethnographically, DRV is a resident of Village B, where the selling in this case occurs. In this interaction, they discuss minor changes in their village—specifically, the abundance of harvested lemons at the village elementary school.

Here, we observe that DRV, while working with their arms, uses the deixis expression 'asoko (there)' during the ongoing conversation. This utterance occurs precisely as he attempts to close the right-side door. Notably, it is difficult to close this door with one arm; it requires the use of both hands to pull down and lock it securely. In Excerpt 2, he uses a pointing gesture to indicate the location verbalized as 'asoko (there)' and 'shoogakkoo (elementary school)' in lines 7-11.

Excerpt 2

```
01 USR SOOJOO ittaka no: dokono
    Did she go to SOOJOO village or where
    drv >>walks to right side of vehicle-->
02 (hoi demo [          ])
    but well
03 DRV [nn byooin, +mo- aa(h)
    mm hospital, or- aah
    drv -->+closes lower
    right door-->
04 byoo(h)in, +ka, +[asoko ka+:,
    hospital or there
05 USR [a (          )
    oh
    drv -->+,,,+points---+,,,-->
```

```
06 + (byooin[ka)
    hospital
07 DRV [.h +ima sa:,
    now you know
    drv -->+try to close upper right door-->
    -->+points with left hand
    with right hand
    placed on door-->
08 (0.2)
09 USR +n?
    hm?
    drv -->+hold-->
10 (0.1)
11 DRV asoko shoogakkoo n +toko de:
    at that elementary school
    -->+,,,-->
12 USR +are, [remo-
    that lemo-
13 DRV [yooke aru no wa nan-
    what is with having so many-
    drv -->+places both hands
    on upper right door-->
14 (0.2)
15 USR remon shori.
    Lemon processing.
16 (0.4)
17 DRV +remon nanka shori n shiy- ano:
    Are they processing lemons
    -->+closes upper right door-->
18 (0.2)
19 USR hoijake are wa ne (.) m- mae kara,
    so well that you see (.) from before
20 (0.6) ano: +chotto:, (0.4)
    um a little
    drv -->+...-->
21 TAKA+RA no hito rano:
    the people of Takara association
    drv -->+places both hands on head-->>
```



Figure 5: Pointing gesture while verbalizing 'asoko' (there) in lines 7-17

Although DRV initially attempts to close the door (line 7, Figure 5, left), they momentarily suspend the action and point in the direction referred to by 'asoko' (there) (line 7-11, Figure 5, center). At this moment, DRV rests their right arm on top of the door and uses their left hand—the arm closer to the direction of reference—to make the pointing gesture. This suggests that the door-closing task is not reset, but rather temporarily suspended (Raymond & Lerner, 2014).

The door-closing task resumes at line 13. After completing the pointing gesture, the right hand (used to indicate 'asoko'), returns to the door (Figure 5, right). While continuing the conversation and turning their head toward USR,

DRV closes the door. This illustrates how DRV's orientation shifts from cleanup to conversation, expressed through embodied behavior.

5.3 Waiting for Bodily Disruption

As shown above, DRV coordinates their bodily movement and conversation during cleanup work. This coordination is achieved by selectively choosing the orientation at each moment. DRV suspends their door-closing work when engaging in conversation that requires their bodily behavior.

The following excerpt demonstrates a case in which the USR's body positioning disrupts DRV's efficient work. In Excerpts 1 and 2, DRV moves around the vehicle to close its doors. Typically, the preferred sequence is to close the doors in the order: left, back, and right. In many instances, DRV follows this routine. However, in the next case, USR remains standing directly under the back door, preventing DRV's access. Rather than asking USR to move, DRV alters the sequence: they close the left door first, then the right, and finally the back. Excerpt 3 shows the moment when DRV closes the back door.

Excerpt 3

01 DRV u:so(h):(h): hh nani shiyon jaroo
No way:: what are they doing
02 [re+mon ja noni
even though it's a lemon
+walks to rear of vehicle-->
03 USR [shiran.
I don't know
04 (0.1)
05 USR hh h sore ga,
The thing is
06 (0.7)
07 USR koko no ma+e ka(ra) ni kai hodo
About twice recently
drv -->+reaches hand upper door-->
08 shiyon yo.
they've been doing it
09 (0.2)
10 DRV ee +ee @hoija-, kyo@o mo, kyoo
no- today too, today
-->+puts hand on door-->
usr @.....@steps back-->
11 mo to+ori yotta yo:, @[kuru+ma.
they passed by too, the car
12 USR [kyoo mo mata
today again too
-->+lowers door a little-+holds-->25
usr -->@moves bag back
13 USR kooiu yoona shitotta?:
were they doing something like this?
14 DRV un.
yeah.
15 (0.2)
16 DRV kuruma kiyotta mon.
A car was coming
17 (0.1)
18 USR a[a::
I see
19 DRV [washi ga ie-, h. hh.
when I() home-
20 (0.1)
21 USR a ga [na:, (tsundokin) no:,
ah you know, (unclear) the

22 DRV [i- iku toki ni
wh- when going
23 USR hoide so[re wa ne:,
and so that you see
24 DRV [un.
yeah
25 (0.3)+(0.3)
drv -->+closes door-->
26 USR remon o chan- [mae wa shiboru tte
lemons properly- before they squeeze (them)
27 DRV [(nn.)
mm.
28 USR [() shichon+ ka omottara:,
I thought they were doing ()
29 DRV [nn nn.
mm mm
drv -->+
30 DRV n: :n.
mm mm.

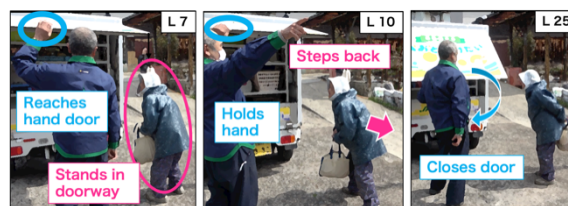


Figure 6: USR stepping away from the doorway while talking with DRV in lines 7-25

USR continues to stand in the rear area, even as DRV approaches the rear of the vehicle and places their hand on the upper part of the back door (line 7, Figure 6, left). After that, USR gradually begins to step back (line 10, Figure 6, center), while talking with DRV. Shortly after USR moves away from the doorway, DRV lowers the door and closes it completely (line 25, Figure 6, right). Following this excerpt, DRV releases their hand from the door, turns fully toward USR, and displays a single orientation toward the conversation.

Across these three excerpts, the driver systematically prioritized the conversation over their tasks to manage both as multiactivity. This practice, realized vocally (Excerpt 1), bodily (Excerpt 2), and temporally (Excerpt 3), enacts orientation toward care shaped by the interaction between participants' backgrounds, the communicative context of a brief encounter, and the immediate situation. These dimensions of social context are difficult to capture without an interdisciplinary, micro-level analysis of the kind presented here.

6. Concluding Remarks

This study documented language use in interactions at a mobile supermarket in a rural community, a communicative setting largely absent from existing language resources. The typology of rural gatherings provides a framework

for systematic documentation in further research, and the community-mediated recording methodology offers a foundation for future work in similar contexts.

The quantitative analysis revealed that conversational activity clusters around the encounter and post-transaction phases, indicating that participants orient to these encounters as occasions for social connection.

The qualitative analysis demonstrated how the driver managed work and conversation simultaneously as multiactivity, adjusting his interactional mode vocally (Excerpt 1), bodily (Excerpt 2), and temporally (Excerpt 3), consistently prioritizing conversation. The cleanup period operated as a buffer zone in which the driver “waited” for the user’s movement, avoided applying pressure, and altered routines. For older users with limited opportunities for social engagement, these moments functioned as meaningful opportunities for community connection. A systematic understanding of such “waiting” practices as a form of social support warrants further investigation across rural areas, demographic groups, and institutional settings.

Regarding our third research question, this communicative setting presents a social context not captured by existing language resources. The temporal structure of interaction is organized by social expectations built through repeated encounters, as shown in the quantitative analysis. Furthermore, as the qualitative analysis revealed, the temporary shopping space was dissolved while the participants were talking after payment. The temporary nature of this setting was noted in previous work (Sakai & Sakaida, 2025). Building on this, the present study reveals the concrete multimodal practices through which this dissolution is accomplished. How participants manage this process is shaped by the particular social context of this setting, including relational history, and cannot be observed in fixed-location service encounters, where the commercial space requires no dissolution. While acquaintanceship can produce extended post-transaction conversation in fixed-location settings (Sakai, 2024b), the practices documented here are qualitatively different: Participants collaboratively managed the dissolution of the commercial space while sustaining conversation (Excerpts 1–3). This interactional challenge arises only in temporarily constituted settings such as mobile supermarkets, shaped by the social conditions of rural depopulation and aging. Documenting these dimensions through interdisciplinary methods contributes to a richer understanding of the diversity of communicative settings, and offers empirically grounded factors, such as relational history, interactional phase, and spatial configuration, that could enrich context-aware

approaches in the computational modeling of social interaction.

7. Limitations

This study is based on a limited number of recordings from a single rural community and focuses on one driver–user pair. Therefore, it may not capture the full diversity of mobile supermarket interactions in Japan. Future research should include multiple cases across different settings to explore variations in interactional practices. In particular, empirical comparisons with urban service encounters and rural fixed-location stores would clarify which features are specific to mobile supermarkets. Beyond expanding the empirical scope, our future work will develop annotation categories that encode the social contextual dimensions identified here, including relational history between participants, spatial configuration of the interactional setting, and sequential position relative to the vehicle’s arrival and departure. We also plan to expand the corpus to cover all four gathering types and to make the annotated data available for linguistic and communication research.

8. Ethical Considerations

The participants were informed of the research objectives and procedures one week before the data collection and provided written consent. Consent withdrawal forms ensured that participation could be terminated voluntarily at any time. The driver confirmed each participant’s willingness before recording. Faces in figures are anonymized, and data will not be shared publicly without additional consent.

9. Acknowledgments

We would like to express our sincere gratitude to the research participants. This work was supported by the JST PRESTO project “Elucidation of Everyday Interactional Support for Older Adults” (JPMJPR2514) and the JSPS KAKENHI project “A Systematic Study of Mobile Supermarket Communication: Linguistic Structure, Interaction, and Social Relationships” (25K16795). We also thank the reviewers for their constructive comments.

10. Bibliographical References

- Goffman, E. (1963). *Behavior in public places: Notes on the social organization of gatherings*. New York: The Free Press.
- Haddington, P., Keisanen, T., Mondada, L., and Nevile, M. (2014). Towards multiactivity as a social and interactional phenomenon. In P. Haddington, T. Keisanen, L. Mondada, and M. Nevile (Eds.), *Multiactivity in Social Interaction: Beyond Multitasking*. John Benjamins, pp. 3--

- 32.
- Harjunpää, K., Mondada, L., & Svinhufvud, K. (2018). The coordinated entry into service encounters in food shops: Managing interactional space, availability, and service during openings. *Research on Language and Social Interaction*, 51(3):271-291.
- Jefferson, G. (2004). Glossary of transcript symbols with an introduction. In G. Lerner (Ed.), *Conversation Analysis: Studies from the First Generation*. John Benjamins, pp. 13–31.
- Kishimoto, K. (2021). The Moment When the Order is fulfilled: How Do Customers Receive the Food They Ordered?. *The Japanese Society for Artificial Intelligence Technical Report, Spoken Language Understanding and Dialogue processing (SIG-SLUD)*, 91: 68–73.
- Koiso, H., Amatani, H., Den, Y., Iseki, Y., Ishimoto, Y., Kashino, W., ... and Watanabe, Y. (2022). Design and evaluation of the corpus of everyday Japanese conversation. In *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC 2022)*, pages 5587–5594.
- Koiso, H., Ishimoto, Y., Iseki, Y., Eguchi, N., Kashino, W., Kawabata, Y., Tanaka, M., Tanaka, Y., and Nishikawa, K. (2024). Overview of the monitor version of CEJC-Child (written in Japanese). In *Proceedings of Language Resources Workshop 2024*, pages 342–248. National Institute for Japanese Language and Linguistics.
- Koiso, H., Tsuchiya, T., Watanabe, R., Yokomori, D., Aizawa, M., and Den, Y. (2016). Survey of conversational behavior: Towards the design of a balanced corpus of everyday Japanese conversation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4434–4439.
- Laurier, E. (2013). Encounters at the counter: The relationship between regulars and staff. In Tolmie, P. & Rouncefield, M. (Eds.), *Ethnomethodology at play*, pages 179–198. Farnham: Ashgate.
- Ministry of Internal Affairs and Communications (2025). *The Current Situation of Depopulation Countermeasures, Fiscal Year 2023*. (written in Japanese). Ministry of Internal Affairs and Communications.
- Mondada, L. (2009). Emergent focused interactions in public places: A systematic analysis of the multimodal achievement of a common interactional space. *Journal of Pragmatics*, 41:1977–1997.
- Mondada, L. (2018). Multiple temporalities of language and body in interaction: Challenges for transcribing multimodality. *Research on Language and Social Interaction*, 51(1):85–106.
- Oshiro, H. (2019). Grocery store and communication place in Osakishimojima Island. *University of Tsukuba Japanese Language and Culture Fieldwork Practicum Report (Domestic Volume)*, 30:27-34.
- Raymond, G. T., and Lerner, G. H. (2014). A body and its involvements: Adjusting action for dual involvements. In *Multiactivity in social interaction*. pages 227-246. John Benjamins Publishing Company.
- Sakai, H. (2024a). Predictability of Encounter and Greeting Practices in Private Grocery Stores. *Japanese Journal of Qualitative Psychology*, 23(special):S112–S118.
- Sakai, H. (2024b). Extending Closings Through See-off Practices in Face-to-Face Apparel Service Encounter. *The Japanese Society for Artificial Intelligence Technical Report, Spoken Language Understanding and Dialogue processing (SIG-SLUD)*, 100: 137–142.
- Sakai, H. & Sakaida, R. (2024). Interaction between the Rural Elderly at a Mobile Supermarket: The Organization of Co-presence during ‘Examining Items’. *The Japanese Society for Artificial Intelligence Technical Report, Spoken Language Understanding and Dialogue processing (SIG-SLUD)*, 101: 102–106.
- Sakai, H. & Sakaida, R. (2025). Research Design Based on Understanding Activities in the Field: Video Recording Targeting Shopping Interaction in a Mobile Supermarket. *The Japanese Journal of Language in Society*. 28(1):250–261.
- Takanashi, K., Maruyama, M., and Aikawa, Y. (2023). Communication analysis of a mobile supermarket in a mountainous area: Aiming rural and local communication studies (Written in Japanese), *Bulletin of the School of Human Cultures, University of Shiga Prefecture*, 54:38–51.

Identifying Linguistically Relevant Communities of Practice on Twitch

Heather Burnett

Laboratoire de Linguistique Formelle (UMR 7110)
CNRS - Université Paris Cité, France
heather.burnett@cncrs.fr

Abstract

This paper argues that, when it comes to modeling language variation and change on the video game streaming platform Twitch, it is necessary to consider “meso-level” communities of practice, i.e. communities of practice that are smaller than the full video game community, yet larger than the usual level of analysis in recent linguistics studies: communities associated with individual Twitch channels. We present a computational method for identifying these linguistically relevant communities of practice and show how this method can be useful for analyzing quantitative patterns of sociolinguistic variation in a corpus composed of the chat transcripts of 15 streamers of the game *Elden Ring: Nightreign*.

Keywords: communities of practice, video game language, Twitch, clustering, corpus linguistics

1. Introduction

This paper presents a method for identifying linguistically relevant *communities of practice* (CofP) online, with a specific focus on the video game streaming platform *Twitch.tv*. By linguistically relevant CofPs, we mean the kind of social structures that are appropriate for understanding how language varies and changes according to the social context.

How the use of socially meaningful linguistic form varies, and how new forms emerge, are major research questions in sociolinguistics. From the early days of quantitative (variationist) sociolinguistics (see Labov (1973)), researchers observed that the use of sociolinguistic variables (grammatical alternations, for example, pronouncing the final consonant in the word *working* as [ŋ] vs [n]) are conditioned by aspects of speakers’ places in their speech community. In the “first wave” of studies of linguistic variation, the “speech community” was taken to be a whole city (New York (Labov, 1966), Montréal (Sankoff and Cedergren, 1972), Detroit (Wolfram, 1969) etc.) and a speaker’s place in this community was established through their demographic properties: usually age, gender, social class and race (see Eckert (2012) for a description of the different “waves” of quantitative sociolinguistic study). However, later more anthropologically oriented studies showed that many sociolinguistic variables do not break down according to demographic properties, and their use is better analyzed as resulting from speakers’ places in smaller, locally defined social structures: communities of practice. Introduced into sociolinguistics from education research by Eckert and McConnell-Ginet (1992), a community of practice is “is an aggregate of people

who come together around mutual engagement in an endeavor. [...] As a social construct, a community of practice is different from the traditional community, primarily because it is defined simultaneously by its membership and by the practice in which that membership engages” (Eckert and McConnell-Ginet, 1992, 464). Eckert (2000)’s foundational work on language variation as a social practice shows how CofPs can be helpful for understanding the distribution of both phonological sociolinguistic variables (vowel pronunciations) and syntactic ones (negation and negative concord) in a Detroit High School.

In the past 30 years, CofPs have become one of the main units of analysis within both quantitative and qualitative sociolinguistics, studying language a wide variety of social settings. In an influential review of the concept, Meyerhoff (2013), building on Wenger (1998), describes three properties defining CofPs. The first is mutual engagement; that is, “the members of a CofP need to get together in order to engage in their shared practices” (Meyerhoff, 2013, 527). The second is that members share some jointly negotiated enterprise: “members get together for some purpose and this purpose is defined through their pursuit of it” (Meyerhoff, 2013, 528). The final defining criterion of a CofP is that members have a shared repertoire, and “these resources (linguistic or otherwise) are the cumulative result of internal negotiations” (Meyerhoff, 2013, 528). It is through this notion of “negotiated shared repertoire” that CofPs have been found to be especially useful for studying certain kinds of language variation and change in the offline world (see, for example, Bucholtz (1999); Holmes et al. (1999), among others).

Communities of practice have also been identi-

fied as relevant for studying language online. Already a decade ago, [Angouri \(2015\)](#) summarized research on digital communication using CofPs; however, one of the most detailed discussions of whether/how this concept can be fruitfully applied to language variation and change in an online setting is [Leuckert and Leuckert \(2020\)](#). In this paper, which focuses on the news and social media platform Reddit, the authors argue that, although there are some differences with offline CofPs, communities associated with individual Reddit forums, known as *subreddits*, generally meet the criteria laid out above for being considered communities of practice. Through an in depth study of three subreddits: *r/LEAGUEOFLEGENDS* (a forum devoted to the video game *League of Legends* (Riot, 2009)); *r/LINGUISTICS* (an academic forum devoted talking about language); and *r/RUPAULSDRAGRACE* (a forum devoted to discussing the TV show *RuPaul's Drag Race* (LogoTV, 2019), [Leuckert & Leuckert](#) argue that participants in these subreddits display mutual engagement, participate in a jointly negotiated enterprise and have a shared repertoire. For example, users on the *r/RUPAULSDRAGRACE* forum use linguistic expressions like *RuSponsibly* and *ruveal*, which are not found in the other subreddits studied.

[Leuckert and Leuckert \(2020\)](#)'s claim that individual subreddits can be communities of practice is both intuitive and well argued. It is also mirrored in other work that studies language (and other kinds of behavior) on a platform with a similar structure: *Twitch.tv*. *Twitch.tv* is a video streaming platform, which pairs live video of an individual broadcasting an activity (called the *streamer*), usually playing a video game, with a synchronous live chat, where audience members (*viewers*) can send public messages to be consumed by both the streamer and the other viewers. Similar to the way in which Reddit is composed of subreddits/forums, Twitch is made up of channels associated with individual streamers. Given that these channels and their viewers are named "communities", and that they are often very clearly sites of mutual engagement (learning more about a video game, fostering a sense of community, etc. ([Hamilton et al., 2014](#); [Gros et al., 2018](#))), analyzing Twitch channels as CofPs is a common when it comes to studying linguistic behavior on the platform, see [Graham \(2019\)](#); [van der Aa \(2021\)](#) among others.

Our goal in this paper is to argue that, while individual Twitch channels undeniably show some of the properties of CofPs, when it comes to modeling language variation and change on the platform, the "micro" level, i.e. channels, is sometimes too fine-grained to properly capture the **shared repertoires** that members of online communities have. To illustrate this point, we present a new quantitative study

of lexical variation in a corpus composed of chat transcripts from 15 streamers of the video game *Elden Ring: Nightreign* (2025, FromSoftware) on Twitch. We show that accurately capturing the distributions of these sociolinguistic variables in the corpus requires a level of analysis that is intermediary between the whole "Elden Ring" community (exemplified by all 15 streamers) and the chats of each individual streamer. Building on recent work applying similarity and cluster analyses to linguistic data, we provide a method for identifying "meso-level" communities of practice that, we show, allow for a better understanding of how new linguistic forms emerge, are used, and change in gaming communities.

2. The Nightreign Corpus

When writing an article studying gaming communities (on Twitch or elsewhere) it is common to start by speaking of a "macro-level" community: *The [Name of Game] Community* (see, for example, [Cochran et al. \(2023\)](#) for *The League of Legends Community*; [Marlatt \(2020\)](#) for *The Fortnite Community*, etc.). This paper will be no exception: its focus will be on the community often described as the *Souls* community, the *Soulsborne* community or the *Elden Ring Community*. These definite descriptions are often used by players who are fans of the series of games developed by the Japanese studio FromSoftware (*Demon's Souls*, *Dark Souls 1*, *Dark Souls 2*, *Dark Souls 3*, *Bloodborne*, *Sekiro*, *Elden Ring*, *Elden Ring: Nightreign*) and games with similar structures. The fact that players often use the singular definite description to refer to the *Soulsborne* community (or similar) suggests that they imagine themselves as being part of a single unified group¹ ([Anderson, 1983](#)), but, in the details, it's not so clear who exactly belongs to this "community". For example, 2022's *Elden Ring* has sold over 30 million copies², but it is unlikely that anyone who uses the term "the *Elden Ring* community" would consider it as having more than 30 million members. To operationalize this macro-level notion of *The Elden Ring Community* in a corpus study, we therefore decided to anchor our study around a particular group of streamers: those who took part in the fourth season of *Elden Ring Bingo* in 2024 (<https://bingobrawlers.com/>). *Elden Ring Bingo* is a competitive game that pits two players against each other to see how fast they can complete a series of objectives in the main single player *Elden Ring* game (completing lines on a

¹See for example https://www.reddit.com/r/Eldenring/comments/pa3iwn/do_you_think_the_soulsborne_community_is_toxic/

²<https://www.thegamer.com/elden-ring-sells-30-million-copies-switch-2/>

Streamer	Followers	Nationality
LilAggy	262k	USA
Ginomachino	162k	Canada
Parkenharbor	144k	Canada
Bushy	98k	USA
Captain_domo	47k	USA/Germany
ItzCBD	47k	USA
star0chris	41k	USA
Vswed	34k	Canada
Adef	32k	USA
Mitchriz	31k	USA
Blanxz	18k	USA
Zoodle	18k	Germany
nuclearpastatom	14k	USA
yojoshertino	14k	USA
DrDoot	7k	USA

Table 1: The 15 streamers featured in the Nightreign corpus (follower count from July 2025)

bingo board). Additionally, most of these streamers perform Elden Ring themed events at the very popular GDQ (*Games Done Quick*) speedrunning events (Blanxz 2024, Mitchriz 2023, 2024, 2025, Captain_domo 2024, 2025, LilAggy 2024, 2025, star0chris 2025, yojoshertino 2025, adef 2025, DrDoot 2025). This makes them identifiable as representatives of the (Western) *Elden Ring Community* to players of other games.

Our corpus is composed of the chat transcripts from the streamers who both participated in the tournament and streamed the newest FromSoftware game, *Elden Ring: Nightreign*, in summer 2025 (see Table 1). All streamers stream in English and enforce an “English only” policy in their chats, regardless of their native language.

More specifically, the *Elden Ring: Nightreign Twitch Chat* corpus is composed of transcripts from all the chats during the period in which *Elden Ring: Nightreign* was actively streamed on Twitch: May 30 (launch of the game) - August 31 2025. The individual chat files were downloaded as .csv files using the *Twitch Chat Downloader* application (<https://www.twitchchatdownloader.com/>). In all, the corpus is composed of the chats from 448 streams, and contains productions from 50 130 users (henceforth known as *chatters*). The *Elden Ring: Nightreign Twitch Chat* corpus is available at https://osf.io/ct3zr/overview?view_only=395154716b8c4d5c9cbb955755a99b8c.

3. Ludolexical variation

One of the best-known features of gaming communities, which is also one of the defining characteristics of communities of practice, is that they often give rise to unique shared linguistic norms.

For gamers, this usually involves having a specific vocabulary for game-play (ludic) elements in the game (lorio, 2010; Ensslin, 2012; Zähres, 2019, among many others). The “Elden Ring Community” is no exception, with community invented terms being widely used, especially when it comes to referring to enemies in the game. For example, consider the message in (1), sent by a user in Zoodle’s chat.

- (1) I love Elden Ring lingo so much, we got nights cav, BBH, BBK, DTS, RadaBeast, Gargs and so much more (chatter in Zoodle stream 2498943069)

As this chatter illustrates, “Elden Ring lingo” involves linguistic practices such as referring to the enemy *Night’s Cavalry* as *nights cav* (employing a morphological process known as clipping), *Bell Bearing Hunter* as *BBH* (acronymization), *Black-Blade Kindred* as *BBK*, *Draconic Tree Sentinel* as *DTS* (also acronymization), *Valiant Gargoyles* as *gargs* (clipping) etc.

In this short paper, we present two examples of variation in “Elden Ring lingo” which, we argue, illustrate our main point: some linguistically relevant communities of practice (at least on Twitch) lie between the macro-level, the *Elden Ring Community*, in this case instantiated by all 15 streamers, and the micro-level, i.e. communities around individual streamers’ channels. More specifically, we will study variation in the naming of the most recent Elden Ring game (*Nightreign*) and references to one of the main antagonists in the game: the *Bell Bearing Hunter*.

3.1. Nightreign

We extracted all references to the game *Nightreign* from the *Elden Ring: Nightreign Twitch Chat* corpus. Through searching with the regular expression *night+*, we extracted all the uses (with various spellings) of the official name of the game: *Nightreign* (2). From inspecting the contexts surrounding the occurrences of *Nightreign* (chat messages shortly before and after the original message), we were able to identify two other variants: the acronyms *NR* (3) and *NTR* (4).

- (2) Hitless **nightreign** runs are about to be crazy (chatter in Blanxz stream 2471884069)
- (3) Chat did **NR** get a new patch today?? (chatter in LilAggy’s stream 2482675698)
- (4) aggy is the best **NTR** bingo player for sure (chatter in LilAggy’s stream 2538401141)

Table 2 shows the distribution of these three variants in the whole corpus. The official name of the game is the most frequent variant, appearing in almost 92% of the cases. However, there is variation:

the acronym *NR* is used around 7% of the time, and the minority variant, *NTR*, is used less than 1% of the time.

Variant	Occurrences
Nightreign	7112 (91.9%)
NR	567 (7.3%)
NTR	59 (0.7%)

Table 2: Variation referring to *Elden Ring: Nightreign*

The appearance of acronyms in a corpus of chat between gamers is not surprising: research on language and video games has shown repeatedly that shortening morphological processes, especially acronymization, is very common (Ensslin, 2012; Záhres, 2019; Jannah and Tawami, 2024, among many others). However, what is less expected, given previous literature, is the existence of two acronym variants: *NR* and *NTR*. A natural hypothesis to explain this state of affairs is that *NTR* is favored by members of a particular streamer’s community, as a way of marking the “in group” status, something that has also been argued to be frequent in gaming communities (Ensslin, 2012; Záhres, 2019). However, if we break the data down by individual streamers’ chats, we see that this is not so clearly the case. Figure 1 shows the raw counts of each variant and Figure 2 shows the percentage of the total distribution of the variant, for the chats of each of the 15 streamers.

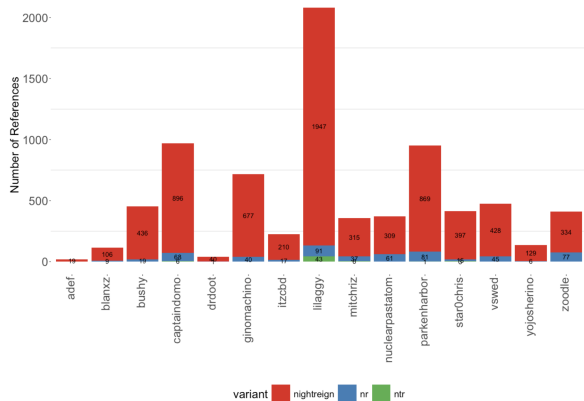


Figure 1: Nightreign vs NR vs NTR by streamer (counts)

This figure shows that *NTR* is used by chatters who watch a number of streamers: Captain_domo, LilAggy, Mitchriz, parkenharbor and star0chris. So a question arises as to whether these five streamers’ chats have something in common that could explain the emergence of the *NTR* variant.

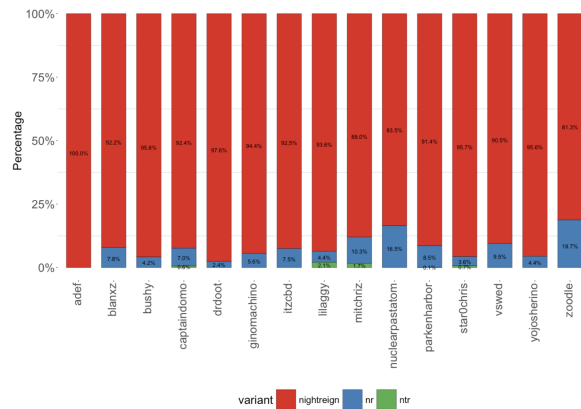


Figure 2: Nightreign vs NR vs NTR by streamer (percentages)

3.2. Bell Bearing Hunter

Our second example is reference to the enemy officially known as the *Bell Bearing Hunter*. Using regular expressions, we extracted all occurrences of use of the variant *bell bearing hunter* from the corpus (5), as well as elliptical versions, such as *bell bearing* (6) and *bell hunter* (7). Inspecting the context around occurrences of *bell bearing hunter*, we further identified other variants³: the acronym *BBH* (8), see also (1), and the expression *Buh Buh Huh*, which orthographically represents the pronunciation of the acronym (9).

- (5) **Bell Bearing Hunter** is the true final boss of this game
(chatter in LilAggy’s stream 2478131530)
- (6) If I was to choose a seed for my runs, I would avoid the wolf and **bell bearing** kkkkk my fights are always longer than I want
(chatter in nuclearpastatom’s stream 2514398400)
- (7) the **bell hunter** SUCKS
(chatter in LilAggy’s stream 2477320404)
- (8) You know, thinking about it, **BBH** would be a fun playable character in this with all his Darth Vader moves.
(chatter in itzCBD’s stream 2516397731)
- (9) **buh buh HUH** isn’t a bad fight solo
(chatter in Captain_domo’s stream 2482368557)

In the corpus, the most frequent variants are the official one (*Bell Bearing Hunter*), the acronym

³We also identified a whole range of variants that occur only once or twice (eg. *Barry the bell bearing meanie*, *Bell bearing Kunter*, *Chad bearing hunter* among others). These variants, although humorous, are unfortunately too rare to include in a quantitative analysis.

BBH and the other “acronym” *BUH BUH HUH*. The distribution of these variants is shown in Table 3.

Variant	Occurrences
Bell Bearing Hunter	289 (25.3%)
BBH	568 (49.8%)
BUH BUH HUH	108 (9.4%)
Other variants	176 (15.4%)

Table 3: Variation referring to the Bell Bearing Hunter

Table 3 shows that the acronym *BBH* is the most frequent (around 50% of occurrences), even more than the official name, which is used in only a quarter of references to the enemy in the corpus. The ‘pronounced’ acronym *BUH BUH HUH* is used in almost 10% of the cases.

As with the references to *Nightreign*, looking at the data at the level of individual streamers’ chats is not very helpful. As Figures 3 and 4 shows, no variant is limited to the community of a single streamer.

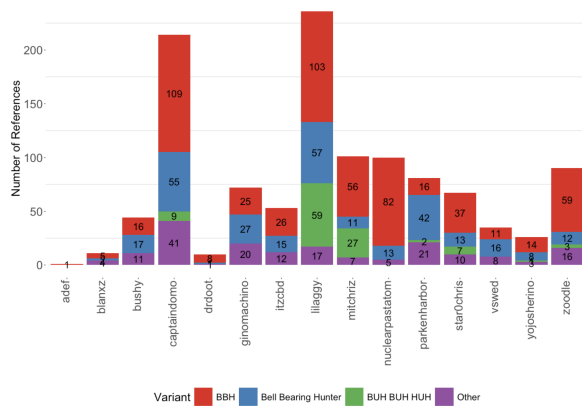


Figure 3: Bell Bearing Hunter vs BBH vs BUH BUH HUH by streamer (counts)

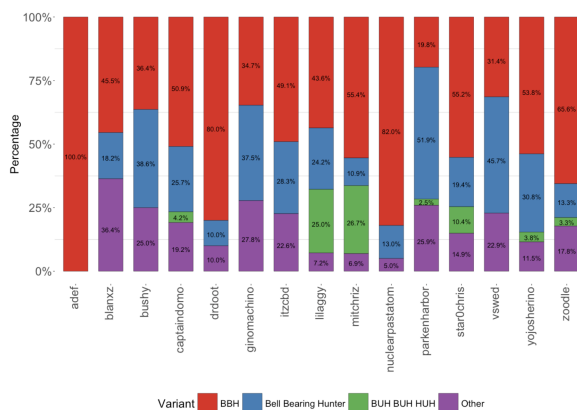


Figure 4: Bell Bearing Hunter vs BBH vs BUH BUH HUH by streamer (percentages)

However, in Figure 4, we can notice that the group of streamers whose chatters use *BUH BUH HUH* is quite similar to the chats where *NTR* is found. Both of these minority variants are found in the chats of LilAggy, Captain_domo, star0chris, parkenharbor and Mitchriz. In the next section, we will lay out a method for identifying more precisely what these streamers have in common.

4. Identifying meso-level communities on Twitch

We argue that the problem for studying linguistic variation with treating each streamer’s community as a separate community of practice is that there is significant overlap in the audiences of the streamers. The streamers are all professional Elden Ring players who often collaborate with one another (for example, in the 2024 Bingo Brawlers tournament), so it is not surprising that a viewer who enjoys Elden Ring gameplay would view more than one streamer’s channel. However, chatters’ attentions are not spread out equally across all 15 streamers. For example, in the corpus, there are 3150 unique chatters that post in both LilAggy and Captain_domo’s chats; however, there are only 709 chatters who contribute both to LilAggy and itzCBD’s chat. This is despite Captain_domo and itzCBD having roughly the same number of followers, as shown in Table 1.

We can be more precise about how similar or different the chatting audience of the 15 streamers are through the use of the *Jaccard coefficient*, a similarity measure widely used in machine learning and various areas of quantitative and computational linguistics (see Tan et al. (2006)). For pairs of streamers, we calculated their Jaccard coefficient through taking the cardinality of the intersection of the sets of unique chatters for each streamer, and dividing this cardinality by the cardinality of the union of the two sets of unique chatters. In this way, Jaccard coefficients range from 0 (sets of chatters are completely disjoint) and 1 (sets of chatters are identical)⁴.

Figure 5 shows a heatmap displaying the Jaccard coefficients measuring the chatter overlaps for each pair of streamers in the corpus. In line with the discussion above, we can observe that LilAggy and Captain_domo have a larger Jaccard coefficient than LilAggy and itzCBD (0.13 vs 0.03).

Figure 5 shows that the overlap between streamers’ participating audiences is not enormous: the highest Jaccard coefficient is 0.24. Nevertheless, 0.24 does indicate that, for some pairs of streamers,

⁴The calculation of the Jaccard coefficients was done in R, and the code is available at https://osf.io/ct3zr/overview?view_only=395154716b8c4d5c9cbb955755a99b8c.

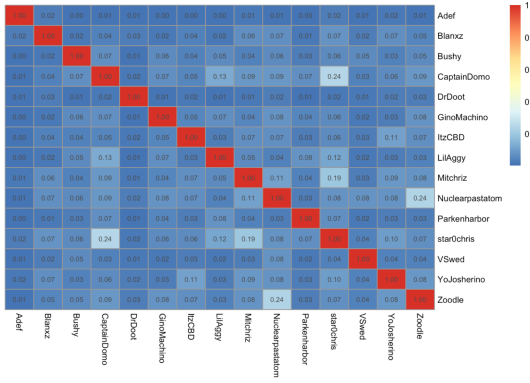


Figure 5: Heatmap of Jaccard coefficients for 15 streamers

a non-trivial part of their participating audience is shared. We would therefore like some way to group multiple streamers together based on this shared audience, to then see whether these new communities of practice are better predictors of linguistic variation.

For this purpose, we employed one of the most common clustering algorithms used in linguistics (see Levshina (2015)): *k*-means clustering (MacQueen, 1967). *k*-means clustering aims to partition (in this case) 15 streamers into *k* groups in a way that maximizes (in this case) Jaccard similarities between members of groups and minimizes similarities between members of different groups. The *k*-means clustering presented in this paper was done in R (with the *cluster* package (Maechler et al., 2026)), and the visualizations were also done in R (with the *factoextra* package (Kassambara and Mundt, 2020)).

Since we are using clustering as an exploratory technique, the difficult question of how many clusters are desired arises. As discussed above, all the Jaccard coefficients are quite similar, so we know that there are no isolated close-knit sub-communities within the group of 15 streamers being studied. This fact is reflected in the clustering results: Figure 6 plots the total within sum of squares value for results of the *k*-means algorithm (calculated using the *fviz_nbclust*(\cdot) function from the *factoextra* package in R), for 1 to 5 clusters. In this figure, there is no “elbow”, i.e. place where adding an extra cluster improves the fit of the model less than any other, at least for few numbers of clusters.

Nevertheless, we see that the algorithm consistently distinguishes between streamers like star0chris, Captain_domo and Mitchriz on the one hand, and nuclearpastatom and Zoodle on the other, as shown by the 2 cluster and 3 cluster visualizations in Figures 7 and 8.

In the next section, we will show that, even if we take a coarse-grained clustering, such as $k = 3$, we can arrive at a better model of ludolexical variation

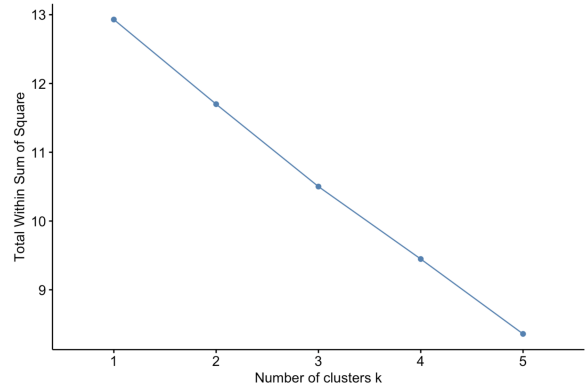


Figure 6: Total sum of squares for $k = 1$ to $k = 5$

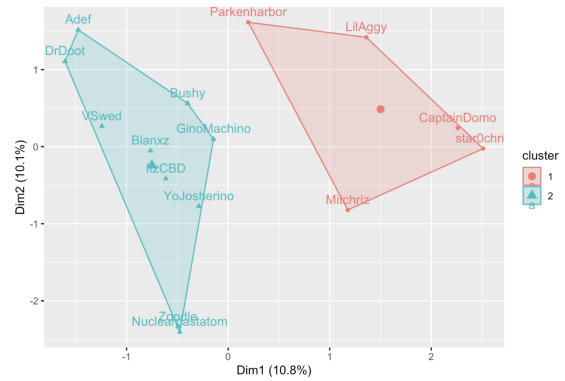


Figure 7: *k*-means clustering, for $k = 2$

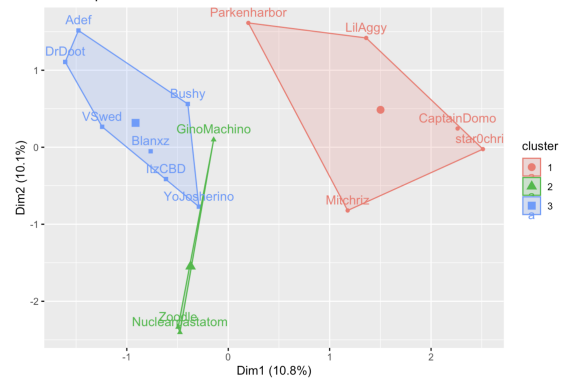


Figure 8: *k*-means clustering, for $k = 3$

than if we simply consider either the whole community (15 channels) or each streamer’s channel independently.

5. Back to ludolexical variation

As shown in Figure 8, the three groups of streamers that were identified based on overlapping chatters are **Cluster A**: Parkenharbor, LilAggy, Captain_domo, star0chris and Mitchriz; **Cluster B**: Adef, Bushy, DrDoot, Vswed, Blanxz, itzCBD and YoJasherino; and **Cluster C**: Zoodle, nuclearpas-

tatom and Ginomachino.

Figure 9 shows the distribution of the variants *Nightreign*, *NR* and *NTR* to refer to the game by cluster. We can now observe that all the occurrences of *NTR* are found within the chats of streamers from cluster A.

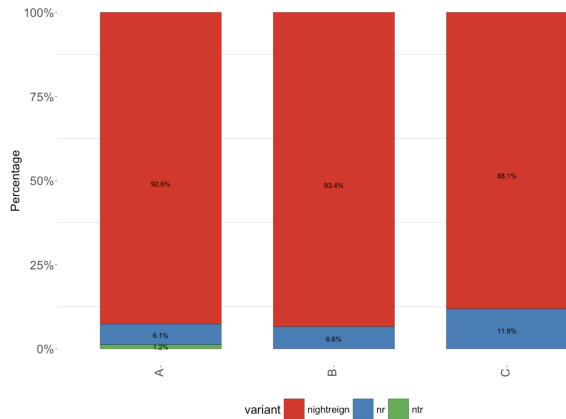


Figure 9: Nightreign vs NR vs NTR by cluster

We have some evidence that this pattern is driven at least in part by individuals who participate in multiple streams. For example, the first use of *NTR* in our corpus comes from a LilAggy stream, where a particular chatter, call them *Chatter A*, writes *NTR* as a variant for *NR* (10). The same chatter than participates in Captain_domo's stream a couple days later, using the new *NTR* variant (11). As example (11) shows, the innovative variant is remarked on by other Captain_domo chatters, who draw attention to the fact that *NTR* also happens to be an acronym for a genre of manga, *Netorare*, which is centered around infidelity and the dark feelings that ensue.

- (10) idk about you guys, but ER **NTR** is at least and 8/10 for me
(Chatter A in LilAggy's stream 2482675698)
- (11) **Chatter A:** I'm playing **NTR** on the other monitor and the double audio is really confusing
Chatter B: @[Chatter A] i wouldn't abbreviate nightreign to **ntr** KEKW
Chatter C: @[Chatter A] HANK, DONT ABBREVIATE NIGHTREIGN. HAAANK
(Captain_domo's stream 2492160644)

The *NTR* variant is then spread to Parkenhabor's chat, when Parkenharbor audience members observe LilAggy chatters using it, during a *Nightreign* collaboration a few days later (12)⁵.

⁵We note here that clearly metalinguistic uses of *NTR* and *Nightreign* were excluded from the quantitative study. Thus, the data in Table 2 only counts actual references to the video game, not to the expression itself or to *Netorare*.

- (12) why the fuck are they talking about **NTR** over there
(chatter in Parkenharbor's stream 2530353580)

In this way, we see the development of a linguistically relevant "meso-level" community of practice centered around Elden Ring gameplay, i.e. communities that are the result of (relatively loose) contact between smaller Twitch communities.

We find a similar pattern when it comes to referring to the Bell Bearing Hunter: as Figure 10 shows, *BUH BUH HUH* is significantly⁶ more frequent in chats of streamers of the A cluster.

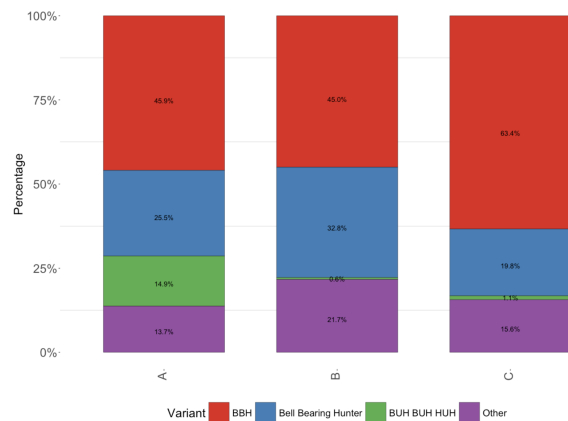


Figure 10: Bell Bearing Hunter vs BBH vs BUH BUH HUH by cluster

And it appears that a similar mechanism to the spread of *NTR* drives this pattern: individuals from one chat will bring variants they learn from that chat to another chat, which then starts using them. For example, a chatter in LilAggy's stream, call them *Chatter D*, uses the *BUH BUH HUH* variant when the Bell Bearing Hunter comes on screen. As example (13) shows, this is a common practice in LilAggy's stream. Chatter D then goes to Mitchriz's stream and informs Mitchriz's audience of the "correct pronunciation" of *BBH*.

- (13) Chatter P: Buhh Buhh HUHH
Chatter Q: super tanky boi
Chatter D: Buhh Buhh HUHH
Chatter R: Buhh Buhh HUHH
(LilAggy stream 2501295308)
- (14) it's pronounced Buh Buh HUH
(Chatter D, in Mitchriz stream 2518225917)

⁶In statistical analysis (generalized linear mixed effects models, using the *lme4* R package (Bates et al., 2015), predicting the use of the *BUH BUH HUH* variant vs another variant, with cluster as fixed effect and streamID as random effect) we find that *BUH BUH HUH* is used significantly less in chats of cluster B ($p < 0.002$) and significantly less in chats of cluster C ($p < 0.001$), compared to the chats in cluster A.

With these two examples, we can see the development of a shared repertoire of linguistic practices (referring to Nightreign as *NTR* and the Bell Bearing Hunter as *BUH BUH HUH*) across communities associated with multiple Twitch channels : those that have many chatters in common.

6. Conclusion

This paper argued that, when it comes to modeling language variation and change on the video game streaming platform Twitch, it is necessary to consider “meso-level” communities of practice, i.e. ones that are smaller than the full video game community, yet larger than communities associated with individual channels. We provided a method, Jaccard coefficients + k-means clustering, which, we argued, could help identify these intermediary communities. We illustrated our proposals through a study of quantitative ludolexical variation in a corpus of chat transcripts from 15 streamers of “The Elden Ring Community”. We showed that lexical innovations that are created in the community associated with a channel can be quickly propagated to the communities of other channels when overlap between the two communities is high, or when the two communities come into contact (for example through streamer collaborations). Our results thus provide new linguistic evidence for what Cochran et al. (2023) calls “nested communities of practice”, CofPs that exist within a larger gaming community of practice, yet may also contain other CofPs.

7. Limitations and ethics statement

The results in this paper raise a number of questions that we are currently unfortunately unable to answer. For example, we focused on the linguistic productions of chatters, neglecting those of the streamers. The reason for this focus was practicality: the chat transcripts were easily obtained through an application; whereas, obtaining a transcript of the streamers’ linguistic productions would be extremely difficult and time consuming, since the stream videos are often 5-10 hours long and automatic transcription of technical gameplay language is mediocre. However, the streamers’ language is clearly relevant to the development of shared linguistic repertoires, so lack of information about the streamers’ behavior is a clear limitation of this study.

Another question raised by our results concerns the precise method for discovering intermediary CofPs. As Figure 6 shows, *k*-means clustering for $k < 6$ does not provide a good fit for our data. Even though we argued that this method was sufficient to shed light on linguistic variation in our corpus, we might wonder whether other methods of identifying

subgroups of streamers or chatters would yield better results. We therefore leave the exploration of this possibility to future research.

Ethics statement: The study reported in this paper used only publicly available data from individuals who, at the time of their participation in public Twitch streams, gave their consent to have their productions recorded and broadcast. No artificial intelligence tools were used at any point in the data, analysis or writing of this paper.

8. Acknowledgments

The author would like to thank Olivier Bonami and two SoCon reviewers for helpful comments and suggestions. This work received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (Grant agreement No. 850539).

9. Bibliographical References

- Benjamin Anderson. 1983. *Imagined Communities: Reflections on the Origins and Spread of Nationalism*. Verso.
- Jo Angouri. 2015. Online communities and communities of practice. In *The Routledge handbook of language and digital communication*, pages 323–338. Routledge.
- Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. [Fitting linear mixed-effects models using lme4](#). *Journal of Statistical Software*, 67(1):1–48.
- Mary Bucholtz. 1999. “why be normal?”: Language and identity practices in a community of nerd girls. *Language in society*, 28(2):203–223.
- Andrew F Cochran, Stephen T Slota, and Michael F Young. 2023. League of legends. *The Literacies of the Esports Ecosystem*, 5:93.
- Penelope Eckert. 2000. *Linguistic Variation as Social Practice*. Blackwell, Oxford.
- Penelope Eckert. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, 41(1):87–100.
- Penelope Eckert and Sally McConnell-Ginet. 1992. Think practically and look locally: Language and gender as community-based practice. *Annual review of anthropology*, pages 461–490.
- Astrid Ensslin. 2012. *The language of gaming*. Bloomsbury Publishing.

- Sage L Graham. 2019. A wink and a nod: The role of emojis in forming digital communities. *Multilingua*, 38(4):377–400.
- Daniel Gros, Anna Hackenholt, Piotr Zawadzki, and Brigitta Wanner. 2018. Interactions of twitch users and their usage behavior. In *International Conference on Social Computing and Social Media*, pages 201–213. Springer.
- William A Hamilton, Oliver Garretson, and Andruid Kerne. 2014. Streaming on Twitch: fostering participatory communities of play within live mixed media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1315–1324.
- Janet Holmes, Maria Stubbe, and Bernadette Vine. 1999. Constructing professional identity: “doing power” in policy units. *Talk, work and institutional order: Discourse in medical, mediation and management settings*, 1:351–385.
- Joshua Boyd Iorio. 2010. *Explaining orthographic variation in a virtual community: Linguistic, social, and contextual factors*. Ph.D. thesis, University of Texas at Austin.
- Salwa Raudhatul Jannah and Tatan Tawami. 2024. Morphological shortening in mobile legend video game. *Mahadaya: Jurnal Bahasa, Sastra, dan Budaya*, 4(2):271–288.
- Alboukadel Kassambara and Fabian Mundt. 2020. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.7.999.
- William Labov. 1966. *The social stratification of English in New York city*. Cambridge University Press.
- William Labov. 1973. *Sociolinguistic patterns*. 4. University of Pennsylvania press.
- Sven Leuckert and Martin Leuckert. 2020. Towards a digital sociolinguistics: Communities of practice on reddit. In *Corpus approaches to social media*, pages 15–40. John Benjamins Publishing Company.
- Natalia Levshina. 2015. *How to do linguistics with R*. John Benjamins Publishing Company.
- J MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297.
- Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. 2026. *cluster: Cluster Analysis Basics and Extensions*. R package version 2.1.8.2 — For new features, see the ‘NEWS’ and the ‘Changelog’ file in the package source).
- Rick Marlatt. 2020. Capitalizing on the craze of fortnite: Toward a conceptual framework for understanding how gamers construct communities of practice. *Journal of Education*, 200(1):3–11.
- Miriam Meyerhoff. 2013. Communities of practice. *The handbook of language variation and change*, pages 428–447.
- Gillian Sankoff and Henrietta Cedergren. 1972. Sociolinguistic research on french in montreal. *Language in Society*, 1(1):173–174.
- Pang-Ning Tan, Michael Steinbach, Vipin Kumar, et al. 2006. Cluster analysis: basic concepts and algorithms. *Introduction to data mining*, 8:487–568.
- JTA van der Aa. 2021. “it’sa first!”: A case study on identity expression through language on twitch in a game of teamfight tactics. B.S. thesis, Utrecht University.
- Etienne Wenger. 1998. *Communities of practice: Learning, meaning, and identity*. Cambridge University Press.
- Walt Wolfram. 1969. *Black/White Speech Differences Revisited: A Preliminary Report*. Center for Applied Linguistics, Washington DC.
- Frederic Záhres. 2019. “pentakill inc.”-a case study of shortening processes in ludolects. In *Communication forms and communicative practices. New perspectives on communication forms, affordances and what users make of them*, volume 15. Peter Lang.

Implicit Cultural Identity Signals in Language: Detection and Effects in Negotiation Dialogue

Bin Han, Danah Yun, James Hale, Jonathan Gratch

University of Southern California

Los Angeles, CA, USA

{binhan, dnyun, jahale}@usc.edu, gratch@ict.usc.edu

Abstract

Language transmits cultural identity even without explicit disclosure, shaping how individuals perceive and engage in interpersonal tasks. We investigate these identity signals within the KODIS (KObe DISpute) corpus, an English-only corpus of anonymous text-chat negotiations involving participants from the US, UK, Mexico, and South Africa. We test whether a speaker's country can be inferred from dialogue using zero-shot LLMs and embedding-based classifiers. Our results demonstrate that cultural identity is reliably detectable in text, with embedding models achieving far higher accuracy. Critically, while objective negotiation outcomes remain consistent across groups, these subtle identity cues significantly alter participants' subjective feelings about the interaction. These findings suggest that cultural identity-related signals are embedded in language and may be relevant for analyzing negotiation dialogue.

Keywords: Cultural Identity, Sociolinguistic Cues, Social Interaction, Person Perception

1. Introduction

Language use can convey cues about a speaker's background and social identity (Gumperz, 1982). Word choice and expression shape impressions about a person's background and social identity, and these impressions influence how interaction unfolds (Koschate et al., 2021). In social psychology, this process is known as *Person Perception* (Keith, 2013). Interaction is influenced not only by what is said but also by how the speaker is perceived (Kunda and Thagard, 1996). In natural dialogue, speakers rarely state their cultural background directly. Even so, patterns of language use may convey subtle identity-related signals.

Identifying these signals from text is a meaningful research goal, not just a step toward downstream analysis. In many natural dialogue datasets, especially those from online disputes or archived negotiations, cultural background information is missing or unreliable. A classifier that infers cultural origin from language makes culture observable and measurable in large unlabeled corpora where metadata is not available (Tetreault et al., 2013). Negotiation is a useful context to study this problem. Prior dialect identification work mostly focuses on neutral or everyday language, but negotiation dialogue is different. It involves strategic interaction, power differences, and emotionally charged face threats (Lawler and Yoon, 1995). These conditions may strengthen or weaken cultural signals, but this has not been tested empirically.

This paper examines whether traces of cultural identity emerge in task-based dialogue when speakers do not explicitly disclose their background. It also examines whether these signals reflect

surface-level differences in expression or are associated with differences in interaction outcomes. To explore this, we analyze English negotiation dialogues collected under the same task conditions from participants in four countries (Hale et al., 2025). The analysis focuses on whether a speaker's country can be inferred from dialogue alone and whether the strength of this inference is related to negotiation behavior.

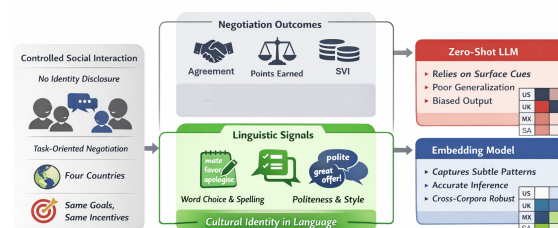


Figure 1: Overview of the study design and evaluation framework.

Specifically, we address the following research questions:

- **RQ1:** To what extent can a speaker's country of origin be inferred from dialogue when cultural identity is neither relevant nor intentionally disclosed?
- **RQ2:** Which computational approach better captures these signals in negotiation dialogue: zero-shot LLMs or embedding-based classifiers?
- **RQ3:** Are these inferred identity signals associated with differences in negotiation outcomes, as reflected in within-country versus between-country interactions?

2. Related work

2.1. Cultural identity in language

Sociolinguistic research has shown that social and cultural identity can be reflected in patterns of language use, such as lexical choice, spelling conventions, and pragmatic style (Bucholtz and Hall, 2005; Labov, 2006). These patterns often emerge without explicit self-disclosure and can be observed across different communicative contexts (Bucholtz and Hall, 2005). Prior work suggests that listeners may form social impressions based on such linguistic cues, even when identity is not directly relevant to the task (Giles and Powesland, 1975).

2.2. NLP for social variable detection

Prior work in NLP has examined the prediction of social variables from text, including demographic and cultural attributes (Tetreault et al., 2013). These studies are motivated by the observation that such social variables are systematically reflected in linguistic patterns, and have proposed a range of computational approaches, most commonly using feature-based representations of text (Zampieri et al., 2018), with more recent work applying large language models for social attribute prediction (Al-Nuaimi et al., 2024). However, these studies have largely focused on static text genres such as essays or social media posts, overlooking interactive task-oriented dialogue. Negotiation is a particularly compelling yet underexplored setting: it is fundamentally a social interaction shaped by perceptions of group membership and identity (Tajfel et al., 2001), where in-group favoritism and inter-group bias can influence communication style and subjective evaluations (Brewer, 1999; De Dreu and Carnevale, 2003). However, the application of NLP methods to negotiation dialogue remains largely unexplored, even though strategic pressure and emotional intensity may influence how cultural identity is expressed.

3. Method

3.1. KODIS Dataset

We use English negotiation dialogues from the KODIS dataset. The dialogues are collected from an online negotiation experiment in which participants interact anonymously to resolve a purchase dispute (Hale et al., 2025). Participants do not know each other’s country of origin, and there is no requirement or incentive to reveal it during the interaction. All dialogues follow the same task scenario: a buyer requests a refund after receiving an incorrect basketball jersey from a seller. The scenario is designed to elicit emotionally charged exchanges,

as the buyer has been wronged and may express frustration or anger during the dispute.

Each dialogue is associated with a buyer country label among four classes: *U.S.*, *U.K.*, *Mexico*, and *South Africa*. We construct a balanced dataset with 80 dialogues per country (320 total), where both buyer and seller are from the same country.

3.2. Models

We analyze only the buyer side of each dialogue (speaker-specific setting). We use 5-fold stratified cross-validation with an 80% train and 20% test split in each fold.

- **LLM (Zero-shot).** We use `gpt-4o-mini` model as a zero-shot classifier. The prompt provides a short list of linguistically grounded cues (e.g., spelling variants, lexical choice, delivery terminology, and interactional tone) to guide the prediction. The model is queried with a temperature of 0.3 to reduce output variability.¹
- **Embedding + Logistic Regression.** Buyer utterances are embedded using `text-embedding-3-large` (3072-dimensional representations). A multinomial logistic regression classifier is trained on these embeddings using L_2 regularization with a maximum of 1000 iterations. The random seed is fixed for reproducibility.
- **Embedding + SVM.** The same precomputed embeddings are used as input features. We train a Support Vector Machine with an RBF kernel, using the library’s standard regularization and kernel parameters.
- **Embedding + Random Forest.** The same embedding representations are used to train a Random Forest classifier. We use the default number of trees and feature-sampling strategy provided by the library.

3.3. Lexical Cues by Country

We analyze TF-IDF logistic regression coefficients to identify lexical features associated with each country. Features are ranked by their class-specific coefficient magnitude, and the top positively weighted cues are selected for inspection. The model assigns high weights to words and phrases that reflect country-specific linguistic style.

¹The full prompt is provided in the Appendix.

3.4. Within- vs. Between-Country Interaction Analysis

We compared negotiation outcomes between within-country and between-country dyads. All dialogues were collected under identical task conditions with matched incentives and roles. We examined resolution rates, buyer and seller points, joint points, and all facets of the Subjective Value Inventory (SVI) (Curhan et al., 2006). For this analysis, we included all dyads with complete outcome information ($N = 2076$ dialogues). Dyads were categorized as within-country or between-country based on whether the buyer and seller shared the same country label. Resolution outcomes were analyzed using a chi-square test of independence, while buyer points, seller points, joint points, and SVI measures were analyzed using mixed-effects ANOVAs with Role as a within-dyad factor and Match as a between-dyad factor.

4. Result

4.1. Performance

Table 1: Model performance comparison.

Method	Accuracy	F1-Score
LLM	0.431	0.374
LogReg	0.625	0.627
SVM	0.616	0.628
RandomForest	0.615	0.612
Chance (4-class)	0.250	—

In Table 1, embedding-based methods substantially outperform the LLM classifier across all metrics. Logistic Regression achieves 62.5% accuracy and an F1-score of 0.627, compared to 43.1% accuracy and 0.370 F1 for the LLM.

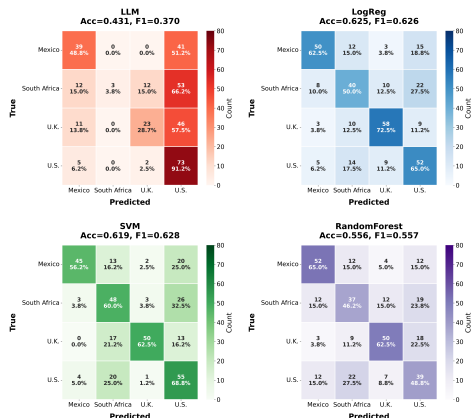


Figure 2: Confusion matrices for four models on the four-country classification task

The confusion matrices in Figure 2 further clarify why this gap emerges. The LLM exhibits a strong

prediction bias toward the U.S. class: for true U.S. instances, it predicts U.S. 91.2% of the time, but it also incorrectly maps a large proportion of Mexico (51.2%), South Africa (66.2%), and U.K. (57.5%) samples into the U.S. category.

In contrast, embedding-based classifiers show balanced decision boundaries across all four classes. Logistic Regression correctly identifies 62.5% of Mexico, 50.0% of South Africa, 72.5% of the U.K., and 65.0% of U.S. samples, with substantially reduced cross-country confusion. This balanced error distribution leads to a much higher macro F1-score. Similar patterns are observed for SVM (F1=0.628) and Random Forest (F1=0.557), all of which avoid the systematic overprediction seen in the LLM.

Limitations of Zero-Shot LLM Classification:

Analysis of the explanations (“reasons”) generated by the LLM alongside its predictions reveals two main sources of error. First, the model shows a strong bias toward the U.S. class, frequently interpreting generic negotiation language (e.g., references to refunds or shipping) as evidence of American origin, regardless of the speaker’s actual country. Second, the LLM relies primarily on explicit surface-level cues, such as spelling variants or named institutions, and fails to capture more subtle discourse-level differences when these markers are absent. These tendencies are reflected in the model’s own reasons, which often cite generic lexical cues rather than distinctive linguistic patterns.

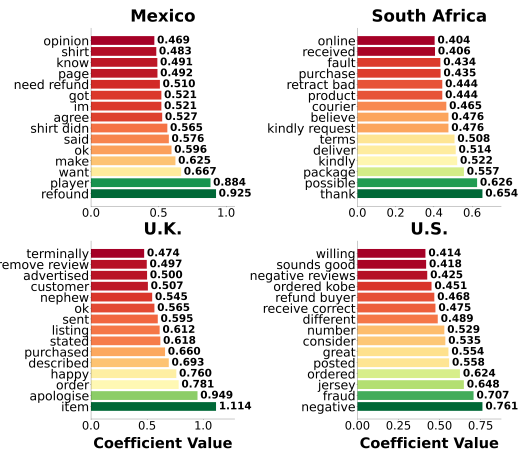


Figure 3: Lexical features with the highest coefficients for each country in the logistic regression model.

4.2. Corpus-level Linguistic Signals

Figure 3 shows distinctive words for each country, identified based on the highest positive logistic regression coefficients from the country-level classifier. These features reflect corpus-level stylistic tendencies rather than fixed cultural traits, and may

be shaped by task context, role expectations, and L2 language use. U.K. dialogues contain politeness markers and British spelling variants (e.g., “apologise”, “colour”). U.S. dialogues emphasize direct and transactional language (e.g., “refund”, “fraud”). Mexican dialogues show simplified phrasing and L2-influenced constructions (e.g., “im”, “ok”). South African dialogues contain procedural and regulatory terms (e.g., “courier”, “terms”).

4.3. Within- vs. Between-Country

Table 2: Result of Mixed ANOVA Analysis

DV	Effect	<i>p</i>	Direction
Points	Role	<0.001	Buyer > Seller
SVI Instrumental	Match	.040	Between > Within
SVI Instrumental	Role	.022	Buyer > Seller
SVI Relationship	Match	.004	Between > Within
SVI Relationship	Role	.017	Buyer > Seller

We analyzed interaction outcomes using mixed ANOVAs with Match (within-country vs. between-country) as a between-dyads factor and Role (buyer vs. seller) as a within-dyads factor. Resolution outcomes were analyzed separately using a chi-square test. Table 2 summarizes the main effects of Match and Role on negotiation outcome measures. Significant effects were observed for Role on points and for Match on instrumental and relationship-focused subjective value.

5. Discussion

Regarding RQ1, our results suggest that cultural identity-related signals may emerge through linguistic style even when such information is not explicitly disclosed. The results support this assumption: despite the absence of direct references to nationality, both zero-shot LLMs and embedding-based models were able to infer speakers’ countries from dialogue patterns alone. This suggests that social identity may be reflected not only in explicit content but also in subtle stylistic features such as orthographic choices, hedging, and discourse structure.

For RQ2, embedding-based classifiers captured identity signals more robustly than the zero-shot LLM in this setting. Although zero-shot LLMs are often effective in many NLP settings, they were less robust than embedding-based classifiers in our specific negotiation classification setup. One possible reason is that negotiation dialogue contains fewer explicit nationality markers than more conventional dialect identification settings. Error analysis further indicated that the LLM often relied on a small number of salient lexical cues and, when such cues were weak or absent, tended to default to predicting the U.S. class. This observation suggests that, in

our setting, the zero-shot model was more sensitive to sparse surface cues, while embedding-based models may have captured more distributed stylistic patterns.

For RQ3, the comparison between within-country and between-country interactions revealed an interesting dissociation: despite the competitive and emotionally charged nature of the negotiation task, objective outcomes did not significantly differ across conditions, whereas some subjective evaluations did. A more conservative interpretation is that identity-related cues may shape how interactions are subjectively experienced, even when they do not translate into measurable differences in objective outcomes. In such contexts, even emotionally intense exchanges may be interpreted more charitably in cross-country interactions, shaping relational perceptions without altering material negotiation outcomes. Together, these findings suggest that social signals expressed through language may influence how interactions are experienced and evaluated, rather than the objective results of the negotiation itself.

6. Conclusion

Overall, our findings show that implicit cultural identity cues can be detected in negotiation language. Embedding-based classifiers consistently outperformed zero-shot LLM predictions, suggesting that distributed stylistic patterns may provide more reliable signals for identity inference than surface-level reasoning. These results suggest that culture classification can serve as a useful measurement step for making country-linked variation observable in naturalistic dialogue data, enabling the study of cultural effects even when explicit metadata is unavailable. They also provide a starting point for examining when cultural signals become salient or are overridden in contexts shaped by strategic interaction and emotional intensity.

Several limitations should be noted. First, participants were not directly asked about their partner’s nationality or identity perception, making it difficult to determine whether identity inferences were consciously recognized or only implicitly processed. Second, although we compared multiple classifiers and a zero-shot LLM, additional model families or fine-tuned approaches may yield different patterns of robustness and generalization. Future work could examine how emotional dynamics interact with identity signaling, particularly by considering the role of tone, politeness, and attribution in text-based communication. It would also be valuable to investigate whether the strength or accuracy of inferred identity signals acts as a moderator that shapes partner responses and the unfolding of interaction dynamics.

Acknowledgements

Research was sponsored by the Air Force Office of Scientific Research under grant FA9550-23-1-0320. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References

- Khaled AlNuaimi, Gautier Marti, Mathieu Ravaut, Abdulla AlKetbi, Andreas Henschel, and Raed Jaradat. 2024. Enriching datasets with demographics through large language models: What's in a name? *arXiv preprint arXiv:2409.11491*.
- Marilynn B Brewer. 1999. The psychology of prejudice: Ingroup love and outgroup hate? *Journal of social issues*, 55(3):429–444.
- Mary Bucholtz and Kira Hall. 2005. Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, 7(4-5):585–614.
- Jared R Curhan, Hillary Anger Elfenbein, and Heng Xu. 2006. What do people value when they negotiate? mapping the domain of subjective value in negotiation. *Journal of personality and social psychology*, 91(3):493.
- Carsten KW De Dreu and Peter J Carnevale. 2003. Motivational bases of information processing and strategy in conflict and negotiation.
- Howard Giles and Peter F Powesland. 1975. *Speech style and social evaluation*. Academic Press.
- John J Gumperz. 1982. *Language and social identity*. 2. Cambridge University Press.
- James Anthony Hale, Sushrita Rakshit, Kushal Chawla, Jeanne M Brett, and Jonathan Gratch. 2025. Kodis: A multicultural dispute resolution dialogue corpus. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12771–12785.
- Kenneth D. Keith. 2013. *Person Perception*, pages 991–993. John Wiley & Sons, Ltd.
- Miriam Koschate, Elahe Naserian, Luke Dickens, Avelie Stuart, Alessandra Russo, and Mark Levine. 2021. Asia: Automated social identity assessment using linguistic style. *Behavior Research Methods*, 53(4):1762–1781.

Z. Kunda and P. Thagard. 1996. Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review*, 103:284–308.

William Labov. 2006. *The social stratification of English in New York city*. Cambridge University Press.

Edward J Lawler and Jeongkoo Yoon. 1995. Structural power and emotional processes in negotiation: A social exchange approach.

Henri Tajfel, John Turner, William G Austin, and Stephen Worchel. 2001. An integrative theory of intergroup conflict. *Intergroup relations: Essential readings*, pages 94–109.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, pages 48–57.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, Chris van der Lee, Stefan Grondelaers, Nelleke Oostdijk, Dirk Speelman, Antal van den Bosch, Ritesh Kumar, Bornini Lahiri, and Mayank Jain. 2018. *Language identification and morphosyntactic tagging: The second VarDial evaluation campaign*. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 1–17, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Appendices

A. LLM Prompt

LLM Prompt Snippet

System:

You are a helpful assistant that identifies the country of origin from a dialog.

User:

Below is a dialog from a customer service dispute over the purchase of an American basketball player jersey. Both buyer and seller are from the same country: U.S., U.K., Mexico, or South Africa.

Use the following linguistic clues to help decide:

- Spelling: “colour/organisation” (U.K.) vs “color/organization” (U.S.)

- Word choice: “advert” (U.K.) vs “ad” (U.S.)
- Spanish L2 typos: “recieved”, “missunderstanding” (Mexico)
- Delivery terms: Royal mail (U.K.), shipping (U.S.), courier (South Africa), parcel service (Mexico)
- Tone and references: PROFECO (Mexico), Consumer Protection Act (South Africa)
- Ignore currency formatting and the fact that the item is U.S. merchandise

Think step by step using these clues, but do not show your reasoning.

Dialog:

<Speaker text inserted here>

Output format:

Prediction: <country>, Reason: <reason>

How Much Does Persuasion Strategy Matter? LLM-Annotated Evidence from Charitable Donation Dialogues

Tatiana Petrova, Stanislav Sokol, Radu State

Interdisciplinary Centre for Security, Reliability and Trust (SnT)

University of Luxembourg

{tatiana.petrova, radu.state}@uni.lu, stanislav.sokol.001@student.uni.lu

Abstract

Which persuasion strategies, if any, are associated with donation compliance? Answering this requires fine-grained strategy labels across a full corpus and statistical tests corrected for multiple comparisons. We annotate all 10,600 persuader turns in the 1,017-dialogue PersuasionForGood corpus (Wang et al., 2019), where donation outcomes are directly observable, with a taxonomy of 41 strategies in 11 categories, using three open-source large language models (LLMs; Qwen3:30b, Mistral-Small-3.2, Phi-4). Strategy categories alone explain little variance in donation outcome (pseudo $R^2 \approx 0.015$, consistent across all three annotators). Guilt Induction is the only strategy significantly associated with *lower* donation rates ($\Delta \approx -23$ percentage points), an effect that replicates across all three models despite only moderate inter-model agreement. Reciprocity is the most robust positive correlate. Target sentiment and interest predict whether a donation occurs but show at most a weak correlation with donation amount. These findings suggest that strategy identification alone is insufficient to explain persuasion effectiveness, and that guilt-based appeals may be counterproductive in prosocial settings. We release the fully annotated corpus as a public resource.

Keywords: persuasion strategies, donation dialogues, LLM annotation, sentiment analysis, PersuasionForGood

1. Introduction

Charitable donation conversations are a natural setting for studying persuasion: a persuader attempts to convince a target to donate money, and the outcome (donated or not, and how much) is directly observable. The PersuasionForGood corpus (Wang et al., 2019)¹ provides 1,017 such dialogues collected via Amazon Mechanical Turk, where persuaders try to convince targets to donate part of their task earnings to Save the Children. Understanding which strategies help or hinder can inform the design of prosocial dialogue systems and evidence-based training for charitable fundraisers.

Subsequent work has improved strategy classification accuracy (Saha et al., 2021) and analyzed target resistance (Tian et al., 2020), but the question of which strategies, if any, are statistically associated with donation outcomes remains only partially addressed. Wang et al. (2019) tested strategy–donation associations via logistic regression on 252 annotated dialogues with 10 strategies, finding only “Donation information” significant ($p < 0.05$) without multiple-comparison correction. The limited sample (29% of the corpus), 10-strategy scheme, and absence of correction for multiple testing leave room for a more comprehensive analysis. To our knowledge, no prior work has tested individual strategy–outcome associations across the full corpus with a fine-grained taxonomy and multiple-comparison corrections.

¹We use the publicly available PersuasionForGood dataset (Wang et al., 2019).

We address this gap by (1) defining a hierarchical taxonomy of 41 persuasion strategies in 11 categories, grounded in Cialdini’s principles of influence (1984) and Marwell and Schmitt’s compliance-gaining strategies (1967); (2) annotating all 10,600 persuader turns using three open-source LLMs (Qwen3:30b as primary annotator, Mistral-Small-3.2 and Phi-4 as robustness checks) and all 10,332 target turns using Qwen3:30b alone (for sentiment and interest labels), following current best practice (Carlson and Burbano, 2025; Abdurahman et al., 2025); and (3) conducting bivariate tests with multiple-comparison corrections and multivariate logistic regression to assess strategy–donation associations. Our contributions include corpus-scale evidence on strategy–donation associations (including the Guilt Induction backfire and the limited predictive power of strategy categories), a three-model robustness design, and a fully annotated resource covering all 1,017 dialogues.

2. Related Work

PersuasionForGood and persuasion in NLP. Wang et al. (2019) introduced the corpus with 10 strategy labels (e.g., logical appeal, emotional appeal, credibility appeal) and an RCNN-based classifier. Saha et al. (2021) improved classification with BERT-based models, while Tian et al. (2020) analyzed target resistance strategies. Chen and Yang (2021) proposed weakly supervised identification of 8 persuasion strategies in online contexts.

Persuasion strategy taxonomies. Our taxonomy builds on Cialdini’s (1984) six principles of influence and Marwell and Schmitt’s (1967) 16 compliance-gaining strategies, organizing 41 strategies into 11 categories to enable analysis at both category and individual strategy level (Section 3.1).

LLM-as-annotator. Recent guidance recommends testing multiple models and treating model choice as a researcher degree of freedom (Gilardi et al., 2023; Pangakis et al., 2024; Carlson and Burbano, 2025; Abdurahman et al., 2025); we follow this practice (Section 3.2).

3. Methodology

3.1. Taxonomy

Starting from Cialdini’s (1984) principles of influence and Marwell and Schmitt’s (1967) compliance-gaining strategies, supplemented by work on fear appeals, framing, and emotional manipulation, we compiled 45 candidate strategies. Pilot annotation revealed that several were poorly distinguishable (e.g., overlapping moral and value-based appeals). After iterative merging and refinement, we arrived at 41 persuasion strategies in 11 categories, plus 9 conversation management labels (e.g., Greeting, Acknowledgement). Each strategy has a textual definition, characteristic markers, and decision rules (see released code).

The most frequent persuasion category is Norms / Morality / Values ($n = 1,331$, 12.6% of all persuader turns), followed by Rational / Impact Appeal (9.2%) and Framing & Presentation (7.3%). Overtly coercive strategies (Threat / Pressure, Urgency / Scarcity) are nearly absent, together accounting for only 0.2% of turns. Guilt Induction, while psychologically manipulative, is placed under Norms / Morality / Values because it operates through moral obligation rather than direct coercion.

3.2. Annotation Procedure

We annotate all 10,600 persuader utterances across 1,017 dialogues using three open-source LLMs deployed locally via Ollama: **Qwen3:30b** (Alibaba; primary annotator), **Mistral-Small-3.2** (Mistral AI), and **Phi-4** (Microsoft, 14B). These models were selected on three criteria. First, they represent distinct developer families (Alibaba, Mistral AI, Microsoft), reducing the risk that findings reflect idiosyncratic biases of a single training pipeline, in line with the robustness-check methodology of Carlson and Burbano (2025) and Abdurahman et al. (2025): one model serves as the primary annotator and the others as independent replications. Second, all three are fully open-weight models deployable locally, ensuring that the annotation pipeline

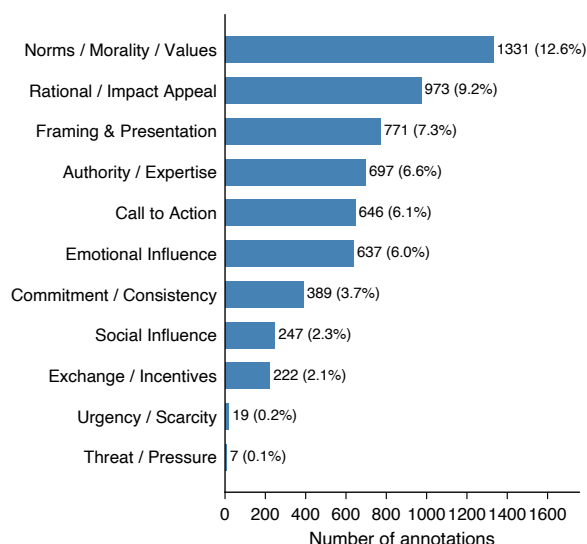


Figure 1: Distribution of 11 persuasion strategy categories (Qwen3:30b). Each bar shows the number of persuader turns assigned to the category; percentages indicate the category’s share of all $N=10,600$ persuader turns. Conversation Management turns (44.0%) are omitted.

is fully reproducible without dependence on proprietary APIs. Third, by including models of different parameter scales—Phi-4 (14B) and Qwen3 (30B) as the smallest and largest – we can assess whether the larger primary annotator’s capacity drives results; Table 4 shows that key effects hold across model sizes. Qwen3:30b is designated as primary because it assigns valid labels to all 10,600 turns without errors, produces non-degenerate distributions across all 11 categories (unlike Mistral, which assigns <1% to Framing & Presentation), and achieves the highest macro-level agreement with Wang et al.’s gold standard.

All three models use the same two-step hierarchical prompt: the system message instructs the model to act as a “hierarchical persuasion strategy classification system”; the user message presents (1) the persuader utterance to classify, (2) up to 5 previous dialogue turns as context, and (3) the full category→strategy hierarchy with definitions. The model first selects the parent category, then selects the specific strategy within that category, returning a structured JSON response. Temperature is set to 0.1 for near-deterministic output.

We also annotate all 10,332 target (persuadee) utterances using Qwen3:30b for two dimensions: *sentiment* (negative / neutral / positive, coded as $-1, 0, +1$) and *interest in donation* (not interested / neutral / interested, coded as $0, 1, 2$), capturing the target’s expressed engagement with the donation topic independently of affective tone: a turn may be affectively neutral yet indicate genuine curiosity about the charity or a willingness to consider donat-

ing. We treat mean target sentiment and interest as *covariates* rather than primary predictors: both are measured *during* the conversation and thus reflect the target’s evolving response to persuasion rather than pre-existing dispositions; causal direction between strategies, target responses, and donation cannot be established from observational data alone (see Limitations).

3.3. Annotation Quality

We evaluate annotation quality at two levels, following best practices for LLM-based annotation (Pangakis et al., 2024). No human inter-annotator agreement study exists for a 41-label persuasion strategy task; the expected human ceiling for this taxonomy is unknown.

Cross-taxonomy validation. We compare our annotations against Wang et al.’s (2019) gold standard on 300 dialogues (3,047 turns). Because the 41-strategy and 10-strategy schemes are structurally different, we map both to three macro-categories (*persuasive appeal*, *persuasive inquiry*, *non-strategy*), obtaining moderate agreement (Cohen’s $\kappa = 0.507$, macro $F_1 = 0.703$; “moderate” on the Landis and Koch 1977 scale). This is lower than Wang et al.’s human $\alpha > 0.70$ but was obtained zero-shot without task-specific training.

Expert verification. An expert in persuasion dialogue reviewed a stratified sample of 100 persuader turns covering all strategy labels annotated by the LLM, confirming correct classification in 84 of 100 cases (84%). The 16 disagreements predominantly involved boundary cases between semantically adjacent strategies (e.g., Emotional Appeal vs. Empathy Appeal, Moral Appeal vs. Self-feeling Appeal). As this is verification rather than independent blind annotation, we report accuracy rather than Cohen’s κ .

Inter-model agreement. On all 10,600 persuader turns, pairwise Cohen’s κ between the three models ranges from 0.38 to 0.54 at the strategy level (“fair” to “moderate”) and from 0.44 to 0.62 at the category level. At the macro level (persuasion vs. conversation management), agreement is higher: $\kappa = 0.66$ – 0.75 , with raw agreement of 84–88%. Three-way exact match is 34.1% for strategies and 47.5% for categories. The models diverge most on fine-grained labels (e.g., Rational Appeal vs. Credibility Appeal) but converge on functional classification. The key downstream findings, in particular the Guilt Induction backfire effect, replicate across all three annotators (Section 4.7).

4. Analysis and Results

We conduct analyses at two levels of granularity. At the **category level** (Section 4.1), we test whether the presence of each of the 11 strategy categories in a dialogue is associated with donation outcome. At the **individual strategy level** (Sections 4.2–4.3), we restrict tests to strategies appearing in at least 20 dialogues ($n \geq 20$), a minimum-frequency threshold adopted to ensure reliable chi-square inference; 22 of 41 strategies meet this criterion (full counts in Appendix A). In both analyses we apply chi-square tests with Bonferroni and Benjamini-Hochberg (FDR) corrections for multiple comparisons, followed by multivariate logistic regression to assess independent effects (Section 4.5). Target sentiment and interest are included as covariates rather than primary predictors, as their role is detailed in Section 3.2; bivariate associations with donation outcome are reported in Section 4.4.

The overall donation rate is 53.6% (545 / 1,017 dialogues), with a mean donation of \$2.17 (\$4.05 among donors only). Dialogues contain an average of ~ 10 persuader turns, typically employing 4–5 distinct persuasion strategies per dialogue (mean = 4.4, $SD = 1.6$, range 0–10); persuasion in this setting is thus multi-strategy by nature. All results below use Qwen3:30b (primary annotator) unless otherwise noted.

4.1. Strategy Categories Do Not Independently Predict Donation

We test whether the presence of each strategy category in a dialogue is associated with donation outcome using chi-square tests with Bonferroni correction for 11 comparisons. None of the 11 strategy categories reach statistical significance (all $p_{Bonf} > 0.05$). The closest is Commitment / Consistency ($p_{Bonf} = 0.146$), followed by Rational / Impact Appeal ($p_{Bonf} = 0.280$). This null result replicates across all three annotators (Table 4).

Because dialogues are long (avg. ~ 10 turns) and contain many strategies simultaneously, most categories appear in most dialogues (Norms / Morality / Values in 72%, Rational / Impact Appeal in 56%), making binary presence/absence a coarse signal.

4.2. Guilt Induction is Associated with Lower Donation

At the individual strategy level, we test 22 strategies with $n \geq 20$ dialogues, applying both Bonferroni and Benjamini-Hochberg (FDR) corrections. Guilt Induction is the only strategy significantly associated with *lower* donation likelihood (Table 1). Dialogues containing Guilt Induction ($n = 104$) have a 32.7% donation rate, compared to 56.0% for dialogues without it ($n = 913$; $\chi^2 = 19.4$, $\phi = 0.14$,

$p_{Bonf} < 0.001, p_{FDR} < 0.001$). The mean donation amount is also $3.5\times$ lower (\$0.67 vs. \$2.34, Mann-Whitney U $p < 0.001$). This effect replicates across all three annotators: $\Delta = -23.0$ pp with Mistral-Small-3.2 and -23.9 pp with Phi-4, both significant at $p_{raw} < 0.05$ (Table 4).

To illustrate, a typical Guilt Induction utterance from a non-donated dialogue reads: “Kids are dying from hunger every minute. Don’t you want to help stop that?” In contrast, a Reciprocity utterance from a donated dialogue: “That’s great! Every bit helps, I will match your donation myself.” Guilt induction threatens the target’s autonomy; reciprocity creates a mutual exchange.

This pattern is consistent with psychological reactance theory (Brehm, 1966): when individuals perceive that their freedom of choice is threatened, they resist rather than comply (Figure 2a).

This association is correlational and may partly reflect reverse causality (see Discussion). Our logistic regression (Section 4.5) shows that Guilt Induction remains a negative predictor (odds ratio [OR] = 0.60, $p = 0.029$) even when controlling for target sentiment and interest.

4.3. Reciprocity and Commitment/Consistency are Positive Predictors

Reciprocity is the strategy most robustly associated with *higher* donation rates: 72.2% of dialogues containing Reciprocity ($n = 72$) result in donation, vs. 52.2% without ($n = 945$; $\chi^2 = 10.0$, $\phi = 0.10$, $p_{Bonf} = 0.034$, $p_{FDR} = 0.016$). The positive direction replicates in all three models ($\Delta = +5.8$ to $+20.1$ pp), reaching significance in two of three (Table 4). Commitment and Consistency, while not significant under Bonferroni correction, also reaches significance under FDR ($p_{FDR} = 0.016$; 63.5% vs. 51.2%). For both strategies, reverse causality cannot be excluded: persuaders may deploy them after the target has already signaled willingness.

4.4. Target Sentiment and Interest Predict Donation but Not Amount

The target’s expressed sentiment and interest are the variables most strongly associated with whether a donation occurs (both $p < 0.001$; Table 1, Figure 2b). Donated dialogues have higher mean target sentiment (0.44 vs. 0.27 on a -1 to $+1$ scale) and higher mean interest (1.24 vs. 1.09 on a 0 to 2 scale). At the turn level, donated dialogues contain more positive target turns (49.5% vs. 38.6%) and fewer negative turns (6.1% vs. 12.0%).

However, neither sentiment nor interest shows a statistically significant linear correlation with the donation *amount* among those who did donate (Pearson $r = 0.034$, $p = 0.424$ and $r = -0.001$,

	Pres.	Abs.	p_B	p_{FDR}
<i>Strategy–donation (22 tests, χ^2)</i>				
Guilt	32.7%	56.0%	<.001	<.001
Reciprocity	72.2%	52.2%	.034	.016
Commit./Cons.	63.5%	51.2%	.049	.016
<i>Target response–donation (MWU)</i>				
	Don.	No don.	p	
Sentiment	.44	.27	<.001	
Interest	1.24	1.09	<.001	
<i>Target response–amount (Pearson)</i>				
	r		p	
Sentiment	.034		.424	
Interest	−.001		.981	

Table 1: Bivariate associations with donation outcome (Qwen3:30b). *Top*: donation rates (%) in dialogues where the strategy is present (Pres.) vs. absent (Abs.); p_B : Bonferroni, p_{FDR} : Benjamini-Hochberg correction over 22 tests. Effect sizes are small ($\phi = 0.14$ for Guilt, 0.10 for Reciprocity). *Middle*: mean target sentiment (-1 to $+1$) and interest ($0-2$) in donated (Don.) vs. not-donated (No don.) dialogues, Mann-Whitney U test. *Bottom*: Pearson r between target response and donation amount (donors only). Cross-model robustness is reported in Table 4.

$p = 0.981$, respectively). Spearman rank correlation detects a small monotonic association for sentiment ($\rho = 0.112$, $p = 0.009$) but not for interest ($\rho = 0.059$, $p = 0.172$), suggesting a weak non-linear link between target sentiment and donation amount that the linear measure misses. The “Negative” sentiment group contains only 23 dialogues (3 donations), so the 13.0% rate in Figure 2b carries a wide confidence interval and should be interpreted with caution.

4.5. Logistic Regression: Multivariate Analysis

The bivariate tests above examine each predictor in isolation. To assess whether strategy effects survive when controlling for other predictors, we fit logistic regression models with donation (binary) as the dependent variable (Table 2).

Model 1: Strategy categories only. With the 9 strategy categories present in ≥ 20 dialogues as binary predictors, the model is significant overall (log-likelihood ratio [LLR] $p = 0.011$) but explains little variance (pseudo $R^2 = 0.015$). Commitment / Consistency (OR = 1.41, $p = 0.014$) and Rational / Impact Appeal (OR = 1.36, $p = 0.018$) are positive predictors; no category is a significant negative predictor.

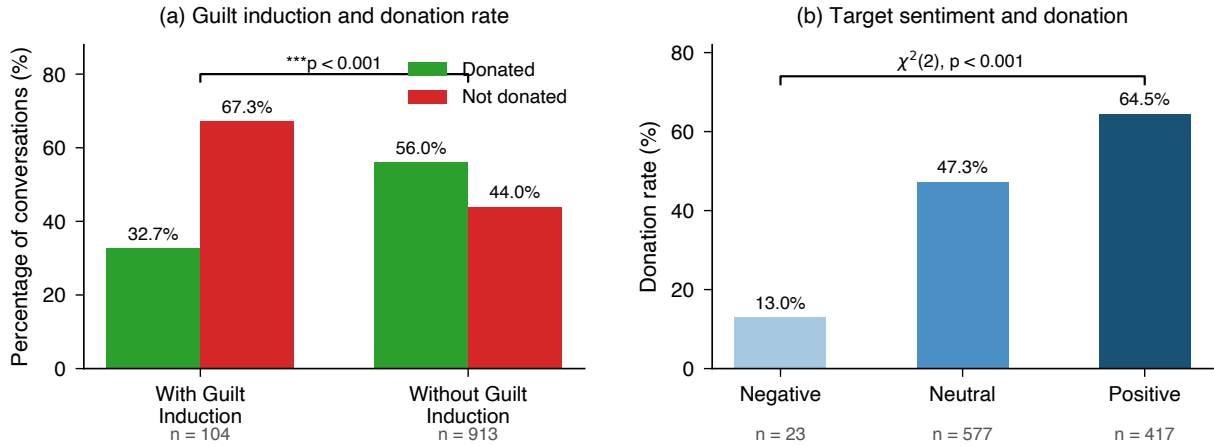


Figure 2: (a) Donation rates in dialogues containing Guilt Induction ($n=104$) vs. dialogues without ($n=913$); difference of -23.3 pp ($\chi^2=19.4, p < 0.001$). This effect replicates across all three annotators (Table 4). (b) Donation rate by predominant target sentiment across a dialogue’s target turns ($\chi^2(2)=44.3, V=0.21, p < 0.001$).

Model 2: Categories + sentiment + interest.

Adding mean target sentiment and interest improves fit substantially (pseudo $R^2 = 0.082$; likelihood ratio test vs. Model 1: $\chi^2 = 94.4, p < 10^{-6}$). Sentiment (OR = 4.24) and interest (OR = 3.11) yield the largest effects (both $p < 0.001$). Commitment / Consistency loses significance ($p = 0.014 \rightarrow 0.454$), suggesting either confounding or mediation through target sentiment. Call to Action emerges as a negative predictor (OR = 0.70, $p = 0.009$).

Model 3: Parsimonious model (exploratory).

As an exploratory check, we fit a compact model with only Guilt Induction, Reciprocity, mean sentiment, and mean interest (the variables with the strongest bivariate signals). This model achieves nearly the same fit (pseudo $R^2 = 0.080$, Akaike information criterion [AIC] = 1302, area under the ROC curve [AUC] = 0.67) as the full model (AIC = 1314). All four predictors are significant: sentiment (OR = 3.94, $p < 0.001$), interest (OR = 2.77, $p < 0.001$), Reciprocity (OR = 2.41, $p = 0.002$), and Guilt Induction (OR = 0.60, $p = 0.029$). For comparison, a categories-only model achieves AUC = 0.58.

4.6. Strategy–Response Sentiment Link

To explore why certain strategies are associated with donation, we pair each persuader turn carrying a persuasion strategy label with the immediately following target turn ($n = 5,387$ persuader–target pairs) and compute the mean target sentiment elicited by each strategy (Table 3). The corpus-wide average response sentiment is +0.37. Guilt Induction (+0.02) elicits responses far below the corpus average, as do Fear Appeal (+0.19) and Unity (+0.23). In contrast, Commitment and Con-

Predictor	OR	95% CI	p
<i>Model 3 (parsimonious): pseudo $R^2 = 0.080, AUC = 0.67$</i>			
Sentiment	3.94	[2.30, 6.74]	<.001
Interest	2.77	[1.64, 4.67]	<.001
Reciprocity	2.41	[1.38, 4.20]	.002
Guilt Ind.	0.60	[0.37, 0.95]	.029

Table 2: Logistic regression predicting donation (binary), parsimonious Model 3. OR: odds ratio (> 1 = higher donation probability). Sentiment: mean target sentiment per dialogue (-1 to $+1$). Interest: mean target interest per dialogue (0 – 2).

sistency (+0.54), Reciprocity (+0.51), and Foot-in-the-door (+0.51) elicit the most positive responses. Reciprocity co-occurs with positive engagement; the affective response pattern points to a possible mediating role in the strategy–donation link, though formal mediation analysis would be needed to confirm this.

4.7. Cross-Model Robustness

To assess whether our findings depend on the choice of annotator, we replicate the full analysis pipeline with Mistral-Small-3.2 and Phi-4 annotations (Table 4). The Guilt Induction backfire effect is the most robust finding: the effect direction and magnitude ($\Delta \approx -23$ pp) are consistent across all three model families, despite the models identifying different numbers of Guilt turns (Qwen: 104 dialogues, Mistral: 35, Phi-4: 36) and achieving only moderate pairwise agreement ($\kappa = 0.38$ – 0.54). To probe whether this consistency reflects a core set of “obvious” guilt turns, we examine the intersection: only 16 dialogues are flagged by all three models (Jaccard = 0.14), and 63 are flagged by Qwen

Strategy	n	Mean sent.	% neg.
<i>Bottom 3 (lowest mean sentiment)</i>			
Guilt Induction	114	+0.02	27.2%
Fear Appeal	69	+0.19	21.7%
Unity	43	+0.23	23.3%
<i>Top 3 (highest mean sentiment)</i>			
Commit. & Cons.	182	+0.54	7.1%
Reciprocity	67	+0.51	7.5%
Foot-in-the-door	41	+0.51	7.3%
<i>Corpus avg.</i>	5,387	+0.37	—

Table 3: Mean target sentiment (−1 to +1) in the turn immediately following each persuader strategy ($n = 5,387$ persuader–target pairs). Only persuasion strategies with $n \geq 20$ pairs shown; Conversation Management labels excluded. % neg.: proportion of negative target responses.

alone. The donation rate is low for both subsets (37.5% for all-three-agree, 30.2% for Qwen-only; Fisher $p = 0.56$ for the difference), indicating that the effect is not confined to extreme cases; even borderline guilt turns identified by a single model are associated with lower donation rates (any-guilt union: 29.9% vs. no-guilt: 56.7%, $p < 0.001$). The positive Reciprocity association replicates in direction across all three models and reaches significance in two of three. The null result for categories (pseudo $R^2 \approx 0.015$) and the dominance of sentiment and interest in the full model (pseudo $R^2 \approx 0.08$) are stable across all annotators.

Finding	Qwen	Mistral	Phi-4	Repl.
Guilt Δ (pp)	−23.3 ^{***}	−23.0 [*]	−23.9 ^{**}	3/3
Recip. Δ (pp)	+20.1 ^{**}	+5.8	+10.4 [*]	2/3
Categories R^2	.015	.011	.016	3/3
+ Sent./Int. R^2	.082	.079	.077	3/3

Table 4: Cross-model robustness. Δ : difference in donation rate (pp) between dialogues with vs. without the strategy. R^2 : McFadden pseudo R^2 . Significance: * $p < .05$, ** $p < .01$, *** $p < .001$ (uncorrected p_{raw} ; multiple-comparison correction is applied only to the primary Qwen model in Table 1). Repl.: models where the finding reaches significance (p_{raw}) or matches direction (for R^2).

5. Discussion and Conclusion

Our results indicate that persuasion effectiveness cannot be reduced to “strategy X leads to donation.”

First, strategy categories have limited predictive power (pseudo $R^2 = 0.011$ – 0.016 across all three annotators), challenging the assumption that strategy identification alone captures persuasion effectiveness. With 4–5 strategies per dialogue, binary

presence/absence is inherently coarse; future work should model strategy *sequences* and *combinations*.

The moderate inter-model agreement ($\kappa = 0.38$ – 0.54) is expected for a 41-label zero-shot task with many semantically adjacent strategy pairs. Key findings replicate regardless of this disagreement (Table 4), while weaker effects, notably Reciprocity (significant in two of three models), should be interpreted with more caution. Following Carlson and Burbano (2025), effects replicating across all three annotators provide stronger evidence than single-model results.

Second, the Guilt Induction backfire effect has a practical implication: prosocial dialogue systems may benefit from avoiding guilt-based appeals. Temporal analysis supports this: splitting the 104 guilt-containing dialogues by the position of the first guilt turn, donation rates decline monotonically from early (45.9%, $n=37$) through mid (34.3%, $n=35$) to late guilt (15.6%, $n=32$), compared to the no-guilt baseline of 56.0% ($\chi^2(3) = 26.7$, $p < 0.001$). After Bonferroni correction, only late guilt differs significantly from the baseline ($p_{adj} < 0.001$). This pattern may reflect reverse causality (persuaders may resort to guilt after sensing resistance), but even early guilt underperforms the baseline numerically.

Resource contribution. We release the full annotated dataset (20,932 turns across 1,017 dialogues): strategy labels from all three annotators for all 10,600 persuader turns and sentiment/interest labels for all 10,332 target turns, together with prompt templates, analysis scripts, and validation code.² The full taxonomy is in Appendix A.

6. Limitations

Our LLM annotations are produced without fine-tuning; inter-model agreement on fine-grained labels is moderate ($\kappa = 0.38$ – 0.54), and annotation noise may attenuate downstream estimates. Key findings (Guilt backfire, low category R^2) are robust to annotator choice, but weaker effects (e.g., Reciprocity) vary across models. Each turn receives one label, though turns may contain multiple strategies; forced single-label annotation may undercount co-occurring strategies. Sentiment and interest are measured *during* the conversation, so they function as concurrent mediators rather than exogenous predictors; causal mediation analysis would be needed to disentangle strategy effects from target response effects. The released data lacks individual worker identifiers, so non-independence across dialogues cannot be ruled out. Finally, all

²Code and data: <https://github.com/persuasion-nlp/persuasion-strategies>

models explain at most 8% of variance (pseudo $R^2 = 0.08$); all reported associations are correlational. Expert verification was conducted by a single annotator with expertise in persuasion and dialogue research; while the 84% accuracy on a stratified 100-turn sample suggests acceptable label quality, a multi-annotator blind evaluation would yield a more reliable estimate and is left for future work.

7. Ethics Statement

This work analyzes existing publicly available dialogue data (Wang et al., 2019). No new human subjects data was collected. The persuasion strategies we study are from cooperative charitable donation contexts. Findings about persuasion strategy effectiveness could theoretically inform manipulative applications; however, the primary intended use is understanding human persuasion dynamics and improving prosocial dialogue systems.

List of Abbreviations

AIC	Akaike Information Criterion
AUC	Area Under the ROC Curve
CI	Confidence Interval
FDR	False Discovery Rate
LLM	Large Language Model
LLR	Log-Likelihood Ratio
MWU	Mann-Whitney U (test)
OR	Odds Ratio
pp	percentage points

8. Bibliographical References

Suhaib Abdurahman, Alireza Salkhordeh Ziabari, Alexander K. Moore, Daniel M. Bartels, and Morteza Dehghani. 2025. [A primer for evaluating large language models in social-science research](#). *Advances in Methods and Practices in Psychological Science*, 8(2).

Jack W. Brehm. 1966. *A Theory of Psychological Reactance*. Academic Press.

Natalie A. Carlson and Vanessa Burbano. 2025. [The use of LLMs to annotate data in management research: Foundational guidelines and warnings](#). *Strategic Management Journal*. Early view.

Wanqin Chen and Diyi Yang. 2021. Weakly supervised persuasion strategy identification in online persuasion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12704–12711.

Robert B. Cialdini. 1984. *Influence: The Psychology of Persuasion*. William Morrow.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Gerald Marwell and David R. Schmitt. 1967. Dimensions of compliance-gaining behavior: An empirical analysis. *Sociometry*, 30(4):350–364.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2024. Automated annotation with generative AI requires validation. *arXiv preprint arXiv:2306.00176*.

Tulika Saha, S. R. Jayashree, Sriparna Saha, and Pushpak Bhattacharyya. 2021. Towards modeling the style of persuasive strategies in persuasive dialogues. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3873–3883. Association for Computational Linguistics.

Youzhi Tian, Weiyan Shi, Chen Li, and Zhou Yu. 2020. Understanding user resistance strategies in persuasive conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4799–4808. Association for Computational Linguistics.

Xuwei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649. Association for Computational Linguistics.

9. Language Resource References

Wang, Xuwei and Shi, Weiyan and Kim, Richard and Oh, Yoojung and Yang, Sijia and Zhang, Jingwen and Yu, Zhou. 2019. *PersuasionForGood Corpus: 1,017 Persuasion Dialogues for Charitable Donations*. distributed via GitHub. PID <https://gitlab.com/ucdavisnlp/persuasionforgood>.

Category	Strategy	n	%
Norms / Morality / Values	Appeal to Values	543	5.1
	Moral Appeal	521	4.9
	Guilt Induction	129	1.2
	Self-feeling Appeal	138	1.3
Rational / Impact Appeal	Rational Appeal	928	8.8
	Logical Appeal	45	0.4
Framing & Presentation	Framing	751	7.1
	Loss Aversion Appeal	18	0.2
	Bait-and-switch	1	<0.1
	Pretexting	1	<0.1
Authority / Expertise	Credibility Appeal	683	6.4
	Expertise	7	0.1
	Authority	7	0.1
Emotional Influence	Empathy Appeal	265	2.5
	Storytelling	180	1.7
	Emotional Appeal	76	0.7
	Fear Appeal	73	0.7
	Sympathy Appeal	40	0.4
Call to Action	Emotional Manipulation	3	<0.1
	Call to Action	635	6.0
Commitment / Consistency	Liking	11	0.1
	Commitment and Consistency	232	2.2
	Activ. of Personal Commitment	95	0.9
	Foot-in-the-door	45	0.4
Social Influence	Door-in-the-face	17	0.2
	Social Proof	191	1.8
	Unity	53	0.5
Exchange / Incentives	Social Positioning	3	<0.1
	Reciprocity	88	0.8
	Rewarding Activity	71	0.7
	Pre-giving	57	0.5
Urgency / Scarcity	Debt	6	<0.1
	Urgency	17	0.2
Threat / Pressure	Scarcity	2	<0.1
	Threat	5	<0.1
Conversation Management	Aversive Stimulation	2	<0.1
	Greeting / Rapport	1,060	10.0
	Acknowledgement	1,022	9.6
	Charity Awareness Probe	792	7.5
	Non-persuasive Other	608	5.7
	Logistics / Coordination	350	3.3
	Donation Baseline / Habit Probe	325	3.1
	Conversation Closing	297	2.8
	Qualification / Segmentation	138	1.3
Permission / Time Check	69	0.7	

Table 5: Complete taxonomy: 41 persuasion strategies in 11 categories plus 9 conversation management labels (Qwen3:30b annotations). Columns show absolute turn counts and each label's percentage of all $N=10,600$ persuader turns (percentages sum to 100%). Five of the 41 taxonomy strategies (Punishing Activity, Overloading, Confusion Induction, Promise, and Activation of Impersonal Commitment) received zero assignments and are not shown.

A. Full Strategy Taxonomy

Table 5 lists all 41 persuasion strategies and 9 conversation management labels with turn counts.

Do LLMs Ask the Right Questions? Evaluating GPT-Generated Surveys as Instruments for Measuring Social Attitudes

Tina Behzad, Wenbo Li, Reuben Kline, Klaus Mueller

Stony Brook University

{tbehzad, mueller}@cs.stonybrook.edu, {wenbo.li.1, reuben.kline}@stonybrook.edu

Abstract

Understanding human beliefs and social attitudes often relies on carefully designed survey instruments. Recent work has suggested that large language models (LLMs) could automate parts of this process by generating surveys at scale, raising questions about the comparability of such instruments to literature-grounded, human-designed surveys. We present a controlled empirical comparison between GPT-generated surveys and established survey baselines across three social domains: climate change, immigration, and diversity, equity, and inclusion (DEI). GPT-generated surveys were produced using a fixed prompting framework enforcing a 3×3 structure over beliefs, perceptions, and behaviors, while human baselines were assembled from validated instruments to match survey length and construct coverage. We collected responses from U.S.-based participants, who completed both survey types, allowing direct within-subject comparison. We analyze differences in response distributions, clustering behavior, and alignment with self-identified stances. Our results show that GPT-generated surveys capture the same dominant attitudinal divisions as human-designed instruments, while exhibiting differences in the resolution of belief structure and group separation. These findings suggest that LLM-generated surveys are suited for exploratory and large-scale analyses, and can be used to complement expert-designed instruments.

Keywords: survey design, attitude measurement, human–AI comparison

1. Introduction

Surveys are a central tool for studying human beliefs, values, and social attitudes across disciplines such as psychology, political science, and sociology. Carefully designed survey instruments are not merely collections of questions, but theory-informed measurement tools that encode assumptions about which constructs matter, how they should be operationalized, and how responses should be interpreted (Dillman et al., 2014; Boynton and Greenhalgh, 2004; Synodinos, 2003). As a result, survey design choices such as question framing, construct coverage, and scale structure play a critical role in shaping the conclusions drawn about public opinion and social behavior (Knäuper et al., 1997; Lenzner, 2012).

Recent advances in large language models (LLMs) have raised the possibility of automating or augmenting survey design. Prior work has shown that LLMs can generate fluent survey questions (Maiorino et al., 2023), simulate respondent behavior (Argyle et al., 2023; Bisbee et al., 2024), and scale social data collection (Wuttke et al., 2025), facilitating rapid pilot studies, iterative instrument refinement, and large-scale processing of unstructured responses, tasks that are typically resource-intensive and slow in survey workflows.

In this work, we present a systematic comparison between GPT-generated surveys and human-designed surveys as a baseline for measuring beliefs and stances on social topics. Such comparisons are essential for determining whether LLMs can be responsibly used as tools in questionnaire

design without compromising measurement validity or analytic conclusions. We focus on three timely and socially consequential domains: climate change, immigration, and diversity, equity, and inclusion (DEI). We evaluate whether surveys generated using a fixed, structured prompt template produce coherent and reliable instruments across topics and elicit response patterns comparable to those obtained using established instruments. Rather than assessing surface-level similarities in wording or face validity, we examine how the two survey types differ in their ability to capture belief variation, support meaningful clustering of respondents, and align with self-identified stances.

Our contributions are threefold. First, we introduce a framework for generating surveys using LLMs that explicitly controls construct coverage and survey length, enabling direct comparison with human-designed instruments. Second, we conduct an empirical evaluation of LLM-generated and human-designed surveys across multiple social domains, using responses collected online from 150 U.S.-based participants per topic. Third, we analyze where and how GPT-generated surveys diverge from human baselines in ways that matter for interpreting social beliefs, highlighting both their potential and their limitations as measurement tools.

Taken together, our findings contribute to ongoing discussions at the intersection of NLP and psychology by clarifying when LLM-generated surveys may be useful and where caution is warranted. More broadly, this work underscores the importance of treating survey generation not as a purely linguistic task, but as a measurement problem with sub-

stantive implications for how social attitudes are modeled and understood.

2. Related Work

This section reviews foundational approaches to survey design, emerging work on LLM-based survey generation, and common approaches to quantify and evaluate survey quality.

2.1. Measuring Social Attitudes

Survey instruments are central to measuring beliefs and social attitudes, but their validity depends on how abstract constructs are operationalized into items, response scales, and questionnaire context (Groves et al., 2011). Classic work in survey methodology demonstrates that seemingly small design choices such as question wording, ordering, and framing, can systematically shift responses and even change the meaning respondents assign to a question (Schuman and Presser, 1996). More broadly, survey methodology emphasizes that surveys are subject to multiple sources of error (e.g., comprehension, recall, judgment, and response mapping), and that careful questionnaire construction is necessary for producing interpretable and reliable measurements (Groves et al., 2011).

Survey instruments aimed at measuring social attitudes typically begin with the definition of an underlying construct space that specifies which dimensions of belief are to be measured and how they relate to one another. Rather than treating attitudes as unitary variables, a large body of work in psychology and social science conceptualizes attitudes as multidimensional, often distinguishing between cognitive beliefs, affective evaluations, and behavioral or policy-oriented intentions (Ajzen, 1991; Eagly and Chaiken, 1993). This tripartite view has informed the design of many survey instruments, particularly in domains where attitudes are complex, value-laden, and socially contested.

Operationalizing construct spaces into survey items has been approached through multiple methodological strategies. Some surveys rely on multi-item, construct-based scales, in which parallel items are designed to probe distinct facets of an underlying construct (Cronbach and Meehl, 1955), while others adopt segmentation-oriented designs that prioritize a small number of high-signal questions to differentiate respondent groups rather than exhaustively measure all dimensions (Chryst et al., 2018). In applied and large-scale survey settings, these design choices are typically shaped by a combination of theoretical grounding, prior empirical usage, and practical constraints such as respondent burden and survey length (Groves et al., 2011). Across approaches, a common principle is

the use of structured and balanced item layouts that explicitly sample across dimensions to support interpretability and comparability, facilitating downstream analyses such as clustering and segmentation (Saris and Gallhofer, 2014).

2.2. LLM-Based Survey Generation

The emergence of LLMs has created new opportunities across the survey lifecycle, spanning instrument design, administration, and response analysis. Prior work has explored LLMs as generators of synthetic survey responses (Argyle et al., 2023; Hämäläinen et al., 2023), as tools for survey question selection, editing, and drafting (Rothschild et al., 2024), and as aids for annotating and analyzing open-ended responses (Valenzuela et al., 2025; Jansen et al., 2023).

Early work on question generation demonstrated that LLMs can support researchers in editing and paraphrasing survey questions, as well as in re-designing response options for multiple-choice items (Rothschild et al., 2024). More recently, a growing body of work has shifted toward using LLMs to generate entire questionnaires. For example, Adhikari et al. (2025) study how LLMs can be prompted to produce questionnaire items that adhere to specified formats, constructs, and design constraints, highlighting their potential for rapid survey creation and customization. Related studies have examined the use of LLM-generated surveys in applied settings such as communication research (Valenzuela et al., 2025), market research (Keane and McNaughton, 2025), and personality measurement (Oeljeklaus et al., 2025).

In contrast to prior work that emphasizes topic-specific applications or task-tailored prompting, we investigate whether LLMs can be used in a more generalizable manner by employing a single, universal prompt to generate questionnaires across diverse social topics, and assess how reliably these instruments compare to expert-designed surveys.

2.3. Evaluating Survey Quality

A common approach to evaluating survey quality in the social sciences is to examine whether responses support the discovery of meaningful latent structure and are predictive of downstream intentions, behaviors, or policy-relevant outcomes. A substantial literature demonstrates that clustering and latent-class analyses of survey data can recover meaningful population segments that differ systematically in behaviors, participation patterns, and policy preferences, motivating their use as analytic tools rather than purely descriptive techniques. For example, climate-opinion research has shown that audience segments derived from survey responses capture heterogeneity in engagement be-

haviors and policy support, enabling prediction and targeted communication strategies (Chryst et al., 2018; Maibach et al., 2011; Leiserowitz et al., 2021). Similarly, political science research has used latent class and profile models to identify participation and voting types that predict turnout and modes of civic engagement more effectively than item-level analyses (Greaves et al., 2015; Johann et al., 2020; Oser, 2022). More broadly, foundational work on attitude-behavior relationships establishes that survey-measured latent constructs are meaningful precisely because they support the prediction of intentions and behaviors across domains (Ajzen, 1991; McEachan et al., 2016). Within this context, recent work on LLM-generated surveys has primarily evaluated generated instruments through comparisons with expert-designed questionnaires, often using psychometric criteria. Studies have assessed whether LLM-generated items reproduce the factor structure, internal consistency, and convergent validity of established scales, treating human designed surveys as reference measurement models (Adhikari et al., 2025; Terry et al., 2025; Oeljeklaus et al., 2025). While these evaluations are typically tied to study-specific measures and predefined scales, we adopt similar principles within a more general evaluation framework. In particular, we extend assessment beyond construct replication to examine latent structure, respondent segmentation, stability, and predictive utility, thereby evaluating the robustness and analytic usefulness of the induced belief structure across domains.

3. Methodology

This section describes the experimental design and analysis procedures used in this study. We employ a comparative setup in which GPT-generated surveys are evaluated alongside literature-grounded, human-designed surveys targeting the same social topics. The study proceeds in four stages:

1. **Survey Generation** (Section 3.1): GPT-generated surveys were produced using a controlled prompting framework specifying construct structure, formatting constraints, and generation parameters.
2. **Human Baseline Construction** (Section 3.2): Literature-grounded survey instruments were assembled from established sources to match the length and conceptual coverage of the GPT-generated surveys.
3. **Response Collection** (Section 3.3): U.S.-based participants completed both survey types under randomized ordering, along with additional survey-quality assessments.

4. **Comparative Analysis** (Section 3.4): We analyzed differences in survey structure and response patterns using quantitative measures of construct coverage, clustering behavior, and alignment with self-identified stances.

For the purpose of this paper we focus on three timely and socially consequential domains: climate change, immigration, and diversity, equity, and inclusion (DEI). These topics were selected for their broad scope and frequent presence in everyday discourse, making it likely that individuals across diverse backgrounds hold opinions on them, while also helping to reduce topic-specific confounds and enhance the generalizability of our findings.

3.1. GPT-Generated Survey Construction

To structure the GPT-generated surveys, we adopt a 3×3 construct space that systematically varies both the substantive dimension of the topic and the type of attitude being elicited. This design choice is motivated by established survey methodology research, which emphasizes that social attitudes are multidimensional and are more reliably measured through structured sets of questions rather than single items (Tourangeau et al., 2000; Saris and Gallhofer, 2014; Dillman et al., 2014). By explicitly crossing topical facets with distinct attitudinal orientations, the 3×3 design promotes broader construct coverage while maintaining a controlled and interpretable structure. At the same time, this layout balances expressiveness with cognitive and methodological constraints, avoiding overly long or unstructured surveys while enabling consistent comparison across survey conditions.

GPT-generated surveys were created using a fixed prompting template applied consistently across topics. Survey generation was performed using the OpenAI API with GPT-4.1 model (temperature of 0.7). A system prompt instructed the model to assume the role of an expert survey methodologist and social scientist, with the goal of producing concise, signal-rich surveys suitable for persona discovery and cluster analysis. The user prompt specified the target topic and imposed strict design constraints, including a 3×3 structure spanning beliefs, perceptions, and behaviors; neutral, non-leading wording; and a uniform 5-point Likert response format. Additional constraints required questions to be non-redundant, accessible to a general adult audience, and oriented toward maximizing differentiation between respondents when analyzed jointly. To support interpretability, the model was also instructed to include a brief parenthetical note describing the latent dimension probed by each question. The full prompt template alongside the final survey questions are provided in the supplementary material.

3.2. Human Survey Baseline

To enable meaningful comparison with the GPT-generated surveys, we constructed human survey baselines with a comparable scope and length. Because literature-grounded surveys on social topics often vary in granularity and focus, we combined items from multiple established survey instruments to end up with at least 8 questions. Item selection prioritized alignment with the same high-level constructs targeted in the GPT surveys, namely beliefs, perceptions, and behaviors, while preserving the original wording and intent of each question as used in prior work. This approach allows us to compare surveys of similar length and conceptual coverage, while maintaining fidelity to human-designed instruments grounded in existing social science research. We detail the construction of the human baseline survey for the climate change here. The procedures for the other two topics and the full survey questions are available in the supplementary material. Throughout the paper, we refer to this survey as the human baseline.

For the climate change domain, we constructed the human survey baseline by drawing on widely used and well-validated survey instruments in the literature. The core of the baseline survey is based on the Six Americas Short Survey (SASSY) (Chryst et al., 2018), which has been extensively used to segment public attitudes toward climate change and is supported by a large body of empirical reports and downstream analyses.

To complement this core instrument and expand construct coverage without duplicating question content, we incorporated the Single-Item Measure of Belief in Climate Change (Berger et al., 2025), which captures causal attribution and evaluative belief about climate change in a compact and validated form. We further added two non-redundant items drawn from the American National Election Studies (ANES) (American National Election Studies, 2021), and one question from the climate change perceptions questionnaire (Poortinga et al., 2019) to capture perceptions of the current impacts of climate change, such as its perceived effects on severe weather events and temperature patterns in the United States.

3.3. Response Collection

Participants were recruited via the Cloud Connect platform and were required to be at least 18 years old and residing in the United States. Data collection was conducted using Qualtrics, and the study protocol was approved by our university’s Institutional Review Board (IRB). For each topic (climate change, immigration, and DEI), we recruited approximately 150 participants. Participants were compensated at an hourly rate of \$11; the average

completion time was approximately 6 minutes. No personally identifiable information was collected, and detailed demographic statistics are provided in the supplementary material.

For each topic, participants first reviewed and signed an informed consent form. They were then asked to self-identify their stance using a predefined set of labels accompanied by brief descriptions (full label lists provided in the supplementary material). This self-identification served two purposes: (1) to support downstream analysis of how survey responses cluster relative to self-reported stances, and (2) to facilitate the collection of diverse viewpoints. To this end, once a threshold number of participants was reached for a given stance label, additional participants selecting that label were screened out to maintain a more balanced distribution of perspectives.

Each participant completed both the GPT-generated survey and the corresponding human survey baseline for the topic. Survey order was randomized to control for ordering effects. Following each survey, participants answered a short four-item questionnaire assessing perceived survey quality, including clarity, bias, relevance, and over-specificity.

3.4. Evaluation Metrics

To compare the quality of human baseline and GPT-generated surveys, we employ a set of complementary evaluation metrics that assess latent structure, robustness, diversity preservation, and behavioral relevance. All analyses that require clustering use k -means, a standard unsupervised clustering algorithm, applied to standardized survey responses.

3.4.1. Clustering Quality and Stability

We treat clustering quality as a central aspect of survey quality. In many social science applications, surveys are not used solely to analyze individual items in isolation, but to induce latent groups or typologies of respondents that summarize heterogeneity in beliefs, attitudes, or preferences. Such typological analyses are foundational in fields including political science, sociology, and psychology, where survey responses are routinely clustered to identify ideological blocs, belief systems, or behavioral profiles (Gross and Manrique-Vallier, 2012; Ahlquist and Breunig, 2012; Thomas et al., 2019). From this perspective, a high-quality survey should support coherent, stable, and interpretable group structure, as deficiencies in clustering quality can propagate to downstream tasks such as belief modeling, behavioral prediction, and policy analysis.

We evaluate internal clustering quality using the Silhouette Score (Rousseeuw, 1987) and Davies–Bouldin Index (DBI) (Davies and Bouldin, 1979),

computed on cluster assignments obtained via unsupervised k -means clustering. The Silhouette Score assesses cluster separation at the level of individual respondents by comparing within-cluster cohesion to nearest-cluster separation. For a point i , it is defined as

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

where $a(i)$ is the average distance from i to all other points in the same cluster, and $b(i)$ is the minimum average distance from i to points in any other cluster. The overall Silhouette Score is computed as the mean of $s(i)$ across all N points. The score ranges from -1 to 1 , with higher values indicating stronger separation between clusters and values near zero suggesting overlapping structure.

We also report inertia, which measures overall within-cluster dispersion. Inertia is defined as the sum of squared distances between each point and the centroid $\mu_{c(i)}$ of its assigned cluster:

$$\text{Inertia} = \sum_{i=1}^N \|x_i - \mu_{c(i)}\|^2,$$

where x_i denotes the feature vector for point i and $c(i)$ denotes its cluster assignment. Lower inertia indicates more compact clusters, though it does not account for between-cluster separation.

The DBI evaluates cluster compactness relative to centroid separation and is defined as

$$\text{DBI} = \frac{1}{K} \sum_{k=1}^K \max_{l \neq k} \frac{\sigma_k + \sigma_l}{d(c_k, c_l)},$$

where σ_k denotes within-cluster scatter for cluster k and $d(c_k, c_l)$ denotes the distance between cluster centroids. Lower DBI values indicate more compact and well-separated clusters, while higher values indicate poorer separation. Because these metrics emphasize complementary geometric properties, separation vs. compactness, agreement between them provides stronger evidence of coherent structure than reliance on a single index.

Since internal clustering metrics can be sensitive to sampling variation, we additionally assess *stability under resampling* by repeatedly bootstrapping participants, reclustering with k -means, and computing the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) between runs. The ARI between two partitions U and V is defined as

$$\text{ARI} = \frac{\text{RI} - E[\text{RI}]}{\max(\text{RI}) - E[\text{RI}]}$$

where RI denotes the Rand Index, which measures pairwise agreement between two clusterings by evaluating all pairs of samples and counting the

proportion assigned either to the same cluster in both partitions or to different clusters in both partitions. The ARI corrects this measure for chance agreement. ARI values range from -1 to 1 , with 0 indicating chance-level agreement and 1 indicating identical partitions. Higher ARI under resampling indicates more stable and robust latent structure.

External Structural Agreement. We evaluate the external validity of the induced cluster structure in two complementary ways. First, to assess whether different survey instruments for a single topic recover similar respondent groupings, we apply unsupervised k -means clustering independently to each survey and examine agreement between the resulting partitions using transition matrices. This analysis tests whether the two surveys capture a shared underlying organization of beliefs.

Second, we assess alignment with self-identified persona labels, which participants selected at the beginning of the study to describe their own stance. We compare survey-induced cluster assignments to these self-reported personas using ARI (Hubert and Arabie, 1985) and Normalized Mutual Information (NMI) (Strehl and Ghosh, 2002). Let C_1 denote the clustering obtained from survey responses and C_2 denote the partition induced by participants' self-reported persona labels. NMI is defined as

$$\text{NMI}(C_1, C_2) = \frac{2I(C_1; C_2)}{H(C_1) + H(C_2)},$$

where $I(C_1; C_2)$ measures the mutual information between the two partitions and $H(C_1)$ and $H(C_2)$ denote their respective entropies. NMI ranges from 0 (no shared information) to 1 (identical partitions). Moderate alignment suggests that the derived clusters capture meaningful distinctions while remaining non-redundant with persona categories.

3.4.2. Predictive Validity

A key criterion for evaluating survey quality is the extent to which survey responses capture information that is relevant for downstream behaviors and attitudes. In social science and behavioral research, surveys are routinely used not only to describe beliefs, but to predict consequential outcomes such as political participation, policy support, and social engagement (Ajzen, 1991; Fishbein and Ajzen, 2011; Sciarini and Goldberg, 2016; Lazarsfeld and Franzen, 1945). From this perspective, predictive performance provides a direct test of whether a survey encodes behaviorally meaningful signal. We assess predictive validity by evaluating how well survey responses and survey-induced clusters predict downstream outcomes, including self-reported policy support, action likelihood, and social discussion. Cluster-based predictors use k -means

assignments as discrete representations of latent structure. Predictive performance is measured using cross-validated R^2 , defined as

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}.$$

Comparing feature-based models with cluster-based models reveals how much predictive signal is preserved under unsupervised compression.

4. Results

Due to space constraints, we focus our empirical evaluation on the climate change surveys, which serve as a representative testbed for comparing human-designed and GPT-generated instruments. Results for the remaining two topics follow similar patterns unless otherwise noted and are reported in the supplementary material. All reported results are computed using the same set of respondents, each of whom completed both the human baseline survey and the GPT-generated survey on climate change. We begin our analysis by examining mean responses to each survey question, stratified by self-identified climate personas. For the climate change topic, participants were asked to select one of the following labels: **Advocate for Avoiding Climate Alarmism**, **Climate Policy Skeptic**, **Neutral Observer of Climate Change**, **Advocate for Balanced Climate Action**, **Advocate for Proactive Climate Policy**, and **Climate Action Advocate**. These personas were designed to capture ordered positions along a spectrum of climate beliefs, reflecting increasing levels of concern about climate change and stronger support for mitigation and policy intervention. Each label was accompanied by a brief description (see Supplementary Material), and wording was deliberately framed in neutral or positive terms to avoid discouraging participants from selecting any category.

Figure 1 shows the mean response for each question in the human baseline survey (top) and the GPT-generated survey (bottom), with bars colored by persona. Across nearly all questions in both surveys, mean responses exhibit a clear and ordinally meaningful pattern: as one moves along the persona spectrum toward more climate-concerned identities, mean responses change monotonically in the expected direction.

The monotonic structure indicates that both surveys capture a coherent underlying belief gradient rather than noisy variation. The human baseline shows more consistent spacing between personas, suggesting stable differentiation, whereas the GPT-generated survey exhibits attenuated separation across some questions, indicating partial merging of adjacent personas.

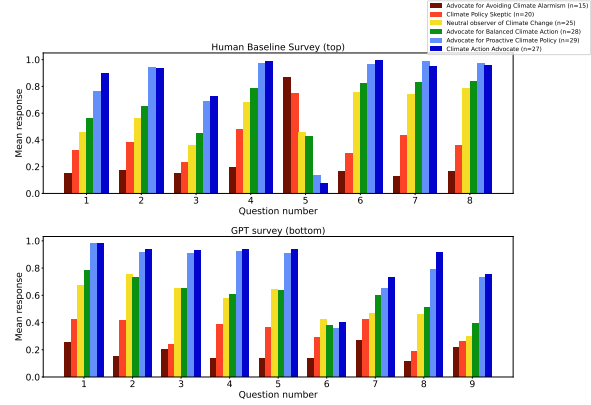


Figure 1: Climate change surveys: Normalized mean response per question by persona label. The top panel shows the human baseline survey and the bottom panel shows the GPT-generated survey. Bars are colored by self-identified climate persona, ordered from least to most climate-concerned. Across both surveys, mean responses vary monotonically along the persona spectrum, indicating an ordinally meaningful alignment between survey responses and underlying climate belief positions.

Overall, these results validate both the self-identified personas and the survey instruments, showing that persona labels correspond to systematic response patterns and that both surveys align with these distinctions.

4.1. Clustering Quality and Stability

We next examine the number of clusters supported by each survey using internal clustering metrics. Based on both the Silhouette Score and the DBI, the optimal number of clusters is $k = 2$ for both the human baseline and the GPT-generated climate change surveys (as indicated by the Silhouette curve in Figure 2). Although we do not include DBI plots due to space constraints, both metrics consistently indicate that a two-cluster solution yields the most compact and well-separated structure for both surveys.

Although this may seem inconsistent with the six self-identified personas, we do not view it as contradictory. Instead, both surveys appear to capture a dominant, coarse-grained division in climate beliefs, consistent with prior work showing that attitudes often exhibit hierarchical structure (Goldberg, 2006). Our initial analyses indicate that each of the two major clusters can be further subdivided into more nuanced subgroups using hierarchical or probabilistic approaches such as Latent Profile Analysis (LPA) (Masyn, 2013). Exploring such multi-level structure is a promising direction for future work; in the remainder of this paper, we focus on the two-cluster solution.

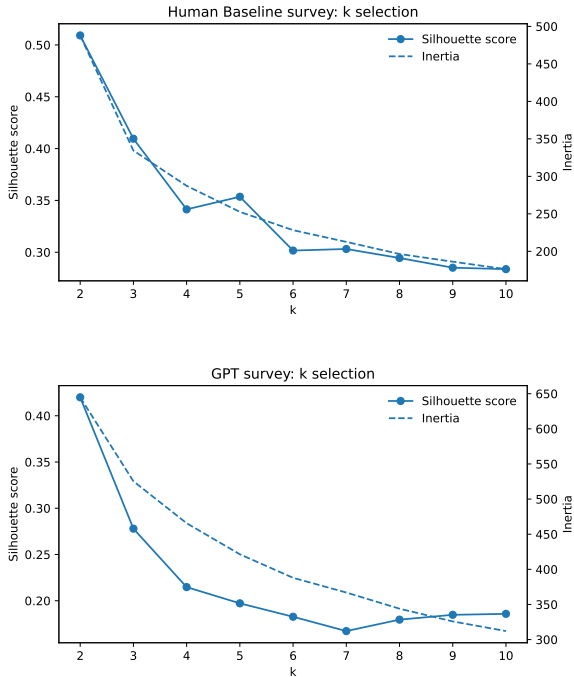


Figure 2: Selection of the number of clusters (k) using silhouette score and inertia for the human baseline(top) and GPT-generated(bottom) surveys.

We assess the robustness of the induced cluster structure using stability under resampling. For each survey, we perform 300 bootstrap runs, each time sampling 80% of respondents without replacement and reclustering using k -means with $k = 2$. Stability is quantified using the ARI between cluster assignments across runs.

Both surveys exhibit high stability. The human baseline survey achieves a mean ARI of 0.87 (median = 1.00), with the central 80% of values ranging from 0.58 to 1.00. The GPT-generated survey exhibits slightly higher average stability, with a mean ARI of 0.93 (median = 0.92) and a 10th–90th percentile range of 0.80 to 1.00. While both surveys occasionally produce lower-agreement runs, as reflected in minimum ARI values, the predominance of high ARI scores indicates that the two-cluster solutions are robust to substantial perturbations of the data. These results suggest that the dominant two-cluster structure recovered by both surveys reflects a stable latent division in climate beliefs rather than an artifact of sampling variability.

We next examine the degree to which the human baseline and GPT-generated surveys recover similar latent group structure. This analysis is important because high between-survey agreement indicates that both surveys are capturing a shared underlying belief organization, whereas low agreement may signal measurement distortion or instrument-specific artifacts.

Human Baseline/GPT	Cluster 0	Cluster 1
Cluster 0	0.86	0.14
Cluster 1	0.02	0.98

Table 1: Transition matrix between clusters induced by the human baseline and the GPT-generated surveys. Rows correspond to human baseline clusters columns correspond to GPT clusters.

The transition matrix in Table 1 shows strong agreement between the two surveys. Most respondents remain on the diagonal (86% and 98%), with minimal cross-cluster reassignment. This indicates that the GPT-generated survey recovers the same primary division in climate beliefs identified by the human baseline, with only limited redistributions.

We evaluate external semantic validity by examining how unsupervised cluster assignments align with persona labels that reflect respondents’ positions on climate change. As our primary analysis, we focus on alignment with self-identified personas, which were collected directly from participants and represent positions along a climate belief spectrum.

Using ARI and NMI, we observe modest but systematic alignment between clusters and self-identified personas. The human baseline survey achieves an ARI of 0.12 and an NMI of 0.22, while the GPT-generated survey achieves slightly higher alignment (ARI = 0.15, NMI = 0.27).

Because persona labels include six categories while clustering yields only two groups, perfect alignment is neither expected nor feasible. The low ARI largely reflects this difference in granularity, whereas the non-trivial NMI indicates shared information beyond chance, suggesting both surveys capture a dominant ideological division underlying finer persona distinctions.

Inspection of cluster–persona distributions reveals a clear qualitative pattern. In both surveys, one cluster is enriched for self-identified personas expressing higher climate concern (e.g., *Advocate for Proactive Climate Policy* and *Climate Action Advocate*), while the other cluster is enriched for more skeptical or disengaged personas (e.g., *Climate Policy Skeptic* and *Advocate for Avoiding Climate Alarmism*), with neutral personas distributed between clusters. This structure suggests that unsupervised clustering recovers a meaningful ideological ordering.

Taken together, these results suggest that both survey instruments capture socially meaningful climate belief structure without trivially reproducing persona labels. Alignment increases as persona labels are coarsened, indicating that both surveys reliably recover a dominant ideological division, while differences between instruments emerge primarily in how much finer-grained nuance they preserve beyond this shared structure.

Survey / R^2	Policy (Feat.)	Policy (Clust.)	Action (Feat.)	Action (Clust.)
Climate Change				
GPT	0.486	0.368	0.158	0.176
Human Baseline	0.554	0.426	0.391	0.144
Immigration				
GPT	0.345	0.154	0.537	0.488
Human Baseline	0.285	0.313	0.525	0.287
DEI				
GPT	0.504	0.450	0.497	0.421
Human Baseline	0.501	0.592	0.331	0.244

Table 2: Predictive performance (R^2 , mean 5-fold cross-validated values) using Ridge regression for feature-based (Feat.) and cluster-based (Clust.) models across all three topics.

4.2. Predictive Validity

We evaluate predictive validity by assessing how well each survey predicts two downstream outcomes: support for policy and engagement in taking action. For each outcome, we first remove survey items that directly measure that outcome (e.g., policy-support questions when predicting policy support, action-oriented questions when predicting engagement) to prevent information leakage. The remaining survey questions are used as predictive features. In parallel, clusters are derived using only these non-outcome questions with similar clustering method as section 4.1. For each outcome, we compare models trained on survey responses (feature-based models) to models that use only the induced cluster label as a predictor (cluster-based models). All models are estimated using Ridge regression, and predictive performance is measured using 5-fold cross-validated R^2 . We report results for all three topics since predictive performance varies by domain and modeling choice rather than favoring a single instrument uniformly (see Table 2).

For climate change, feature-based models using the human baseline survey outperform those using the GPT-generated survey for both policy support ($R^2 = 0.554$ vs. 0.486) and action likelihood ($R^2 = 0.391$ vs. 0.158). However, cluster-based performance narrows this gap, with GPT clusters slightly outperforming human clusters for action (0.176 vs. 0.144), suggesting both surveys recover a similar coarse behavioral division.

For Immigration, the pattern shifts. GPT feature-based models slightly outperform the human baseline survey for policy prediction (0.345 vs. 0.285), while action prediction is comparable at the feature level (0.537 vs. 0.525). Notably, GPT clusters substantially outperform human baseline clusters for action prediction (0.488 vs. 0.287), indicating stronger segmentation for this domain.

For DEI, feature-based policy prediction is nearly identical (0.504 vs. 0.501), while GPT features yield

higher predictive performance for action (0.497 vs. 0.331). At the cluster level, the pattern reverses for policy: human baseline clusters outperform GPT clusters (0.592 vs. 0.450), whereas GPT clusters perform better for action (0.421 vs. 0.244), revealing domain-dependent differences in how belief structure translates to behavioral outcomes.

Overall, these results indicate that both survey types recover behaviorally relevant structure, but their predictive strengths vary by topic and by level of representation (item-level versus cluster-level). Rather than demonstrating uniform superiority of one instrument, the findings suggest domain-dependent differences in how belief structure translates into downstream behavioral prediction. Future work should focus on understanding the mechanisms underlying these differences. Identifying these sources of divergence may enable principled refinement of LLM-generated survey instruments and help close predictive gaps where they emerge.

5. Discussion

Across three domains, GPT-generated surveys recover the dominant ideological structure identified by human baseline instruments. Clustering quality, stability, and cross-survey agreement indicate similar high-level respondent partitions. Differences arise primarily in how belief variation is represented beyond this dominant split. Dimensional efficiency and predictive analyses suggest that human baseline surveys often preserve a stronger fine-grained signal at the item level, while GPT-generated surveys can produce competitive segmentation in certain domains. Alignment with self-identified personas suggests that both surveys capture a shared ideological backbone underlying more nuanced distinctions. Further examination of this layered structure remains an important direction for future work.

These findings suggest that divergence between instruments lies less in the existence of structure and more in its resolution and translation into downstream prediction. Future work should examine how prompt design influences survey quality, evaluate generalization across additional topics, explore finer-grained clustering solutions, and investigate when predictive gaps are substantively meaningful.

Overall, LLM-generated surveys show strong promise for recovering coarse belief structure. However, the current evidence does not support their use as full replacements for expert-designed instruments without human oversight. In particular, our evaluation is limited to three social domains, and performance may vary for more technical, less salient, or more culturally specific topics. Instead, LLM-generated surveys may serve as useful scaffolds for exploratory analysis, pilot studies, or supervised survey development.

6. Acknowledgements

This work was partially supported by a Stony Brook University Climate Change Seed Grant. We gratefully acknowledge the Department of Political Science at Stony Brook University for providing funding for the crowdsourced user studies.

7. Ethical Considerations

This study involved human participants and was approved by the relevant Institutional Review Board (IRB). All participants provided informed consent prior to participation and were recruited through an online research platform. Participants were required to be 18 years of age or older and residing in the United States. No personally identifiable information was collected, and all analyses were conducted on anonymized data. Participants were compensated at a rate aligned with platform norms.

The survey topics examined (climate change, immigration, and DEI) are socially and politically sensitive. To minimize potential harm, survey questions were designed to be neutral, non-leading, and accessible to a general adult audience. Participants were free to discontinue participation at any time without penalty.

Because this study evaluates LLM-generated survey instruments, we also note the potential risks associated with deploying automated surveys in real-world settings. LLM-generated instruments may inadvertently reflect biases present in training data or oversimplify complex belief structures. Our findings do not advocate replacing expert-designed surveys with fully automated systems; rather, we emphasize the importance of human oversight when using LLMs in survey design and interpretation.

8. Bibliographical References

Christopher L Aberson. 2007. Diversity, merit, fairness, and discrimination beliefs as predictors of support for affirmative-action policy actions 1. *Journal of Applied Social Psychology*, 37(10):2451–2474.

Divya Mani Adhikari, Alexander Hartland, Ingmar Weber, and Vikram Kamath Cannanure. 2025. Exploring llms for automated generation and adaptation of questionnaires. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, pages 1–23.

John S Ahlquist and Christian Breunig. 2012. Model-based clustering and typologies in the social sciences. *Political Analysis*, 20(1):92–112.

Icek Ajzen. 1991. The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2):179–211.

American National Election Studies. 2021. [Anes 2020 time series study](#).

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Sebastian Berger, David Hauser, Anna Lange, and Sander van der Linden. 2025. Measuring belief in climate change with a single item. *Global Environmental Psychology*, 3:1–22.

James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. 2024. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416.

Petra M Boynton and Trisha Greenhalgh. 2004. Selecting, designing, and developing your questionnaire. *Bmj*, 328(7451):1312–1315.

Breanne Chryst, Jennifer Marlon, Sander Van Der Linden, Anthony Leiserowitz, Edward Maibach, and Connie Roser-Renouf. 2018. Global warming’s “six americas short survey”: Audience segmentation of climate change views using a four question instrument. *Environmental Communication*, 12(8):1109–1122.

Lee J Cronbach and Paul E Meehl. 1955. Construct validity in psychological tests. *Psychological bulletin*, 52(4):281.

David L. Davies and Donald W. Bouldin. 1979. [A cluster separation measure](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

Don A Dillman, Jolene D Smyth, and Leah Melani Christian. 2014. *Internet, phone, mail, and mixed mode surveys: The tailored design method*. John Wiley & Sons Inc.

Alice H Eagly and Shelly Chaiken. 1993. *The psychology of attitudes*. Harcourt brace Jovanovich college publishers.

Martin Fishbein and Icek Ajzen. 2011. *Predicting and changing behavior: The reasoned action approach*. Psychology press.

Lewis R Goldberg. 2006. Doing it all backwards: The development of hierarchical factor structures from the top down. *Journal of Research in Personality*, 40(4):347–358.

- Lara M Greaves, Danny Osborne, and Chris G Sibley. 2015. Profiling the fence-sitters in new zealand elections: A latent profile model of political voting blocs. *New Zealand Journal of Psychology*, 44(2):43.
- Justin H Gross and Daniel Manrique-Vallier. 2012. A mixed-membership approach to the assessment of political ideology from survey responses. In *Individual Presentation, Society for Political Methodology, 29th Annual Summer Meeting, Chapel Hill, NC. Citeseer*.
- Robert M Groves, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau. 2011. *Survey methodology*. John Wiley & Sons.
- Perttu Hämäläinen, Mikke Tavast, and Anton Kunari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19.
- Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification*, 2(1):193–218.
- Bernard J Jansen, Soon-gyo Jung, and Joni Salmiinen. 2023. Employing large language models in survey research. *Natural Language Processing Journal*, 4:100020.
- David Johann, Markus Steinbrecher, and Kathrin Thomas. 2020. Channels of participation: Political participant types and personality. *PloS one*, 15(10):e0240671.
- Darsel Keane and Rod B McNaughton. 2025. Using generative ai to enhance psychometric scale development in market research. *International Journal of Market Research*, page 14707853251384769.
- Bärbel Knäuper, Robert F Belli, Daniel H Hill, and A Regula Herzog. 1997. Question difficulty and respondents' cognitive ability: The effect on data quality. *JOURNAL OF OFFICIAL STATISTICS-STOCKHOLM-*, 13:181–199.
- Paul F Lazarsfeld and Raymond H Franzen. 1945. Prediction of political behavior in america. *American Sociological Review*, 10(2):261–273.
- Anthony Leiserowitz, Connie Roser-Renouf, Jennifer Marlon, and Edward Maibach. 2021. Global warming's six americas: a review and recommendations for climate change communication. *Current Opinion in Behavioral Sciences*, 42:97–103.
- Timo Lenzner. 2012. Effects of survey question comprehensibility on response quality. *Field Methods*, 24(4):409–428.
- Edward W Maibach, Anthony Leiserowitz, Connie Roser-Renouf, and CK Mertz. 2011. Identifying like-minded audiences for global warming public engagement campaigns: An audience segmentation analysis and tool development. *PloS one*, 6(3):e17571.
- Antonio Maiorino, Zoe Padgett, Chun Wang, Misha Yakubovskiy, and Peng Jiang. 2023. Application and evaluation of large language models for the generation of survey questions. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5244–5245.
- Katherine E. Masyn. 2013. Latent class analysis and finite mixture modeling. In Todd D. Little, editor, *The Oxford Handbook of Quantitative Methods*. Oxford University Press.
- Rosemary McEachan, Natalie Taylor, Reema Harrison, Rebecca Lawton, Peter Gardner, and Mark Conner. 2016. Meta-analysis of the reasoned action approach (raa) to understanding health behaviors. *Annals of Behavioral Medicine*, 50(4):592–612.
- NORC at the University of Chicago. 2023. [General social survey 2022](#).
- Lydia Oeljeklaus, Stefan Höft, and Daniel Danner. 2025. Comparing psychometric properties of expert-developed and ai-generated personality scales. *Psychological Test Adaptation and Development*.
- Jennifer Oser. 2022. Protest as one political act in individuals' participation repertoires: Latent class analysis and political participant types. *American Behavioral Scientist*, 66(4):510–532.
- Wouter Poortinga, Lorraine Whitmarsh, Linda Steg, Gisela Böhm, and Stephen Fisher. 2019. Climate change perceptions and their individual-level determinants: A cross-european analysis. *Global environmental change*, 55:25–35.
- David M Rothschild, James Brand, Hope Schroeder, and Jenny Wang. 2024. Opportunities and risks of llms in survey research. *Available at SSRN*.
- Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Willem E Saris and Irmtraud N Gallhofer. 2014. *Design, evaluation, and analysis of questionnaires for survey research*. John Wiley & Sons.

Howard Schuman and Stanley Presser. 1996. *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage.

Pascal Sciarini and Andreas C. Goldberg. 2016. [Turnout bias in postelection surveys: Political involvement, survey participation, and vote overreporting](#). *Journal of Survey Statistics and Methodology*, 4(1):110–137.

Alexander Strehl and Joydeep Ghosh. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617.

Nicolaos E Synodinos. 2003. The “art” of questionnaire construction: some important considerations for manufacturing studies. *Integrated manufacturing systems*, 14(3):221–237.

John Terry, Gerald Strait, Steve Alsarraf, Emily Weinmann, and Allison Waychoff. 2025. Artificial intelligence in scale development: evaluating ai-generated survey items against gold standard measures. *Current Psychology*, 44(20):16339–16350.

Emma F Thomas, Simon M Bury, Winnifred R Louis, Catherine E Amiot, Pascal Molenberghs, Monique F Crane, and Jean Decety. 2019. Vegetarian, vegan, activist, radical: Using latent profile analysis to examine different forms of support for animal welfare. *Group Processes & Intergroup Relations*, 22(6):836–857.

Roger Tourangeau, Lance J Rips, and Kenneth Rasinski. 2000. *The psychology of survey response*. Cambridge University Press.

Sebastián Valenzuela, Stephan Winter, and Sebastián Rivera. 2025. Using large language models for survey research in communication: opportunities and challenges. *Communication and Change*, 1(1):14.

Alexander Wuttke, Matthias Aßenmacher, Christopher Klamm, Max Lang, and Fraue Kreuter. 2025. Ai conversational interviewing: Transforming surveys with llms as adaptive interviewers. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 179–204.

A. Full GPT Prompts

This appendix provides the complete prompt templates and generation parameters used to produce the GPT-generated surveys analyzed in this study. We include the full system and user prompts verbatim to support transparency, reproducibility, and careful interpretation of the survey generation process.

```
MODEL = "gpt-4.1"
TEMPERATURE = 0.7
MAX_TOKENS = 800

SYSTEM_PROMPT = ""
You are an expert survey methodologist
and social scientist.
Your task is to design concise,
signal-rich surveys suitable for
persona discovery and cluster analysis.
""

USER_PROMPT_TEMPLATE = ""
I want to understand how people think,
feel, and act with respect to the
following topic:

Topic: "{topic}"

Design a concise, signal-rich survey
that captures the major latent
dimensions of human stance on this
topic.

Survey requirements:
1. The survey must contain exactly 9
questions arranged as a 3 by 3
structure:
- 3 belief / cognitive questions
- 3 perception / affect questions
- 3 behavior / intention / policy
questions

2. Each question must:
- Be answerable on a 5-point Likert
scale
- Be neutral and non-leading
- Be understandable to a general
adult audience
- Avoid technical jargon

3. The questions should be:
- Non-redundant
- Designed to maximize
differentiation between
respondents
- Capable of revealing distinct
personas when analyzed jointly

Output format:
```

- Group questions into three sections (Beliefs, Perceptions, Behaviors)
- Number questions from 1 to 9
- For each question, include a brief parenthetical note describing the latent dimension it probes

Only output the survey. Do not include analysis or commentary.

""

B. Full Survey Questions and Self-Identified Labels

B.1. Climate Change

B.1.1. Persona Labels and Descriptions

The following self-identification question was presented to the users after signing the consent form.

Which of the following labels and descriptions best reflect your view on climate change?

- **Advocate for Avoiding Climate Alarmism**
You are a person who does not believe that climate change is a real or serious issue, particularly not one caused by human activity. You believe climate-related policies are unnecessary and may harm economic growth or personal freedoms. You generally oppose environmental regulations and favor deregulation.
- **Climate Policy Skeptic**
You are a person who believes climate change is happening but questions the degree of threat and the effectiveness of government-led solutions. You may worry that some climate policies go too far or could harm economic opportunity and personal freedoms.
- **Neutral observer of Climate Change**
You are someone who recognizes that environmental concerns exist but doesn't follow climate debates closely. You may adopt eco-friendly practices when convenient, but climate action is not a central concern in your daily life.
- **Advocate for Balanced Climate Action**
You are a person who believes climate change is real and worth addressing, but solutions should be practical and economically responsible. You support action that balances environmental goals with innovation, affordability, and individual freedom.
- **Advocate for Proactive Climate Policy**
You believe climate change is real and requires action, but think solutions should balance environmental responsibility with economic and social considerations. You support investments

in renewable energy, resilience, and policy reform that is both effective and sustainable.

- **Climate Action Advocate**

You are someone who sees climate change as an urgent challenge requiring bold action. You support ambitious policy reforms, investment in renewables, and efforts to hold corporations accountable. You may also adopt lifestyle changes to reduce your personal impact and promote environmental responsibility.

B.1.2. Human Baseline Survey

The human baseline survey consisted of items drawn from established instruments. The first four questions were adapted from SASSY (Chryst et al., 2018), and Question 5 was taken from the Climate Change Perception Questionnaire (Poortinga et al., 2019). These were followed by items V241749 and V242321 from the ANES survey (American National Election Studies, 2021), as well as a single-item measure of belief in climate change (Berger et al., 2025). The resulting survey items are listed below. Question 6 was one of the questions used in section 4.2 for predicting policy support.

1. How important is the issue of climate change to you personally?
1. Not at all important, 2. a little important, 3. moderately important, 4. very important, 5. extremely important
2. How worried are you about global warming?
Very worried 2. Somewhat worried 3. Not very worried 4. Not at all worried
3. How much do you think global warming will harm you personally?
1. A great deal 2. A moderate amount 3. Only a little 4. Not at all
4. How much do you think global warming will harm future generations of people?
1. not at all, 2. only a little, 3. a moderate amount, 4. a great deal
5. How good or bad do you think the impact of climate change would be on people across the world?
1. Extremely bad, 2. Bad, 3. Neither good or bad, 4. Good, 5. Extremely good
6. Do you favor, oppose, or neither favor nor oppose increased government regulation on businesses that produce a great deal of greenhouse emissions linked to climate change?
1. Favor, 2. Oppose, 3. Neither favor nor oppose

For each of the following statements, please indicate the extent to which you agree or disagree using the following scale: 1. Strongly Agree 2. Somewhat Agree 3. Neutral 4. Somewhat Disagree 5. Strongly Disagree.

7. Climate change is currently affecting severe weather events or temperature patterns in the United States.
8. The occurrence of climate change is caused by human activities and will bring largely negative consequences.

B.1.3. GPT Survey

The resulting survey items from the prompt provided in section A are listed below. Questions 7 and 8 were questions used in section 4.2 for predicting policy support and downstream action.

For each statement below, please indicate the extent to which you agree or disagree. Please respond honestly. There are no right or wrong answers.

Response scale (for the next 6 items below): A great deal; A lot; A moderate amount; A little; None at all.

1. How much do you agree that climate change is happening?
2. To what extent do you think human activities contribute to climate change?
3. How confident are you in the scientific information available about climate change?
4. How concerned do you feel about the effects of climate change on the planet?
5. How serious do you believe the consequences of climate change are for people in your country?
6. How hopeful do you feel about society's ability to address climate change?
7. How often do you take actions to reduce your own environmental impact (such as conserving energy, reducing waste, or choosing sustainable products)?
 1. Always 2. Often 3. Sometimes 4. Rarely 5. Never.
8. How likely are you to support policies or laws aimed at reducing climate change, even if they have economic costs?
 1. Extremely likely 2. Very likely 3. Moderately likely 4. Slightly likely 5. Not at all likely.

9. How likely are you to discuss climate change or related topics with others in your social circle?
 1. Extremely likely 2. Very likely 3. Moderately likely 4. Slightly likely 5. Not at all likely.

B.2. Immigration

B.2.1. Persona Labels and Descriptions

The following self-identification question was presented to the users after signing the consent form. **Which of the following labels and descriptions best reflect your view on Immigration?**

- **Advocate for Border Integrity**
You prioritize secure borders, the enforcement of immigration laws, and the preservation of national identity. You believe immigration should be tightly controlled.
- **Advocate for Selective Immigration**
You support legal immigration that serves the country's strategic goals. You favor policies that emphasize skills, economic contribution, or national security, while being cautious about broad or open-ended immigration reforms.
- **Neutral Observer**
You recognize immigration as a complex issue and prefer not to take a fixed stance. You believe immigration policies should be shaped by context, evidence, and evolving societal needs rather than ideology.
- **Advocate for Balanced Immigration**
You support immigration policies that balance national interests with humanitarian values. You favor fair and efficient systems that welcome newcomers while maintaining social cohesion and economic stability.
- **Pragmatic Immigration Supporter**
You support humane and efficient immigration processes that weigh both national interests and human dignity. You believe reforms should be practical and inclusive, ensuring economic contribution while offering protection to those in need.
- **Advocate for Global Mobility**
You are a person who strongly supports welcoming immigration policies that offer opportunities to those seeking a better life. You advocate for inclusive, well-resourced systems that help immigrants integrate into society and contribute fully.

B.2.2. Human Baseline Survey

For the immigration domain, we relied on items drawn from ANES ([American National Election](#)

Studies, 2021), which includes a broad set of questions capturing multiple aspects of public attitudes toward immigration, including perceived impacts, fairness, and policy preferences. From this item pool, we selected nine non-redundant questions (specifically ANES items V201424, V201427, V202240, V202419, V202246, V202420, V202418, V201417, and V201421) to construct a human survey baseline with coverage across belief, perception, and behavior-oriented dimensions. This selection allowed us to preserve the diversity of immigration-related attitudes represented in ANES while matching the length and overall structure of the GPT-generated surveys. The resulting survey items are listed below. Questions 8 and 9 were the questions used in section 4.2 for predicting policy support.

For each statement below, please indicate the extent to which you agree or disagree. Please respond honestly. There are no right or wrong answers.

Response scale (for the next seven items below): Strongly Agree; Somewhat Agree; Neutral; Somewhat Disagree; Strongly Disagree.

1. A wall should be built at the U.S. border with Mexico.
2. It is important that everyone in the United States learn to speak English.
3. There should be a path to citizenship for unauthorized immigrants who obey the law, pay a fine, and pass security checks.
4. America's culture is generally harmed by immigrants.
5. The children of detained immigrants should be separated rather than kept with their parents in adult detention centers.
6. Immigrants increase crime rates in the United States.
7. Immigrants are generally good for America's economy.
8. Which comes closest to your view about what government policy should be toward unauthorized immigrants now living in the United States?
 - Make all unauthorized immigrants felons and send them back to their home country.
 - Have a guest worker program that allows unauthorized immigrants to remain in the United States in order to work, but only for a limited amount of time.

- Allow unauthorized immigrants to remain in the United States and eventually qualify for U.S. citizenship, but only if they meet certain requirements like paying back taxes and fines, learning English, and passing background checks.
- Allow unauthorized immigrants to remain in the United States and eventually qualify for U.S. citizenship, without penalties.

9. What should happen to immigrants who were brought to the U.S. illegally as children and have lived here for at least 10 years and graduated high school here?

- Should be sent back where they came from.
- Should be allowed to live and work in the U.S.

B.2.3. GPT Survey

The resulting survey items from the prompt provided in section A are listed below. Questions 7, 8, and 9 were questions used in section 4.2 for predicting policy support and downstream action.

For each statement below, please indicate the extent to which you agree or disagree. Please respond honestly. There are no right or wrong answers. Response scale (for all items below): Strongly Agree; Somewhat Agree; Neutral; Somewhat Disagree; Strongly Disagree.

1. Immigration has a positive effect on the economy of the receiving country.
2. The culture of this country is strengthened by the presence of people from different backgrounds.
3. The number of immigrants entering my country in recent years is higher than is good for society.
4. I feel comfortable interacting with people who have recently moved here from other countries.
5. News stories about immigration often make me feel anxious or concerned.
6. I worry that immigration could make it harder for me or my family to find opportunities.
7. I would support policies that make it easier for people from other countries to live and work here.
8. I would be willing to sign a petition supporting the rights of immigrants.
9. I believe the government should prioritize strict enforcement of immigration laws.

B.3. DEI

B.3.1. Persona Labels and Descriptions

The following self-identification question was presented to the users after signing the consent form. **Which of the following labels and descriptions best reflect your view on DEI?**

DEI stands for Diversity, Equity, and Inclusion: principles aimed at promoting fair treatment and meaningful participation for all people, especially those historically marginalized or underrepresented.

- **Advocate for Meritocracy**

You are someone who believes that fairness means treating everyone the same, regardless of identity. You are cautious about DEI initiatives that prioritize group-based outcomes, as you worry they may compromise neutrality or create unintended divisions. You prefer merit-based systems and evaluate policies based on their effectiveness and broad impact.

- **Advocate for Individual-Centered Fairness**

You value fairness on an individual level, believing that opportunities should be tailored to people's qualifications and efforts rather than group membership. You may support inclusive efforts when they focus on shared human dignity and not identity-based metrics.

- **Neutral Observer** You are someone who hasn't thought much about DEI and doesn't feel strongly either way. You believe in treating everyone fairly, but DEI isn't something you follow closely or prioritize. You tend to focus on your day-to-day responsibilities and judge DEI efforts based on what you experience locally.

- **Advocate for Balanced Opportunity** You support efforts to promote fairness and inclusion, but believe they should be measured and context-sensitive. You favor approaches that expand access without compromising merit, cohesion, or organizational effectiveness.

- **Advocate for Inclusive Culture** You are someone who believes that creating a respectful, inclusive culture helps everyone thrive. You support structured DEI efforts—like internal training, inclusive practices, and broader representation—as essential for fostering belonging, engagement, and team success.

- **Advocate for Structural Equity** You are someone who is deeply committed to systemic equity. You work to dismantle institutional barriers and amplify marginalized voices. You advocate for transparent policies, structural reform, and sustained investment to ensure long-term, accountable change.

B.3.2. Human Baseline Survey

For the DEI domain, we reviewed several established survey instruments measuring attitudes toward affirmative action, diversity, merit, and discrimination, including debate-style items from large-scale surveys such as the General Social Survey (GSS) (NORC at the University of Chicago, 2023) and ANES (American National Election Studies, 2021), as well as multi-item scales from the social psychology literature. While GSS and ANES items are widely used, they primarily capture overall support or opposition through single questions with limited construct differentiation.

For the human survey baseline, we therefore drew on the Diversity, Merit, Fairness, and Discrimination (DMFD) Belief Scales (Aberson, 2007), which decompose DEI-related attitudes into multiple interpretable constructs using parallel item structures. To ensure comparability with the GPT-generated surveys, we selected three out of four constructs, resulting in a nine-item instrument matched in length and overall structure to the GPT-generated 3×3 surveys. The resulting survey items are listed below. Questions 1, 2, and 3 were the questions used in section 4.2 for predicting policy support.

For each statement, please indicate the extent to which you agree or disagree. Please respond honestly. There are no right or wrong answers.

Affirmative action refers to policies or practices designed to increase opportunities for historically underrepresented groups in areas such as education and employment. These measures may include considering factors like race, ethnicity, or gender as part of admissions or hiring decisions, with the goal of promoting diversity and addressing past discrimination.

Response scale (for all items below): Strongly Agree; Somewhat Agree; Neutral; Somewhat Disagree; Strongly Disagree.

1. Affirmative-action hiring policies are fair.
2. Affirmative-action policies give everyone an equal chance.
3. Affirmative-action policies are unfair to White men.
4. Only the most qualified applicant should be hired, regardless of race or gender.
5. Hiring decisions should be based solely on merit.
6. Considering race or gender in hiring violates the principle of merit.
7. A diverse workforce benefits an organization.

8. Employees from different backgrounds improve problem solving.
9. Diversity makes companies stronger.
10. Racial discrimination in hiring is still common today.
11. Minorities do not yet have equal job opportunities.
12. Many employers would refuse a qualified minority applicant.

B.3.3. GPT Survey

The resulting survey items from the prompt provided in section A are listed below. Questions 7, 8, and 9 were questions used in section 4.2 for predicting policy support and downstream action.

For each statement below, please indicate the extent to which you agree or disagree. Please respond honestly. There are no right or wrong answers.

Response scale (for all items below): Strongly Agree; Somewhat Agree; Neutral; Somewhat Disagree; Strongly Disagree.

1. Having people from different backgrounds improves group performance.
2. Efforts to promote fairness and inclusion are necessary to address existing inequalities.
3. Opportunities should be based solely on individual merit, regardless of background or identity.
4. I feel comfortable interacting with people whose backgrounds or perspectives are different from my own.
5. I perceive that people like me are treated fairly in most settings (e.g., work, school, public spaces).
6. I often notice unfair treatment or bias based on identity (such as race, gender, ability, etc.) in my everyday life.
7. I am likely to support policies or programs that aim to increase representation of underrepresented groups.
8. I am likely to speak up or take action if I witness discrimination or unfair treatment.
9. I actively seek out information or experiences that expose me to different perspectives.

C. User Demographic Information

We recruited participants online for each topic-specific survey. After data cleaning and attention checks, the final sample consisted of 144 respondents for the climate change survey, 170 for the immigration survey, and 167 for the DEI survey. Participants were restricted to U.S.-based respondents and varied across gender, race/ethnicity, education level, and age group. The demographic distribution for each topic is summarized in Table 3.

Demographic (%)	Climate	Immigration	DEI
Gender			
Female	50.7	44.7	52.1
Male	48.6	52.9	46.1
Other	0.7	2.4	1.8
Race / Ethnicity			
White	72.2	78.2	72.5
Black or African American	11.8	10	13.8
Asian	8.3	4.7	6.6
Other / Multiracial	7.7	7.1	7.1
Education			
High school or less	15.2	17.1	13.4
Some college	30.6	34.7	30.6
Bachelor's degree	38.9	30.6	35
Graduate degree	15.3	17.6	21
Age Group			
18–24	9	11.2	7.8
25–34	24.3	26.5	29.9
35–44	29.2	32.4	28.7
45–54	13.9	14.1	12
55+	23.6	15.8	21.6

Table 3: Participant demographic characteristics (Climate: $N = 144$; Immigration: $N = 170$; DEI: $N = 167$).

D. Results

D.1. Immigration

Figure 3 presents the mean response for each question in the human baseline survey (top) and the GPT-generated survey (bottom), with bars colored by immigration persona. Across most questions in both surveys, responses exhibit a clear and ordinal pattern: as one moves along the persona spectrum (from *Advocate for Border Integrity* to *Advocate for Global Mobility*) mean responses shift systematically in directions consistent with increasing support for immigration.

This progression indicates that both instruments capture a meaningful underlying gradient of immigration attitudes rather than purely idiosyncratic variation. In the human baseline survey, persona means generally follow the expected ordering across items, with relatively clear separation between opposing ends of the spectrum, though spacing between adjacent personas varies by question. The GPT-generated survey reproduces the broad ordinal structure across most items, but the dis-

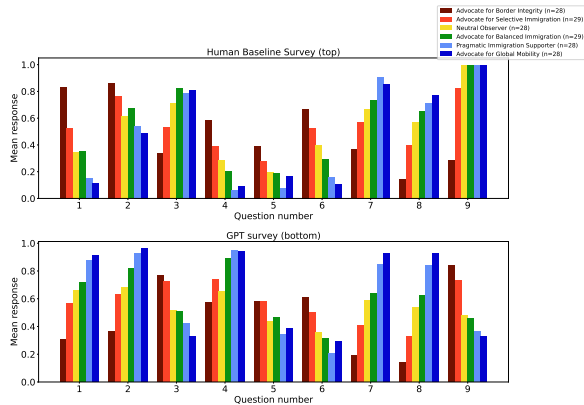


Figure 3: Immigration surveys: Normalized mean response per question by persona label. The top panel shows the human baseline survey and the bottom panel shows the GPT-generated survey. Bars are colored by self-identified immigration persona, ordered from least to most concerned.

tinctions between neighboring personas are more compressed for several questions, and separation is less consistent across the full set of items.

Overall, the results suggest that self-identified immigration personas correspond to systematic response differences, and that both the human-designed and GPT-generated surveys reflect these distinctions in an interpretable, though not perfectly uniform, manner.

D.1.1. Clustering Quality and Stability

We determine the cluster structure supported by each instrument using internal validation criteria. Both the Silhouette Score and the DBI indicate that a two-cluster solution ($k = 2$) provides the best balance of cohesion and separation for the human baseline and GPT-generated immigration surveys (see the Silhouette curve in Figure 4). Although DBI curves are omitted for brevity, they consistently favor the same solution. Accordingly, we proceed with $k = 2$ in order to facilitate direct comparison between instruments.

To evaluate the robustness of this partition, we conduct a resampling-based stability analysis. For each survey, we perform 300 bootstrap iterations, sampling 80% of respondents without replacement and reclustering using k -means with $k = 2$. Stability is quantified via the ARI, which measures agreement between cluster assignments across runs.

Both surveys yield highly stable partitions. The human baseline produces a mean ARI of 0.79 (median = 0.835), with the central 80% of values spanning 0.53 to 1.00. The GPT-generated survey exhibits slightly higher average stability, with a mean ARI of 0.82 (median = 0.86) and a 10th–90th percentile range of 0.58 to 1.00. Although a small

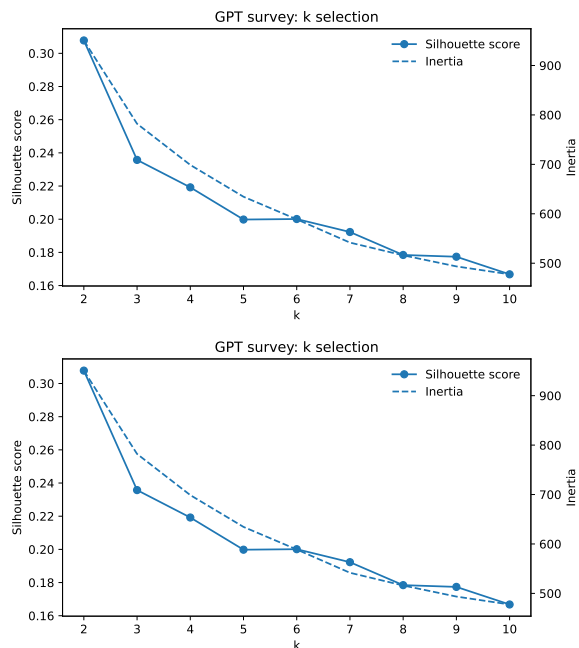


Figure 4: Selection of the number of clusters (k) using silhouette score and inertia for the human baseline (top) and GPT-generated (bottom) immigration change surveys.

Human Baseline/GPT	Cluster 0	Cluster 1
Cluster 0	0.76	0.24
Cluster 1	0.05	0.95

Table 4: Transition matrix between clusters induced by the human baseline survey and the GPT-generated survey. Rows correspond to human baseline clusters; columns correspond to GPT clusters.

number of runs show reduced agreement, the predominance of high ARI values indicates that the two-cluster structure is robust to substantial perturbations of the data.

We next evaluate the extent to which the human baseline and GPT-generated surveys yield comparable latent partitions of respondents. The transition matrix in Table 4 indicates substantial concordance between the two surveys. A large majority of respondents remain on the diagonal (76% and 95%), with relatively few reassigned across clusters. This pattern suggests that the GPT-generated survey recovers the same dominant attitudinal division identified by the human baseline, with only modest reallocation near the boundaries of the partition.

We assess external semantic validity by examining how unsupervised cluster assignments align with participants' self-identified personas, which were collected at the beginning of the study and reflect ordered positions along an immigration belief spectrum. This comparison evaluates whether

the latent groupings induced by each survey correspond to respondents' own stated ideological orientations.

Using the ARI and NMI, we find modest but systematic alignment between clusters and persona labels. The human baseline survey yields an ARI of 0.12 and an NMI of 0.21, while the GPT-generated survey shows slightly higher alignment (ARI = 0.15, NMI = 0.22). Although these values do not indicate one-to-one correspondence, they reflect non-trivial shared structure between unsupervised clusters and self-reported identities.

D.2. DEI

Figure 5 presents the mean response for each question in the human baseline survey (top) and the GPT-generated survey (bottom), with bars colored by DEI persona.

In the DEI domain, the persona patterns are highly structured and largely mirrored across the human baseline (top) and GPT survey (bottom). For several items (especially Questions 1–2 and 7–9) mean responses increase steadily from *Advocate for Meritocracy* through to *Advocate for Structural Equity*, with the highest-endorsement personas consistently at the inclusive-culture/structural-equity end of the spectrum. For other items (most clearly Questions 3–6 in the human baseline), the direction flips: meritocracy-oriented personas show higher agreement and structural-equity personas lower agreement, which is exactly what we would expect for items framed around merit or skepticism toward identity-based interventions. The GPT-generated survey preserves this overall structure—strong separation at the extremes and coherent ordering on most items—though a few questions show small non-monotonicities (e.g., Q5 and Q9) where adjacent personas overlap or the top-end persona is not strictly the maximum. Overall, the figure suggests that both instruments capture the same underlying DEI stance spectrum, with item-dependent polarity reflecting whether an item aligns with meritocratic versus equity-centered framing.

D.2.1. Clustering Quality and Stability

We next identify the cluster configuration favored by each survey using internal validation metrics. Both the Silhouette Score and the DBI converge on a two-cluster solution ($k = 2$) as offering the strongest trade-off between within-cluster cohesion and between-cluster separation for the human baseline and GPT-generated DEI surveys (see the Silhouette curve in Figure 6). We therefore adopt $k = 2$ for subsequent analyses to enable direct comparison across instruments.

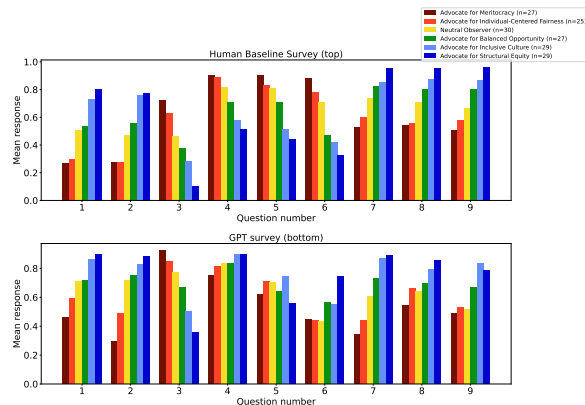


Figure 5: DEI surveys: Normalized mean response per question by persona label. The top panel shows the human baseline survey and the bottom panel shows the GPT-generated survey. Bars are colored by self-identified DEI persona, ordered from least to most concerned.

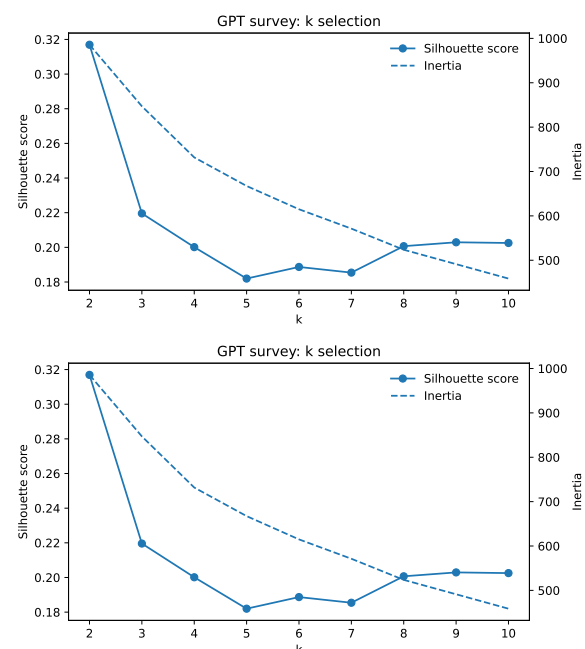


Figure 6: Selection of the number of clusters (k) using silhouette score and inertia for the human baseline (top) and GPT-generated (bottom) DEI change surveys.

As in the other domains, we evaluate the stability of the $k = 2$ solution using the same resampling-based procedure described for other topics. Both instruments exhibit strong stability under resampling. The human baseline achieves a mean ARI of 0.86 (median = 0.92), with the central 80% of values ranging from 0.63 to 1.00. The GPT-generated survey demonstrates comparable robustness, with a mean ARI of 0.83 (median = 0.85) and a 10th–90th percentile interval of 0.60 to 1.00. While a small number of iterations yield lower agreement,

Human Baseline/GPT	Cluster 0	Cluster 1
Cluster 0	0.89	0.11
Cluster 1	0.26	0.74

Table 5: Transition matrix between clusters induced by the human baseline survey and the GPT-generated survey. Rows correspond to human baseline clusters; columns correspond to GPT clusters.

the concentration of high ARI values indicates that the two-cluster structure is resilient to substantial perturbations of the sample.

We examine cross-instrument consistency by comparing the partitions produced by the human baseline and GPT-generated surveys. The transition matrix in Table 5 reveals strong alignment between the two solutions. A substantial proportion of respondents remain on the diagonal (89% and 74%), with relatively limited reassignment across clusters. This pattern suggests that the GPT-generated survey recovers the same primary attitudinal division identified by the human baseline, with differences largely confined to respondents near the decision boundary between clusters.

We assess external semantic validity following the same procedure used for the other two topics. Using the ARI and NMI, we observe modest but systematic alignment between clusters and persona labels. The human baseline survey yields an ARI of 0.07 and an NMI of 0.12, while the GPT-generated survey shows comparable alignment (ARI = 0.07, NMI = 0.12). Although these values are lower than those observed in the climate and immigration domains, they are similar across both instruments, suggesting that each survey captures a comparable degree of shared structure with self-reported DEI personas.

Where Is Politeness in Japanese BERT? A Layerwise Probing and CLS Activation Patching Study

Shusuke Hashimoto, Wenchen Shi

Department of Linguistics, Indiana University Bloomington
{shuhashi, wencshi}@iu.edu

Abstract

Politeness is a key aspect of language use and pragmatics, and Japanese honorifics provide a useful testbed for analyzing how pretrained encoders represent socially meaningful distinctions. We study whether LineDistilBERT encodes Japanese honorific level in its internal representations using the KeiCO corpus, which annotates sentences with four honorific levels. To isolate pretrained representations while enabling task evaluation, we freeze the encoder and train only a lightweight classification head as a minimal readout. We then conduct layerwise linear probing by training multinomial L2-regularized logistic-regression probes on [CLS] representations from each layer, identifying an intermediate layer where honorific level is most linearly decodable. Finally, we test causal leverage via [CLS] activation patching and measure how predictions shift when donor activations are transplanted into receiver sentences. Overall, we find that honorific level is broadly decodable across layers and that [CLS] interventions can systematically steer the classifier’s predictions, with strong depth dependence. These results show how combining linear decodability and causal interventions can yield complementary evidence about how a model encodes socially meaningful distinctions.

Keywords: BERT interpretability, linear probing, activation patching, Japanese honorifics

1. Introduction

Japanese honorifics provide a compact, well-studied testbed for pragmatic representation learning. Speakers choose honorific forms to reflect interpersonal relationships, social hierarchy, and situational formality, and these choices interact with morphosyntax (e.g., honorific verb forms) and conventionalized expressions. This makes honorific level a natural target for asking whether pretrained encoders represent pragmatically relevant distinctions beyond shallow lexical cues.

Several foundational studies have used BERT-based models to investigate pragmatic phenomena (Cho and Kim, 2024; Wise et al.). Prior work has also examined Japanese politeness from a pragmatics perspective (Hill et al., 1986; Liu and Allen, 2014; Pizziconi, 2003). In addition, BERT-based approaches have been applied to Japanese pragmatics (Chia et al., 2024). To our knowledge, however, no prior study has specifically examined how Japanese honorific-level politeness is represented in Japanese BERT-style encoders by combining layerwise probing with [CLS] activation patching. Accordingly, we investigate where honorific level is encoded in a Japanese BERT-style encoder, LineDistilBERT, and whether internal representations can be causally leveraged to steer predictions using the four-level KeiCO honorific corpus. To isolate what is already present in the pretrained encoder representations, we freeze all LineDistilBERT encoder parameters and train only a lightweight classification head on top of [CLS] as a minimal supervised readout for probing and patching.

Our analysis combines layerwise linear probing and targeted causal intervention. We first train multinomial L2-regularized logistic-regression probes on [CLS] representations from each layer to quantify how linearly decodable honorific level is across depth and to identify a dev-selected best layer. We then perform [CLS] activation patching at selected layers by overwriting a receiver sentence’s [CLS] activation with a donor activation and measuring the resulting prediction shifts, together with standard controls.

This paper addresses three research questions. **RQ1** asks to what extent pretrained LineDistilBERT encodes honorific-level distinctions in its internal representations. **RQ2** asks which layer provides the strongest linear decodability of honorific level and how sharply this information is localized around that layer. **RQ3** asks whether the probe-selected best layer plays a causal role in honorific prediction and whether targeted activation patching at that layer can reliably steer honorific predictions.

In summary, we contribute (i) a layerwise decodability profile for Japanese honorific level in a pretrained encoder, (ii) a best-layer evaluation against a frozen-encoder head-only baseline, and (iii) a causal patching analysis with self-, random-, and wrong-layer controls that probes depth-dependent sensitivity.

2. Related Work

We build on two interpretability paradigms for transformer encoders. Probing studies train lightweight

classifiers on frozen representations to quantify linear decodability and its variation across layers (Alain and Bengio, 2018; Belinkov, 2022), and layerwise probe curves have been used to track how linguistic properties become accessible with depth in BERT-style models (Jawahar et al., 2019; Rogers et al., 2020). Activation patching provides a causal counterpart by overwriting internal activations at a chosen site and measuring output changes (Heimersheim and Nanda, 2024; Dumas et al., 2025), with recent work emphasizing careful controls and interpretation (Zhang and Nanda, 2024).

3. Dataset

We use KeiCO corpus (Liu and Kobayashi, 2022), which annotates each sentence with one of four honorific (politeness) levels under a Systemic Functional Linguistics (SFL) framework (Liu and Kobayashi, 2022). Detailed information about the dataset is provided in (Appendix 1). We follow the dataset’s label descriptions, ranging from highly formal honorific constructions (Levels 1–2) to polite language with limited honorific morphology (Level 3) and informal speech with no honorifics (Level 4), and we split the data into stratified 80%/10%/10% train/dev/test partitions.

Our model is LineDistilBERT with a frozen encoder and a lightweight sequence-classification head trained as a minimal readout. For probing, we select the best layer by dev macro-F1 and report final results on the test set.

4. Methods

We analyze pretrained LineDistilBERT with a frozen encoder and a lightweight classification head trained as a minimal readout (Koga et al., 2023). We chose LineDistilBERT because its simplified layer structure makes it easier to interpret in layerwise probing. Our methods combine layerwise linear probing to locate decodable honorific information and [CLS] activation patching to test causal steering. We treat the head-only classifier as the fixed readout function used to compute both baseline and patched predictions, so interventions are evaluated without updating encoder parameters.

4.1. Layerwise probing

For each input sentence, we run the frozen encoder and extract the [CLS] vector from each hidden-state index as a layer-specific sentence representation. Here, hidden-state index 0 corresponds to the embedding output (pre-contextualization), and indices 1–6 correspond to the outputs of successive transformer layers. We adopt the working hypothe-

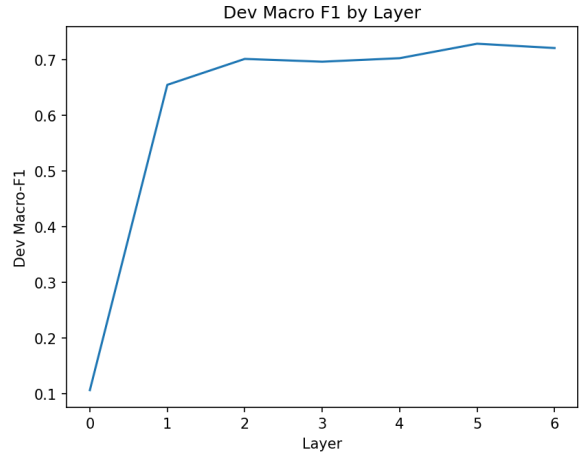


Figure 1: Dev macro-F1 of layerwise [CLS] probes in LineDistilBERT.

sis that honorific-related information is substantially aggregated into [CLS] and is therefore accessible via a single sequence-level embedding, while noting that relevant cues may also be token-local (e.g., honorific-bearing verbs or function morphemes), which we leave to future work.

For each layer l , we train a multinomial L2-regularized logistic-regression probe on $X_{\text{train}}^{(l)} \in R^{N \times H}$ and evaluate on $X_{\text{dev}}^{(l)}$. We select the best layer by dev macro-F1,

$$\text{best_layer} = \arg \max_l \text{MacroF1}_{\text{dev}}(l),$$

and report a layerwise decodability curve together with the dev-selected best_layer .

4.2. Activation patching

To test causal leverage, we perform [CLS] activation patching at selected layers. Let $\text{CLS}_x^{(l)} \in R^H$ denote the [CLS] hidden state at layer l . For a donor input x_d and receiver input x_r , we replace

$$\text{CLS}_{x_r}^{(l)} \leftarrow \text{CLS}_{x_d}^{(l)}$$

during the receiver forward pass and then measure changes in logits and predictions under the same readout. Donors are drawn from Level 1 and receivers from Level 4. Operationally, we implement patching by overwriting the CLS vector in the layer output tensor during the receiver pass. We summarize effects using (i) the transition matrix from baseline to patched predictions, (ii) Δ target-class logit, and (iii) the flip-to-target rate.

We include three standard controls. Self-patch uses $x_d = x_r$ and should yield near-identity behavior. Random-donor patch permutes donor [CLS] vectors within a batch to test whether effects depend on specific donor–receiver pairings. Wrong-

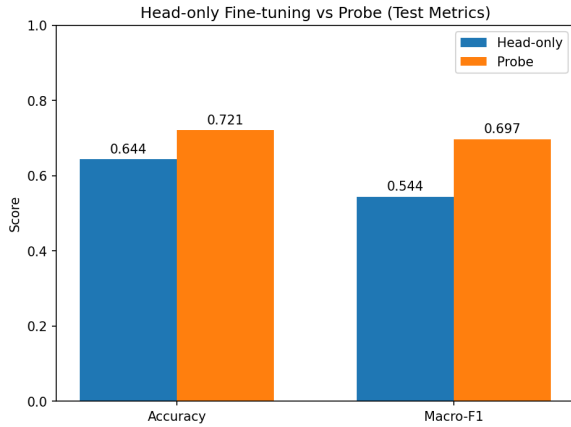


Figure 2: Head-only baseline vs. linear probe performance on the test set

layer patch sweeps non-target layers to assess depth localization of causal sensitivity.

5. Results

5.1. Decodability and best layer

Figure 1 reports dev macro-F1 scores of multinomial logistic-regression probes trained on [CLS] representations extracted from each hidden-state index. Scores rise sharply from Layer 0 to Layer 1 and remain consistently high across Layers 1–6 within a narrow range. This addresses **RQ1** and suggests that honorific-level distinctions are already encoded in pretrained LineDistilBERT representations in a form that is broadly linearly decodable across the encoder stack, rather than emerging only in a single late layer. The best-performing layer on the dev set is Layer 5 (macro-F1 ≈ 0.73), which we designate as the *best layer*. This addresses **RQ2** and indicates that the peak is shallow rather than sharply localized, since neighboring layers, especially Layers 4–6, achieve nearly identical dev macro-F1.

5.2. Best-layer probe vs. head-only baseline

After selecting the best layer on the development set (Layer 5), we trained a multinomial L2-regularized logistic-regression probe on the training [CLS] features from that layer and evaluated it once on the test set. Figure 2 compares test performance between (i) the head-only baseline, where the pretrained encoder is frozen and only the task head is trained, and (ii) the linear probe, which reads out honorific labels from fixed Layer 5 representations. The probe outperforms the head-only baseline on the test set. Together with the shallow

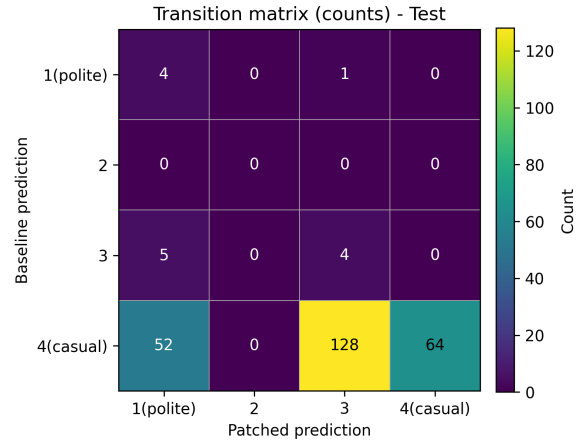


Figure 3: Test-set transition matrix (counts) for CLS activation patching at Layer 5.

best-layer peak in Figure 1, this further supports **RQ2** by showing that the dev-selected best layer yields a strong readout on test data, while nearby layers remain competitive.

5.3. Causal steering at the best layer

To test whether the probed representation is not only linearly decodable but also causally influential for predictions, we intervene on the forward pass via [CLS] activation patching at the selected layer (hidden-state index 5). We overwrite a receiver sentence’s [CLS] vector with a donor [CLS] vector (donors drawn from Level 1; receivers drawn from Level 4) and re-run the head-only classifier. Figure 3 reports the resulting transition matrix on the test set. The dominant effect is a marked reduction in Level 4 predictions after patching. Among baseline Level 4 predictions ($n = 244$), 64 remain Level 4 under patching, while 180 transition to other labels (52 to Level 1 and 128 to Level 3). Thus, patching at this layer changes the predicted label for approximately $180/244 \approx 74\%$ of baseline-4 cases. These systematic shifts directly speak to **RQ3** by showing that targeted patching at the probe-selected best layer can steer honorific predictions for a large fraction of cases.

5.4. Controls and depth dependence

We next evaluate standard controls to clarify how to interpret the steering effect. Self-patch yields near-identity behavior with no off-diagonal movement (Appendix 2), showing that patching itself does not alter the predicted label. Random patch, which permutes donors within a batch, produces a transition pattern highly similar to aligned patching, suggesting that the main effect is not driven by specific donor–receiver pairings (Appendix 2).

Finally, wrong-layer sweeps show strong depth dependence. Early-layer patching yields little to no change, whereas later-layer patching increasingly steers predictions, with the strongest shifts at the final layer (Layer 6) where the classification head consumes [CLS] most directly (see Appendix 3, Figure 10). This pattern suggests that patching effects are amplified by proximity to the classifier input and do not isolate a single honorific-specific causal layer.

6. Discussion

Our results show a nuanced view of where honorific information is accessible and where it influences prediction in a frozen-encoder classifier. First, [CLS] transplantation is a broad intervention. The close similarity between aligned-donor and random-donor patching suggests that overwriting [CLS] can induce large prediction shifts even when donor identity is randomized. This pattern is consistent with at least two hypotheses. One is that many donor [CLS] activations share a common component that tends to push predictions away from the most casual region, making the effect relatively robust to which donor is used. Another is that overwriting [CLS] at this layer acts as a generic but structured perturbation that biases decisions toward non-casual labels, largely independent of donor label.

These hypotheses make different predictions that can be tested without changing the overall intervention framework. If a shared “politeness” component dominates, then the effect should vary systematically with donor label or donor strength (e.g., donors from more polite levels, or donors that the baseline classifier assigns high confidence, should induce larger shifts). If the effect is largely a generic perturbation, then shifts should be comparatively insensitive to donor label and instead correlate more with receiver-side factors such as baseline confidence or juxtaposition to the decision boundary (e.g., low-margin cases should flip more easily). In both cases, label-conditioned randomization provides a direct diagnostic by stratifying donors and receivers and comparing effect sizes across strata.

Two broader implications follow. The high and relatively flat decodability curve across Layers 1–6 suggests that honorific cues are distributed across depth rather than isolated to a single late layer. At the same time, the wrong-layer sweep indicates that causal steering is strongest near the final layer. This mismatch is interpretable because probing and patching capture different notions. Layerwise probing identifies where honorific level is most *linearly decodable* under an external diagnostic classifier, whereas patching measures where overwriting the representation most strongly controls the model’s

own downstream decision. These need not coincide. In an encoder–classification architecture, the classification head directly consumes the final-layer [CLS] representation, so patching at Layer 6 overwrites the classifier’s immediate input and can therefore exert the strongest causal control even if linear decodability peaks slightly earlier.

Taken together, these results suggest that pre-trained LineDistilBERT carries a strong and linearly accessible signal for honorific level across most of the encoder stack, and that a minimal readout can be systematically steered by intervening on internal [CLS] activations. At the same time, the similarity between aligned and randomized patching highlights that full-vector [CLS] transplantation is not cleanly honorific-specific, and that causal steering in this setup may mix honorific-related content with other factors aggregated in [CLS].

7. Conclusion

We studied where honorific-level distinctions are represented in pretrained LineDistilBERT and whether those representations can steer a frozen-encoder classifier. For **RQ1**, layerwise probing showed that honorific level is broadly linearly decodable across most encoder layers. For **RQ2**, dev selection identified an intermediate best layer (Layer 5), but the peak was shallow, with neighboring layers performing similarly, suggesting limited localization. For **RQ3**, [CLS] activation patching at the best layer induced systematic shifts away from the most casual class, demonstrating causal leverage of the probed representation under our readout. Controls clarify the interpretation. Self-patch confirms that hooking does not induce drift, random patch indicates the effect is not driven by donor–receiver pairing, and wrong-layer sweeps show strong depth dependence with the largest shifts near the final layer. These results suggest that honorific-related information is present in hidden states and that [CLS] activations can serve as an effective control point for steering honorific predictions, motivating token-localized sites and label-conditioned causal tests.

By grounding interpretability analysis in a well-studied sociolinguistic system, we illustrate how combining linear decodability and causal interventions yields complementary evidence about what it means for a model to “encode” social meaning. In particular, the dissociation between the layer of maximal linear decodability and the layer of maximal causal influence suggests that representational accessibility and decision control can occupy different loci within the model stack. This distinction matters for interpreting probing results and for designing interventions that target socially meaningful information.

8. Limitation

(1) The nature of Japanese politeness While Japanese politeness is often overtly marked in the surface string, it is also shaped by social relations and situational norms. As a result, some cases remain underspecified without context, and the same surface form can support multiple pragmatic readings. Because our modeling setting provides only sentence-level text as input, sentence-level classification may consequently over-rely on token-level markers that correlate with politeness in the corpus, rather than the underlying situational variables. A stronger test would use context-rich instances (multiple sentences per example) and/or controlled subsets where overt honorific markers are minimized, forcing models to exploit discourse and situational cues rather than recognizing surface forms.

(2) Dataset context and intervention interpretability Our intervention analyses inherit limitations from the dataset and input representation. KeiCO provides topical field labels, which can be useful for controlling topic confounds, but these tags do not explicitly encode interactional context such as speaker–addressee relations, social hierarchy, or situational formality. This matters for [CLS] patching. Because [CLS] aggregates many factors beyond honorific level (topic, semantics, discourse structure), transplanting an entire [CLS] vector from an unrelated sentence can introduce broad distributional changes that are not specific to politeness. The observation that random-donor patching yields effects comparable to aligned-donor patching is consistent with this concern, suggesting that part of the steering may reflect generic perturbation rather than label-specific transfer.

A stronger approach would therefore increase contextual control at the instance level and narrow the intervention target. For example, one can construct minimal pairs that differ only in situational metadata (e.g., explicit formality/relationship tags) and patch across these controlled contexts, or restrict donor/receiver sampling within the same topical field to reduce topic-driven shifts. Another direction is to move from full-vector transplantation toward more targeted interventions such as editing representations along a learned politeness direction, which could better isolate the causal contribution of honorific-related features.

(3) High performance does not necessarily imply understanding Our analyses are motivated by the hypothesis that politeness-related features are important for solving honorific-level classification. However, strong accuracy or decodability does not by itself establish that such features are *necessary* for task success. A direct next step is

an ablation-based causality test, which removes or suppresses candidate politeness-related components and measures the performance change relative to the intact model. This follows a common interpretability logic that quantifies component importance by comparing a full model to an ablated model and inspecting the resulting performance difference (Li and Janson, 2024). Related causal-concept work similarly uses concept ablation (with random-concept controls) to test whether a hypothesized concept actually plays a causal role in predictions (Singla et al., 2021). If performance remains largely unchanged after ablating the "purported" politeness mechanism, then the model may be exploiting alternative correlates in the dataset, and understanding politeness (as operationalized by our identified features) may not be required for this benchmark.

(4) Heterogeneity in patchability across instances Our patching analyses aggregate effects across many donor–receiver pairs, implicitly treating instances as equally patchable. In practice, patching sensitivity may vary. Some donor [CLS] vectors may more easily steer receiver predictions than others, and some receiver sentences may be relatively resistant to donor signals. Future work should characterize this heterogeneity explicitly by, for example, stratifying patch effects by baseline confidence (e.g., max softmax probability / logit margin), sentence length, presence of overt honorific markers, or semantic/domain similarity between donor and receiver, and by reporting instance-level distributions rather than only aggregate transition counts.

9. Acknowledgments

We are grateful to Luke Gessler and Phakphum Artkaeew for their valuable feedback and support throughout this work.

10. References

- Guillaume Alain and Yoshua Bengio. 2018. [Understanding intermediate layers using linear classifier probes](#).
- Yonatan Belinkov. 2022. [Probing classifiers: Promises, shortcomings, and advances](#). *Computational Linguistics*, 48(1):207–219.
- Zheng Lin Chia, Michal Ptaszynski, Marzena Karpinska, Juuso Eronen, and Fumito Masui. 2024. [Initial exploration into sarcasm and irony through machine translation](#). *Natural Language Processing Journal*, 9:100106.

- Ye-eun Cho and Seong mook Kim. 2024. [Pragmatic inference of scalar implicature by LLMs](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 10–20, Bangkok, Thailand. Association for Computational Linguistics.
- Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2025. [Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31822–31841, Vienna, Austria. Association for Computational Linguistics.
- Stefan Heimersheim and Neel Nanda. 2024. [How to use and interpret activation patching](#).
- Beverly Hill, Sachiko Ide, Shoko Ikuta, Akiko Kawasaki, and Tsunao Ogino. 1986. [Universals of linguistic politeness: Quantitative evidence from japanese and american english](#). *Journal of Pragmatics*, 10(3):347–371.
- Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. [What does BERT learn about the structure of language?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.
- Kobayashi Koga, Shengzhe Li, Akifumi Nakamachi, and Toshinori Sato. 2023. [Line distilbert japanese](#). GitHub repository.
- Maximilian Li and Lucas Janson. 2024. [Optimal ablation for interpretability](#).
- Muxuan Liu and Ichiro Kobayashi. 2022. [Construction and validation of a Japanese honorific corpus based on systemic functional linguistics](#). In *Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference*, pages 19–26, Marseille, France. European Language Resources Association.
- Xiangdong Liu and Todd James Allen. 2014. [A study of linguistic politeness in japanese](#). *Open Journal of Modern Linguistics*, 4(5):651–663.
- Barbara Pizziconi. 2003. [Re-examining politeness, face and the japanese language](#). *Journal of Pragmatics*, 35:1471–1506.
- Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. [A primer in BERTology: What we know about how BERT works](#). *Transactions of the Association for Computational Linguistics*, 8:842–866.
- Sumedha Singla, Stephen Wallace, Sofia Triantafyllou, and Kayhan Batmanghelich. 2021. [Using causal analysis for conceptual deep learning explanation](#).
- Matt Wise, Houda Nait, El Barj, and Anna Goldie. No. [Pragmabert: Analyzing pragmatic markers in political speech](#).
- Fred Zhang and Neel Nanda. 2024. [Towards best practices of activation patching in language models: Metrics and methods](#).

Appendix

1. Dataset Overview

Table 1: Dataset statistics (from Liu & Kobayashi, 2022; Table 3).

Polite	Sent.	Avg.len	Wordtokens	Wordtypes	Sentence Example
Level 1	2,584	18.2	47,111	4,744	担当の者をお呼びしました。 "I called the person in charge."
Level 2	2,046	16.4	33,476	3,897	一度、ゆっくりお礼にあがります。 "I'll come by sometime and thank you properly."
Level 3	2,694	15.2	40,980	4,448	あの人のどこが嫌いなんですか？ "What do you dislike about that person?"
Level 4	2,683	13.5	36,233	4,315	これ、うちのオススメ。 "This is our recommendation."
Total	10007	15.8	157806	6465	

2. Self- and Random-Patch Transition Matrices

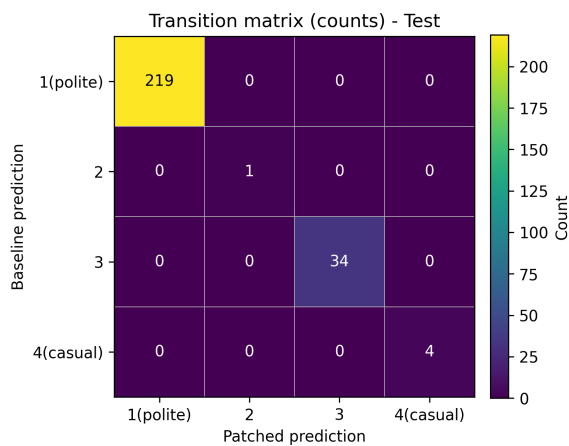


Figure 4: Self-patch transition matrix (counts) for CLS activation patching at Layer 5.

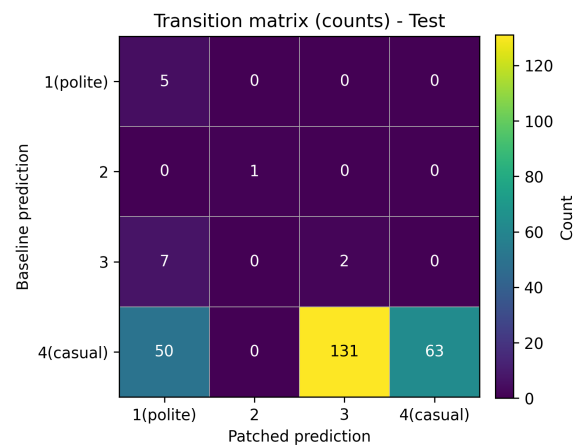


Figure 5: Random-patch transition matrix (counts) for CLS activation patching at Layer 5.

3. Wrong-Layer Patch Transition Matrices

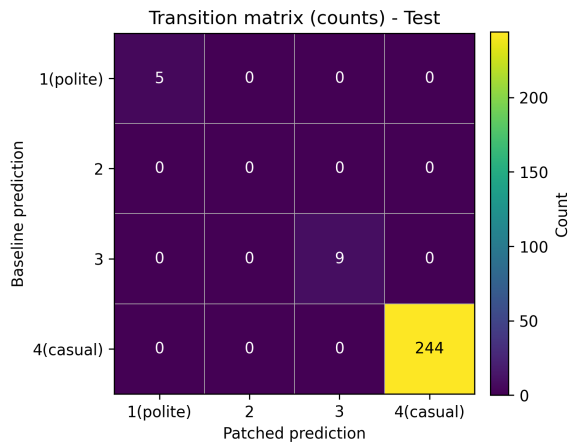


Figure 6: Wrong-layer patch transition matrix (counts) for CLS activation patching at Layer 1.

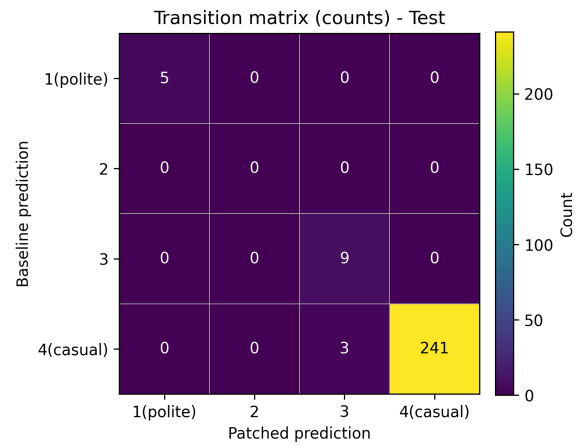


Figure 7: Wrong-layer patch transition matrix (counts) for CLS activation patching at Layer 2.

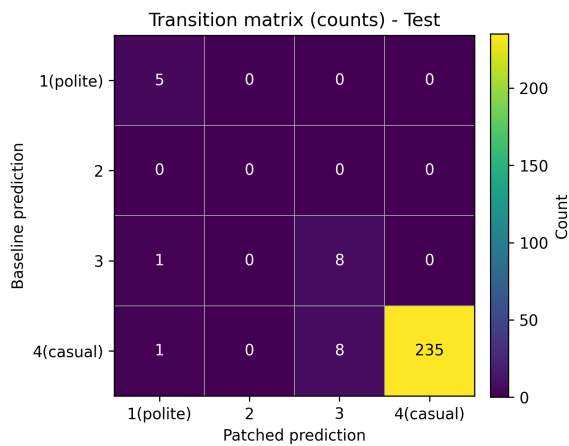


Figure 8: Wrong-layer patch transition matrix (counts) for CLS activation patching at Layer 3.

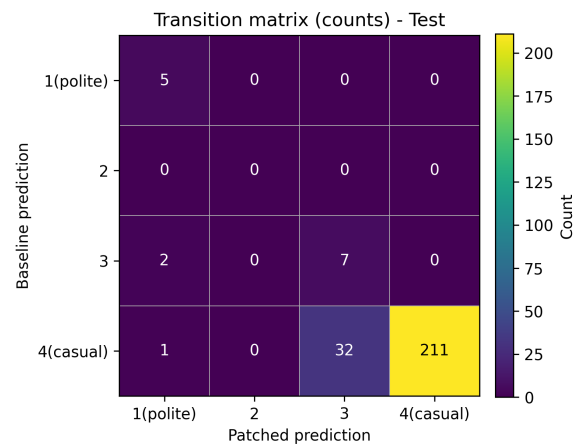


Figure 9: Wrong-layer patch transition matrix (counts) for CLS activation patching at Layer 4.

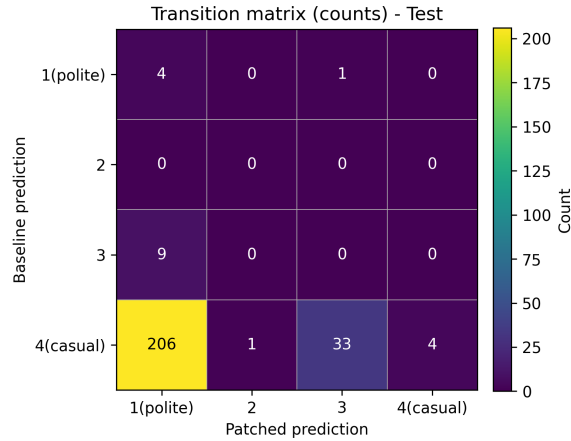


Figure 10: Wrong-layer patch transition matrix (counts) for CLS activation patching at Layer 6.

4. Hyperparameters for Baseline Training (frozen-encoder readout)

Component	Details
Optimizer	AdamW
Learning Rate	2.0×10^{-5}
Weight Decay	0.01
Scheduler	Linear
Batch Size	16
Epochs	5
Tokenizer	line-corporation/line-distilbert-base-japanese
Padding	95th percentile of sentence lengths

5. Test Metrics

Setting	Test metrics
Best-layer patch (hs index = 5)	$N = 258$; Δ target logit = 0.872; flip-to-target = 0.221
Control: Self-patch	$N = 258$; Δ target logit = 0.000; flip-to-target = 0.000
Control: Random-patch (permute within batch)	$N = 258$; Δ target logit = 0.873; flip-to-target = 0.221
Control: Wrong-layer patch (hs index = 1; embedding output)	$N = 258$; Δ target logit = -0.000; flip-to-target = 0.000
Control: Wrong-layer patch (hs index = 2)	$N = 258$; Δ target logit = 0.065; flip-to-target = 0.000
Control: Wrong-layer patch (hs index = 3)	$N = 258$; Δ target logit = 0.135; flip-to-target = 0.008
Control: Wrong-layer patch (hs index = 4)	$N = 258$; Δ target logit = 0.375; flip-to-target = 0.012
Control: Wrong-layer patch (hs index = 6; final layer)	$N = 258$; Δ target logit = 1.694; flip-to-target = 0.833

Rewrite the News: Tracing Editorial Reuse Across News Agencies

Soveatin Kuntur^{1*}, Nina Smirnova^{2*}, Anna Wroblewska¹,
Philipp Mayr², Sebastijan Razboršek Maček³

¹Warsaw University of Technology, Poland

²GESIS – Leibniz Institute for the Social Sciences, Germany

³Slovenian Press Agency, Slovenia

{soveatin.kuntur.dokt, anna.wroblewska1}@pw.edu.pl,

{nina.smirnova, philipp.mayr}@gesis.org, sm@sta.si

*Equal contribution

Abstract

This paper investigates sentence-level text reuse in multilingual journalism, analyzing where reused content occurs within articles. We present a weakly supervised method for detecting sentence-level cross-lingual reuse *without* requiring full translations, designed to support automated pre-selection to reduce information overload for journalists (Holyst et al., 2024). The study compares English-language articles from the Slovenian Press Agency (STA) with reports from 15 foreign agencies (FA) in seven languages, using publication timestamps to retain the earliest likely foreign source for each reused sentence. We analyze 1,037 STA and 237,551 FA articles from two time windows (October 7–November 2, 2023; February 1–28, 2025) and identify 1,087 aligned sentence pairs after filtering to the earliest sources. Reuse occurs in 52% of STA articles and 1.6% of FA articles and is predominantly non-literal, involving paraphrase and compositional reuse from multiple sources. Reused content tends to appear in the middle and end of English articles, while leads are more often original, indicating that simple lexical matching overlooks substantial editorial reuse. Compared with prior work focused on monolingual overlap, we (i) detect reuse across languages without requiring full translation, (ii) use publication timing to identify likely sources, and (iii) analyze where reused material is situated within articles. Dataset and code: <https://github.com/kunturs/lrec2026-rewrite-news>.

Keywords: Text reuse detection, multilingual NLP, cross-lingual journalism, computational social science

1. Introduction

Text reuse has long been a topic of interest in Natural Language Processing (NLP), from early analyses of classical texts such as the Synoptic Gospels (Lee, 2007) to applications like plagiarism detection (Clough et al., 2002) and healthcare communication (D’hondt et al., 2016). However, despite its historical significance, broader attention to the phenomenon has declined in recent years.

Nonetheless, text reuse remains a central practice in multilingual journalism. National news agencies often adapt international reports, paraphrasing, summarizing, or merging multiple foreign sources to produce domestic news articles: factual pieces published on the agency’s newswire and used by other media outlets. Understanding how these articles are constructed sheds light on the invisible processes of news dissemination, editorial decision-making, and cross-lingual information flow, which can help in preparing automated methods for information extraction and the pre-selection of fragments relevant at the national level.

Analyzing text reuse in this setting uncovers the mechanisms of cross-lingual news dissemination and highlights the need for computational methods capable of tracing multilingual content transformation, essential for transparency and efficiency in global journalism. This paper investigates how journalistic content is reused and merged across

languages within real-world news agency workflows. Specifically, we address the following research questions. For English-language news articles published by the Slovenian Press Agency (STA)¹, we examine: **(1)** How is content merged from multiple sources? **(2)** Is text reuse more common in lead paragraphs, the middle of the article or at the end? **(3)** Which article parts are being reused more frequently?

To answer these questions, we propose a weakly supervised framework for identifying and classifying sentence-level reuse in multilingual newswire corpora. Using data from 16 news agencies in seven languages (Italian, English, Polish, French, German, Serbian, and Croatian), we provide empirical evidence on the scale and structure of content reuse, reconceptualizing it as a core element of the global journalistic workflow.

2. Our use case – background on journalists’ work

As part of our research, we observed the work of journalists at the STA, is a national news organization reporting on domestic and international events. It produces news in two languages - Slovenian and English. In this research, we focus only on En-

¹<https://english.sta.si/>

glish news articles. We conducted interviews with journalists from STA and analyzed their editorial workflow. Focusing on the domestic perspective, we examined how STA journalists curate and adapt international news to align with the domestic media landscape. The dataset derived from this workflow is described in detail in Section 4. Slovenia has a comparatively limited representation of journalists abroad; therefore, its national news agency relies on content exchange agreements that provide access to international reports from regions without direct correspondents. Journalists covering foreign events aggregate and synthesize information from multiple agencies to produce domestically relevant news articles. Despite the large volume of available material, only a small subset is selected for publication based on perceived relevance to Slovenia and to the agency’s subscribers. This selection process is guided by journalists’ implicit professional judgment rather than formalized or documented criteria, a judgment that derives from media industry experience, direct subscriber feedback, and indirect feedback based on analyses of article use and reuse in the Slovenian media environment.

3. Literature Review

Journalistic text reuse. Prior work in journalism has focused on building corpora and methods to study reuse and paraphrase in news. Notable resources include the METER (MEasuring Text Reuse) corpus in English (Clough et al., 2002) and COUNTER for Urdu (Sharjeel et al., 2017), which support evaluation of reuse in the press domain. Early methods relied on surface cues such as n -gram overlap, Greedy String Tiling, and sentence alignment (Clough et al., 2001), while later work combined content, structure, and style features in classification models (Bär et al., 2012). More recently, researchers have traced cross-language circulation of news (e.g., using BLAST to detect non-contiguous or editorially transformed reuse in Finnish historical journalism) (Salmi et al., 2021). Taken together, these studies show that newsroom reuse is common and often involves edits and reorganization, but most work remains monolingual and does not model cross-language editorial pipelines or where reuse tends to appear within articles.

Intertextuality vs. editorial reuse. In literary studies, intertextuality refers to echoes and references within a single language, often involving deliberate signaling. Burns et al. (2021) demonstrate that distributional models can surface non-lexical echoes (allusions) in a monolingual literary tradition of Latin epic poetry, providing new tools for literary criticism. Our setting, by contrast, concerns cross-document, cross-language editorial reuse

driven by newsroom practice: editors select, paraphrase, summarize, and merge material from foreign sources, producing many-to-many sentence alignments under time pressure rather than deliberate literary signaling.

Translation-induced intertextuality. McGovern et al. (2025) examine how human and machine translation modulate intertextual links in Biblical texts, using multilingual embeddings to assess whether references are preserved or amplified. We differ in that we do not rely on a fixed catalog of known references. Instead, we automatically identify cross-language sentence correspondences as editors adapt foreign content, and we emphasize where such content appears in target articles over time. In short, we study how translation choices interact with newsroom workflows to shape cross-lingual reuse, rather than evaluating predefined links.

Rhetorical parallelism vs. cross-document alignment. Bothwell et al. (2023) formalize rhetorical parallelism detection (RPD) as an intra-document sequence-labeling task that links stylistically parallel spans. While their focus is on within-text parallelism, the takeaway—that surface-level similarity can miss structural parallels—motivates our approach to cross-document, cross-lingual alignment, where paraphrase and restructuring are common and lexical overlap may be limited. Here, alignment refers to identifying sentences or passages with equivalent meaning across different news texts.

Cross-lingual alignment with multilingual embeddings. Multilingual sentence embeddings (e.g., multilingual SBERT) enable cross-language sentence retrieval without full translation, supporting scalable alignment across languages (Reimers and Gurevych, 2020). We build on this capability but show that editorial reuse in journalism often involves many-to-many mappings and paraphrase or merging, which such embeddings alone struggle to capture. Our approach combines multilingual sentence representations with temporal filtering and discourse-aware analysis (e.g., where content sits within an article) to study not just whether reuse occurs, but how and where it is placed. Since the significance of news positioning has been established in various domains, including its effects on financial markets (Fedyk, 2024).

4. Our Dataset and Approach

Dataset. For this research, two datasets were created. STA provided us with the API, which allowed us to crawl news articles from the STA and foreign

agencies. Thus, we collected a dataset of news articles produced by the STA for the period from October 7th, 2023, to November 2nd, 2023, and from February 1st, 2025, to February 28th, 2025. The first window covers the onset of the Israel-Hamas conflict. The second window lies outside the active conflict timeline, reflecting a comparatively calm period. We processed the STA dataset and excluded irrelevant news categories, such as weather reports and schedules of events, resulting in a final dataset of 1,037 news articles from the STA collection². Furthermore, we collected a dataset of foreign articles for the same periods, provided to the STA by different foreign agencies (FA). The FA dataset comprises 237,551 news articles from 15 foreign agencies in 7 languages. The STA dataset contains 544 news articles from 2023 and 749 from 2025, while the FA dataset includes 120,668 articles from 2023 and 116,883 from 2025. The FA collection is multilingual, with articles written 84,119 in Italian, 49,545 in English, 26,364 in Polish, 25,482 in French, 24,353 in German, 18,378 in Serbian, and 9,310 in Croatian.

Following, the datasets were cleaned from unimportant information, i.e., all HTML tags, emails, and telephone numbers were removed from news article texts in both datasets.

Weakly supervised method. Based on our observations and conversations with journalists, we developed an approach designed to determine which foreign news articles were utilized to create a target STA news article. In the initial step, all the texts from the STA and the foreign corpus were segmented into sentences using the sentence-splitter library³. Additionally, short sentences and sentences without verbs were excluded from the STA corpus⁴. This decision was motivated by both linguistic theory and practical considerations related to the semantic representation of sentences. Linguistic theories distinguish between content words (such as nouns, verbs, and adjectives) and function words (such as prepositions, conjunctions, or articles). The latter serve primarily grammatical purposes, and others contribute the most to the meaning of the sentence (Katz and Fodor, 1963; Pustejovsky, 1996). Furthermore, in English, a meaningful sentence typically requires a verb. This contrasts with some other languages, e.g., those in the Slavic language family, where a meaningful sentence can be built without verbs. Additionally, we selected only sentences with more than 7 tokens to ensure minimal semantic complexity. Specifically,

²Unfiltered STA dataset comprises 1,293 news articles.

³<https://pypi.org/project/sentence-splitter/>

⁴Tokenization and parsing parts of speech were conducted using the spaCy library.

we retained only those sentences that contained adjectives, nouns, verbs, pronouns, proper nouns, and auxiliary verbs, which are crucial for capturing modality. This approach ensured that sentences such as those demonstrated in the following examples do not participate in the similarity mapping:

- *Follow us also on:*
- *Below is a schedule of events for Saturday, 1 February:*
- *7:30am to 2pm: John Doe*

Subsequently, the sentences were transformed into vectors utilizing multilingual sentence transformers (Reimers and Gurevych, 2020). Following, the comparison was restricted to news articles published on the same date. Using cosine similarity, each sentence in the STA news articles was compared with each sentence from the collection of FA news articles. Based on preliminary experiments conducted on the external dataset, we established a similarity score threshold of 60%. The approach was evaluated on the Webis-Wikipedia-Text-Reuse-18 dataset (Alshomary et al., 2018), which consists of documented text reuse cases from Wikipedia articles. From this corpus, we randomly sampled 1,095 text pairs. Given that our news dataset is multilingual, a portion of the selected pairs was translated into six target languages represented in our dataset using a multilingual translation model⁵ (Costa-jussà et al., 2022). The dataset statistics, test results, and detailed description of the approach are presented in Table 4 (Appendix A).

The analysis showed that semantically similar sentences and full texts consistently yielded mean similarity scores above 0.6, whereas dissimilar sentence pairs did not exceed an average score of 0.4. Therefore, sentences exhibiting a cosine similarity greater than this threshold were marked as similar. The PyTorch implementation was used to calculate cosine similarity⁶.

All articles in the STA dataset possess a timestamp, which indicates when an article was created. The timestamp includes year, month, day, hour, minute, second, and microsecond. All articles in the FA dataset also possess such a timestamp, indicating when the article was received by the STA. In the last preprocessing stage, the matched articles were flagged as false positives if the STA articles were created before the corresponding foreign articles were received⁷.

⁵<https://huggingface.co/facebook/nllb-200-distilled-600M>

⁶<https://pytorch.org/docs/stable/generated/torch.nn.CosineSimilarity.html>

⁷Detailed description of the approach and testing can be found at <https://github.com/kunturs/lrec2026-rewrite-news>.

5. Results

The similarity-based approach revealed 8,432 matched sentence pairs between STA (target) and FA (source). After excluding false positives, defined as matches found in STA articles published prior to the receipt of the corresponding FA articles, the number was reduced to 4,004. We further refined the dataset by retaining only the earliest matching FA article for each STA sentence, assuming it is the most likely origin of the content. This final filtering step delivered 1,087 matched sentence pairs⁸. Table 1 demonstrates the number of matched sentences and news articles in each filtering step.

	Articles		Sentences	
	STA	FA	STA	FA
True matches	387	2,070	989	2,731
The earliest matches	—	698	—	875
False positives	419	2,376	1,117	3,288

Table 1: Number of sentences and articles matched using the similarity approach in the STA and FA datasets. An em dash (—) denotes “not applicable”. The earliest matches filtering only applies to the FA data.

As shown in Figure 1–A, sentence reuse is most prevalent in the middle sections of news articles, followed by the ending sections. In contrast, the beginning sections tend to contain the most original content. This pattern remains consistent over time. Similarly, the most reused content comes from the middle section of a news article, as Figure 1-B demonstrates. However, more content are being reused from the beginning sections than from the ending sections. The analysis shows distinct positional reuse patterns between FA and STA. Most reused sentences occur from FA-middle to STA-middle (204 cases) and FA-beginning to STA-middle (217). Beginning-to-beginning reuse is infrequent (31). FA-middle sentences map to STA-beginning (130) and STA-end (144), while FA-beginning ones match with STA-end (135). FA-end sentences mainly abide near the end (112), with fewer mappings to the beginning (42) and middle (72). Overall, reuse concentrates in the middle sections of both article types. The significance of the results was tested using the χ^2 test of independence, resulting in a p-value less than the 0.05 threshold for all variables mentioned above.

To investigate how textual reuse occurs across articles, sentence positional relationships (PRs) between STA and FA news sources were analyzed. A sentence positional relationship (PR) refers to how sentence(s) from an STA article align positionally with sentence(s) from an FA article that share se-

⁸All the analysis was applied to the dataset containing only the earliest matches.

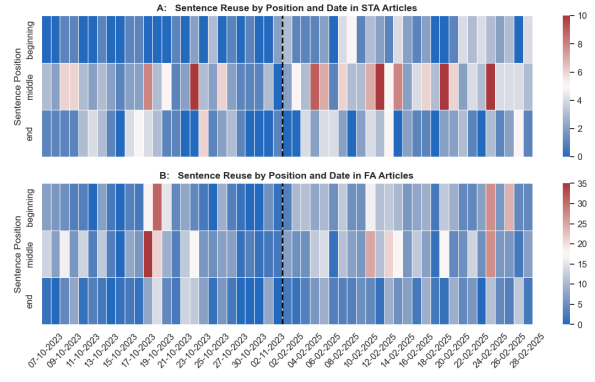


Figure 1: Distribution of reused sentences over time: (A) STA (sentences taken from FA) and (B) FA articles (sentences reused by STA). The plots show how reuse varies by sentence position (with a granularity of: beginning, middle, end of the article) within each article. The color scale on the right indicates the number of reused sentences: zero – blue, through white up to red – ten reused sentences. The dotted line marks the boundary between the years 2023 and 2025 in the data distribution.

mantically similar content. Each pair in the dataset matches one sentence from STA to one from FA. These PRs were categorized into four types based on sentence ID frequency: **1:1** (one-to-one match), **1:many** (one STA sentence aligned with multiple FA sentences), **many:1** (multiple STA sentences aligned with one FA sentence), and **many:many** (multiple PRs on both sides). The PR analysis was conducted in four steps. First, sentence pairs were filtered to ensure reuse was traced back to the earliest available foreign article. Then, sentence-level reuse frequency was counted, pairs were categorized into PR types, and the reuse direction was determined based on temporal order. The majority of sentence reuse followed **many:many** (54.9%) and **many:1** (40.5%) structures, while direct one-to-one matches accounted for only 4.4% of cases. These findings indicate that reuse is rarely literal or isolated; instead, it often involves paraphrasing, summarization, or segmentation across sentence boundaries.

Qualitative analysis of the most frequently reused sentences (corresponding to the red peaks in Figure 1-A) showed that peaks in sentence reuse in the years 2023 (Figure 2-A) and 2025 (Figure 2-B) correspond to intense news cycles dominated, respectively, by a national crisis and by international political reporting. In both cases, repetition resulted not from copying but from the systematic use of shared editorial templates to cover closely related events.



Figure 2: Word cloud illustrating recurrent lexical patterns in reused sentences, highlighting the shared editorial templates underlying coverage during major news cycles: corresponding to the most frequently reused sentences in (A) 2023 and in (B) 2025.

6. Conclusion

In this research, we investigate sentence-level text reuse in multilingual journalism. We propose a weakly supervised framework to detect reuse between English-language articles from the STA and 15 international news agencies publishing in seven languages. Our analysis indicates that 52% of STA news and approximately 1.6% of news articles from international agencies share similar content. Our findings suggest that passages originating from the beginning and middle sections of news articles are reused more frequently than those from the concluding sections. Conversely, the reused content tends to be positioned primarily in the middle and at the end of news articles, suggesting that the beginning sections are more likely to contain original material.

The distribution of PR types suggests that textual reuse between STA and FA news sources follows editorial practices that involve restructuring rather than direct copying. The low proportion of one-to-one PRs indicates that literal reuse is not the dominant strategy. Instead, the prevalence of many-to-many and many-to-one cases points to more flexible reuse patterns, likely shaped by summarization, paraphrasing, and contextual adaptation. These findings reveal the limitations of surface-level similarity approaches and emphasize the need for discourse-aware, multilingual NLP tools. By situating NLP methods within journalistic workflows, the study offers new directions for modeling cross-lingual information flow and encourages interdisciplinary collaboration across language technology, media studies, and computational social science.

7. Limitations

The primary challenge in developing the methodology for detecting text reuse originated from working with a dataset containing multilingual data in seven languages. Given the computational costs and time constraints, we opted against translating all articles into English. Instead, we utilized a multilin-

gual language representation model (Reimers and Gurevych, 2020) and libraries capable of handling data in multiple languages.

In addition to the Webis-Wikipedia-Text-Reuse-18 dataset (Alshomary et al., 2018), we evaluated our method on the Webis-CPC-11 dataset (Burrrows et al., 2013). These corpora were chosen due to their open availability. Webis-CPC-11 contains annotated paraphrased and non-paraphrased sentences. Due to the computational and processing time, for our tests, we used only part of both datasets. Specifically, from the Webis-CPC-11 corpus, a random sample of 1,000 text pairs was drawn, and our similarity approach was applied to this selection. The testing datasets, along with a detailed description of the approach and testing, can be found in the GitHub repository⁹. Additionally, the dataset statistics, test results, and description of the approach are presented in Table 4 (Appendix A). The mean similarity score for both paraphrased and non-paraphrased samples exceeds 0.7, implying that the current approach is ineffective for distinguishing paraphrased from non-paraphrased text. Therefore, this approach cannot be applied to differentiate the stance through which the information was conveyed. This high similarity score may be attributed to the nature of the corpus, in which non-paraphrased sentence pairs, despite differing in final meaning, often remain semantically similar, as illustrated by the example below:

- **Original text:** *"I have heard many accounts of him," said Emily, "all differing from each other: I think, however, that the generality of people rather incline to Mrs. Dalton's opinion than to yours, Lady Margaret." "I can easily believe it."*
- **Non-paraphrased text:** *"I have heard many accounts of him," said Emily, "all different from each other: I think, however, that the generality of the people rather inclined to the view of Ms Dalton to yours, Lady Margaret." That I can not believe.*

We opted for using cosine similarity against other similarity measures, as this approach is commonly used in many NLP tasks, is computationally efficient, and easily interpretable.

Some frequent words were deliberately deleted from Figure 2 due to anonymity issues. The deleted words will be displayed in the accepted version of the paper.

8. Acknowledgements

All authors were funded by the European Union under the Horizon Europe grant OMINO (grant num-

⁹<https://github.com/kunturs/lrec2026-rewrite-news>.

ber 101086321, (Holyst et al., 2024)). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the European Research Executive Agency can be held responsible for them. A.W. and S.K. were also co-financed with funds from the Polish Ministry of Education and Science under the program entitled International Co-Financed Projects. Nina Smirnova additionally received funding from the Deutsche Forschungsgemeinschaft (DFG) under grant number: MA 3964/7-3 (POLLUX Project).

9. Bibliographical References

- Daniel Bär, Torsten Zesch, and Iryna Gurevych. 2012. [Text reuse detection using a composition of text similarity measures](#). In *Proceedings of COLING 2012*, pages 167–184, Mumbai, India. The COLING 2012 Organizing Committee.
- Stephen Bothwell, Justin DeBenedetto, Theresa Crnkovich, Hildegund Müller, and David Chiang. 2023. [Introducing rhetorical parallelism detection: A new task with datasets, metrics, and baselines](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5007–5039, Singapore. Association for Computational Linguistics.
- Patrick J. Burns, James A. Brofos, Kyle Li, Prमित Chaudhuri, and Joseph P. Dexter. 2021. [Profiling of intertextuality in Latin literature using word embeddings](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4900–4907, Online. Association for Computational Linguistics.
- Paul Clough, Robert Gaizauskas, and S. L. Piao. 2002. [Building and annotating a corpus for the study of journalistic text reuse](#). In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).
- Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. [No language left behind: Scaling human-centered machine translation](#).
- Eva D'hondt, Cyril Grouin, and Brigitte Grau. 2016. [Low-resource OCR error detection and correction in French clinical texts](#). In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, pages 61–68, Auxtín, TX. Association for Computational Linguistics.
- Anastassia Fedyk. 2024. [Front-page news: The effect of news positioning on financial markets](#). *The Journal of Finance*, 79(1):5–33.
- Janusz A. Holyst, Philipp Mayr, Michael Thellwall, Ingo Frommholz, Shlomo Havlin, Alon Sela, Yoed N. Kenett, Denis Helic, Aljoša Rehar, Sebastijan R. Maček, Przemysław Kazienko, Tomasz Kajdanowicz, Przemysław Biecek, Bolesław K. Szymanski, and Julian Sienkiewicz. 2024. [Protect our environment from information overload](#). *Nature Human Behaviour*, 8:402–403.
- Jerrold Katz and Jerry Fodor. 1963. The structure of a semantic theory. *Language*, 39:170–210.
- John Lee. 2007. [A computational model of text reuse in ancient literary texts](#). In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 472–479, Prague, Czech Republic. Association for Computational Linguistics.
- Hope McGovern, Hale Sirin, and Tom Lippincott. 2025. [Characterizing the effects of translation on intertextuality using multilingual embedding spaces](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 161–167, Albuquerque, New Mexico. Association for Computational Linguistics.
- James Pustejovsky. 1996. Lexical semantics: the problem of polysemy.
- Nils Reimers and Iryna Gurevych. 2020. [Making monolingual sentence embeddings multilingual using knowledge distillation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Hannu Salmi, Petri Paju, Heli Rantala, Asko Ni-
vala, Aleksu Vesanto, and Filip Ginter. 2021. [The reuse of texts in finnish newspapers and journals, 1771–1920: A digital humanities perspective](#). *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 54(1):14–28.

10. Language Resource References

Alshomary, Milad and Völske, Michael and Wachsmuth, Henning and Stein, Benno and Hagen, Matthias and Potthast, Martin. 2018. [Webis Wikipedia Text Reuse Corpus 2018 \(Webis-Wikipedia-Text-Reuse-18\)](#). Zenodo.

Burrows, Steven and Potthast, Martin and Stein, Benno and Eiselt, Andreas. 2013. [Webis Crowd Paraphrase Corpus 2011 \(Webis-CPC-11\)](#). Zenodo.

Clough, Paul and Gaizauskas, Robert and Piao, Scott S. L. and Wilks, Yorick. 2001. [METER: MEasuring TExt Reuse](#). Association for Computational Linguistics.

Sharjeel, Muhammad and Nawab, Rao Muhammad Adeel and Rayson, Paul. 2017. [COUNTER: corpus of Urdu news text reuse](#).

A. Testing of the similarity approach

To establish similarity thresholds for both sentence-level and document-level approaches, we utilized the Webis-Wikipedia-Text-Reuse-18 corpus, which contains pairs of short texts and their reused counterparts. Each text pair is first segmented into sentences, yielding aligned source (`source text`) and target (`target text`) sentence sequences, as Table 2 demonstrates. We assumed a monotonic alignment, i.e., that sentences correspond sequentially between source and target.

source text	target text
According to the United States Census Bureau, the CDP has a total area of , of which, of it is land and of it (0.05) is water.	According to the United States Census Bureau, the village has a total area of 2.8 square miles (7.1 km), of which, 2.4 square miles (6.2 km) of it is land and 0.3 square miles (0.9 km) of it (12.32) is water.
Wiseman has a continental subarctic climate (Köppen Dfc).	Babylon Village is bordered to the west by West Babylon, to the north by North Babylon, to the east by West Islip, and to the south by the Great South Bay.
As of the census of 2000, there were 21 people, 7 households, and 3 families residing in the CDP.	As of the census of 2010, there were 12,166 people and 4,585 households in the village, with 2.65 persons per household.
The population density was 0.3 people per square mile (0.1 km).	The population density was 4,975.9 people per square mile (2,864.3 km).
There were 30 housing units at an average density of 0.4 sq mi (0.1 km).	There were 4,768 housing units.

Table 2: Example of the sentence alignment in the Webis-Wikipedia-Text-Reuse-18 corpus.

For each text pair, we compute similarity at two levels. First, we obtain a document-level similarity score by encoding the full source and target texts (concatenated sentences) using a sentence embedding model and computing their cosine similarity.

Second, we compute sentence-level similarities by encoding each sentence individually. We then perform pairwise comparisons between all source and target sentence embeddings. Based on the sequential alignment assumption, similarities are divided into two categories: (i) aligned sentence pairs, where source and target sentences share the same index, and (ii) non-aligned sentence pairs, where indices differ. For each category, we calculate the mean cosine similarity across all comparisons.

This procedure yields three scores per text pair: (1) average full-text similarity, (2) average similarity of non-aligned sentences (representing unrelated content), and (3) average similarity of aligned sentences (representing reused content), as Table 3 shows. These aggregated scores are then used to analyze the separation between reused and non-reused text segments and to derive empirical similarity thresholds.

language	full text	different sentences	similar sentences	support
English	0.72	0.37	0.63	500
German	0.73	0.35	0.64	99
French	0.69	0.35	0.59	99
Croatian	0.71	0.37	0.63	99
Italian	0.74	0.37	0.69	100
Polish	0.69	0.35	0.63	99
Serbian	0.72	0.37	0.69	99

Table 3: Similarity scores for the Webis-Wikipedia-Text-Reuse-18 corpus.

The same procedure was applied to the Webis-CPC-11 corpus, which contains both paraphrased and non-paraphrased versions of original texts. Table 4 reports the average full-text similarity scores. We restrict our analysis to document-level similarity only in English, as the results indicate that the proposed method is not sufficiently effective for paraphrase detection at this level. Consequently, we do not extend this approach to sentence-level analysis.

paraphrased	full text	support
no	0.78	500
yes	0.81	500

Table 4: Similarity scores for the Webis-CPC-11

OnCoCo 1.0: A Public Dataset for Fine-Grained Message Classification in Online Counseling Conversations

Jens Albrecht, Robert Lehmann, Aleksandra Poltermann,
Eric Rudolph, Philipp Steigerwald, and Mara Stieler

Technische Hochschule Georg Simon Ohm
Kesslerplatz 12, 90489 Nuremberg, Germany

Abstract

This paper presents OnCoCo 1.0, a new public dataset for fine-grained message classification in online counseling. It is based on a new, integrative system of categories, designed to improve the automated analysis of psychosocial online counseling conversations. Existing category systems, predominantly based on Motivational Interviewing (MI), are limited by their narrow focus and dependence on datasets derived mainly from face-to-face counseling. This limits the detailed examination of textual counseling conversations. In response, we developed a comprehensive new coding scheme that differentiates between 38 types of counselor and 28 types of client utterances, and created a labeled dataset consisting of about 2.800 messages from counseling conversations. We fine-tuned several models on our dataset to demonstrate its applicability. The data and models are publicly available to researchers and practitioners. Thus, our work contributes a new type of fine-grained conversational resource to the language resources community, extending existing datasets for social and mental-health dialogue analysis.

Keywords: Classification, Counseling, Conversations, Dataset, Fine-grained annotation

1. Introduction

Psychosocial counseling covers a wide range of human problems. The WHO 2021 report states that 1 in 8 people worldwide is struggling with mental health problems and disorders (WHO, 2021). In other areas, such as pregnancy, parenting, or debt counseling, the demand for advisory services is similarly increasing (NHS, 2024).

Nowadays, people can seek help on a wide range of psychological and social topics through web forums and dedicated text-based counseling platforms. *Online counseling* is provided on these platforms by, for example, medical professionals, psychologists, or social workers who have undergone specialized training in this method (Kraus et al., 2010). It is characterized by an inherently asymmetric social relationship: counselors and clients occupy different social roles in the sense of interaction role-taking, where participants continuously negotiate positions such as "helper" and "help-seeker". A trained counselor guides the interaction, while the client seeks support under conditions of vulnerability and limited contextual cues. By providing fine-grained role-specific labels for both sides, our dataset OnCoCo 1.0 enables the study of social dynamics in supportive interactions, e.g. how guidance, empathy, and resource activation unfold across counselor-client turns.

Currently, there is an intensifying debate regarding the potential of AI systems to assume the role of online counselors or even therapists (Hatch et al., 2025). Psychotherapy research (Laska et al., 2014) and studies on psychosocial counseling (Grawe, 2000) emphasize that key mechanisms of change

are strongly shaped by the content conveyed in counseling interactions. To gain a more nuanced understanding of whether and how online counseling exerts its effects, it is therefore essential to analyze the specific thematic elements (e.g. agreement on consulting goals or creation of motivation) present within a counseling exchange (Grandeit et al., 2020). Particularly, in German-language research on online counseling, these are characterized by a systematic approach using inductive or deductive coding (Mayring, 2015).

So far, data sets specifically for online counseling have been hardly publicly available, mainly due to privacy-related restrictions or reasons of data protection. In addition, existing category systems for automated content analysis focus mainly on specific counseling approaches, such as Motivational Interviewing (MI) (Miller et al., 2008; Moyers et al., 2016; Xiao et al., 2016; Wu et al., 2022). These systems often focus on face-to-face settings (Malhotra et al., 2022; Wu et al., 2022), which does not capture the full details of relevant aspects in online counseling where text is the only modality. A central challenge is that online counseling is quite different to face-to-face counseling in terms of methodology. To compensate for the lack of auditory and visual cues, it employs specific reading and writing techniques that deviate significantly from those used in face-to-face conversations.

Contribution: With this publication we introduce OnCoCo 1.0 (*Online Counseling Conversations*), a new bi-lingual dataset (German and English) for rich content analysis in psychosocial online counseling. Our dataset extends current conversational

corpora by providing a detailed and ethically curated dataset for bilingual counseling contexts. For the classification, we developed a novel and fine-grained category system, using methods of qualitative social research (Mayring, 2015) and grounded in theoretical models of online counseling (e.g., Grawe, 2000). It is composed of 38 categories for counselor messages and 28 categories for client messages. Furthermore, 2,778 messages were manually created and curated by professional experts and specifically advised students. We fine-tuned several models based on this data to demonstrate its applicability. The best model reaches an accuracy of 80% and an F1 macro score of about 0.78 which is comparable to human performance on this kind of data. The developed category system extends previous work in the sense that it is not limited to just one method within counseling, e.g. MI, but corresponds to the special claim of online counseling - i.e. the mixture of different methods. Hence, our system and the corresponding models are suitable for analyzing all types of texts during counseling or therapy sessions.

Availability and License: The OnCoCo 1.0 dataset, including documentation, metadata and limitations, as well as ready-to-use fine-tuned models are publicly available for research and educational purposes under a CC BY 4.0 license. The data and models can be accessed at <https://huggingface.co/datasets/th-nuernberg/OnCoCoV1>. A full metadata description is also provided via the LRE Map.

2. Relevance of our Research

The ability to analyze unlimited amounts of conversations with a rich system of categories offers completely new possibilities for online counseling research. Fine-grained patterns can be collected and transferred in complex and comprehensive statistical models. This significantly expands our knowledge to relevant aspects of successful counseling.

Social Interaction Pattern Analysis: The fine-grained and role-specific coding scheme enables detailed analyses of counseling interaction patterns and social dynamics. For example, it supports (i) *strategy-response* analyses by studying transitions between counselor categories (e.g., problem clarification vs. motivation building) and subsequent client categories (e.g., engagement, resistance, or requests for help), and (ii) *conversation stage* analyses by comparing category distributions between early and late turns. These analyses are difficult to conduct with coarse tag sets and illustrate how the resource can be used in qualitative and mixed-methods research.

Quality Improvement: Classification models trained on our dataset can be used to identify patterns and success factors for online counseling, such as problem solving, relationship-building, interventions etc. (see Table 1). Text analysis is particularly well suited to explore the micro-level of counseling relationships such as interaction patterns, emotional dynamics, and conversational flow. In face-to-face counseling, conversation and text analyses of recorded sessions have already provided valuable insights into interaction processes and relationship-building (Hancken, 2023). These approaches can be adapted to the context of online counseling. AI-supported analysis also enables the examination of large volumes of counseling transcripts, helping to detect and quantify latent patterns and meaningful structures. This can assist professionals in reflecting on their methods and incorporating successful interaction strategies into their practice or training. Beyond individual counselors, supervisors can also use these insights for case reviews, team reflection and structured monitoring processes to improve quality assurance.

Education: Counselor education can be supported by providing valuable insights and resources. For instance, trainees can receive real-time feedback on their performance by having their counseling sessions automatically analyzed by the model. It could highlight areas in which the counselor needs to work on developing empathy strategies, setting goals, intervention techniques, or activating resources. This feedback can be included into the educational process to direct future development possibilities. Additionally, trainees can learn to identify patterns in interactions between counselors and clients. Since the model may categorize various aspects of a conversation, such as the counselor's use of motivational strategies, educators can pinpoint certain strengths and shortcomings in student performance. In addition to real-world sessions, the model can help to create simulated environments for counselor training. By generating potential interactions between the counselor and the client (or analyzing past simulated dialogues), trainees can practice identifying appropriate responses, intervention strategies, and emotional cues.

Detection of Common Issues: Emerging trends and prevalent issues, such as those that teenagers encountered during the COVID-19 pandemic, can be found with the aid of automated analysis. Early trend detection allows for prompt and focused intervention in social and mental health issues, which shortens the duration of these problems for impacted individuals and lessens their wider societal impact.

Evaluation of Chatbots: Studies based on our dataset enable a fine-grained evaluation of counsel-

ing interactions conducted by counseling chatbots. On this basis, AI-based counseling systems can first be assessed in terms of their counseling quality and subsequently optimized to reach the standard set by human counselors.

Efficient Resource Allocation: In depth data analysis within counseling organizations can support better resource allocation in social work. For example, counselors can be directed to cases requiring urgent attention (e.g., high-risk situations such as suicide or child welfare concerns), thus optimizing the use of human resources and offering better support for clients in urgent need of help.

3. Related Work

In the context of online counseling, content analysis has been investigated in some work during the last decade. (Althoff et al., 2016) developed models to evaluate conversation strategies such as adaptability, progress, and shifts of perspective based on an SMS-data that are not publicly available.

Several other publications are based on Motivational Interviewing (MI) (Miller and Rollnick, 2002). The *Motivational Interviewing Skill Code* (MISC) is a coding scheme designed to analyze both client and counselor behavior in MI sessions (Miller et al., 2008). The MISC category system comprises a total of 28 categories (19 for counselor, 9 for clients). (Xiao et al., 2016) and (Cao et al., 2019) applied machine learning to MISC, but used a simplified set of just 8 counselor and 2 client categories because the support for some categories in their dataset was considered too low.

(Pérez-Rosas et al., 2016) describe the creation of a dataset based on the even simpler *MI Treatment Integrity* (MITI) coding system, which focuses solely on the counselor. Utterances are categorized into MI-adherent (e.g. simple/complex reflection, affirmation) and MI-non-adherent (e.g. direct persuasion or confrontation). Their dataset contains a total of 22,719 coded utterances. Later on, it was used to analyze counseling conversations with regard to empathy and overall quality of counseling (Pérez-Rosas et al., 2017, 2019). Although stated otherwise in the publications, this dataset is currently not available to the public.

Anno-MI, in contrast, is a public dataset that includes transcribed therapy dialogues on MI annotated by professional therapists (Wu et al., 2022, 2023). It consists of 133 videotaped and transcribed conversations with a total of over 9.699 utterances. The annotation scheme is, however, rather coarse, consisting of just the three major therapist behaviours: reflection (simple/complex), question (open/closed), and input (information, advice, giving options, and negotiation). The subcategories (in parentheses) were defined but not used

for annotation because of low inter-coder agreement in a first test. Client messages were also differentiated into three categories, namely change, neutral, and sustain similar to MISC.

Based on this previous work on MI, (Cohen et al., 2024) recently created a new public MI dataset consisting of 242 sessions with 15.627 total messages. Therapist behavior is classified into 11 categories based on MITI.

The dataset *ESConv* (Emotional Support Conversation) as well as a corresponding framework for the modeling of dialogues and corresponding categories were published by (Liu et al., 2021). The coding scheme puts more emphasis on the client side and distinguishes between five types of client problems, seven emotions, and a feedback score between one and five. Eight support strategies are annotated on the counselor side. The publicly available dataset includes 1,053 conversations consisting of over 31,000 utterances.

The public *HOPE* dataset (Malhotra et al., 2022) contains approximately 12,900 statements from 212 therapy sessions in the field of cognitive behavioral therapy (CBT). For the annotation, a category system consisting of 12 categories, organized into three main categories, was developed.

In addition, there exist general dialogue-act frameworks such as ISO 24617-2 (Bunt et al., 2020) which can serve as a shared reference point for relating the task- and counseling-specific annotation schemes discussed above.

Conclusion Many existing datasets are not publicly accessible, and those that are often use overly coarse annotation schemes that prevent detailed analysis, or are based on a face-to-face counseling setting. Furthermore, it is crucial to differentiate between *counseling* (social) and *therapy* (medical), as they use different methodological approaches. Despite these differences, there are numerous similarities that have influenced the subsequent heterogeneous selection of categories in our study.

4. The Category System

To enable a structured and fine-grained analysis of online counseling conversations, we developed a hierarchical category system that captures diverse aspects of both client and counselor messages.

4.1. Objectives

Psychosocial online counseling is a professional support service, but assessing how established counseling factors (Grawe, 2000) are reflected in large-scale text interactions requires scalable, fine-grained analysis beyond traditional resource-intensive methods.

Role	Categories on Level 1	Categories on Level 2	#Subcategories	Samples
Counselor (CO)	Formalities at the beginning		1	115
	Moderation		1	90
	Impact factors	Analysis and clarification of problems	13	873
		Analysis and agreement on objectives	2	56
		Creating motivation	4	131
		Resource activation	5	117
		Help, problem solving	8	71
	Formalities for conclusion and farewell		2	60
	Other statements		2	19
Total		38	1532	
Client (CL)	Formalities at the beginning		4	72
	Empathy		3	95
	Impact factors	Analysis and clarification of problems	8	437
		Analysis and agreement of objectives	2	93
		Creating motivation	2	75
		Resource activation	2	63
		Help coping with problems	6	282
	Formalities for conclusion		2	70
	Other statements		2	59
Total		28	1246	

Table 1: Overview of the category hierarchy and label distribution.

As the basis for the analysis, a category system was developed according to the principles of qualitative social research (Mayring, 2015). In the first step, analytical criteria were deductively derived from literature, primarily the key factors of effective counseling according to (Grawe, 2000), as well as other relevant structural elements of professional online counseling. For example, several categories derived from the theory of MI were integrated (Miller and Rollnick, 2012). Care was taken to ensure compatibility with the respective datasets mentioned in Section 3, in order to maintain optimal comparability. These individual elements were mapped to clear definitions, and typical anchor examples for corresponding text segments were established.

Using the first draft of the category system, suitable texts were annotated by specially trained human coders. In iterative team meetings, the results were compared and the category system was inductively developed further. For example, very similar categories were merged, or new categories were introduced to capture content that had not been adequately represented up to that point.

4.2. Structure

Based on these principles we created a new category system, consisting of 38 categories for counselor (CO) messages and 28 categories for client (CL) messages. The category system is hierarchically structured into at most five levels of detail. Table 1 shows the first two levels of the category system together with the respective number of cate-

gories at the lowest level (Level 5) and the number of annotated samples in the dataset. The names of all level-5 categories can be found in Table 4.

With 38 different message types, the counselor side is more detailed than the client side. This is by design given that assessments of counseling quality and method adherence are driven foremost by counselor messages. We use fine-grained categories for the most interesting type of messages for analysis, the so-called *Impact Factors*.

For example the impact factors at the counselor's level-2 subcategory *Resource activation* are further distinguished into *Question about possible support resources*, *Request about problem statement*, *Suggestion for activating resources at professional level*, *Suggestion for activating resources at family level*, and *Suggestion for activating resources at friendship level*. This kind of fine distinction between semantic categories allows for very detailed analysis. However, many categories were defined at level four without further subdivision levels, because the level of detail was considered sufficient for the types of analyses we aim for. For example, all MI-related categories can be found on level 4.

The highest resolution of five levels has the subcategory *Impact factors* → *Analysis and clarification of problems*, because we want to analyze in detail the different types of questions and methods which counselors use to get to the core of the client's problem. Some examples can be found in Table 2.

We created unique short codes for each of the categories for simpler data processing. Each cat-

Category Code	Writer	Text	Category (Hierarchical Path)
CO-IF-AC-RF-RLS-L	Counselor	To advise you better, it would be helpful to know what kind of hobbies or interests you have.	Impact factors → Analysis and clarification of problems → Reflection (Fact) → Request about living situation → Leisure
CO-IF-AC-RF-RTP-*	Counselor	Can you describe the feeling you have afterwards a bit more closely? Is it more disgust or does it seem somehow wrong to you what you did?	Impact factors → Analysis and clarification of problems → Reflection (Fact) → Targeted, precise request
CO-IF-HP-*.ITFE-*	Counselor	It is completely normal for a young person to start developing sexual desires and fantasies during puberty.	Impact Factors → Help, Problem Solving → Technical or factual explanations
CL-E-*.ECP-*	Client	It's more about my brother than about me.	Empathy → Concern for another person
CL-IF-ACP-*.FPA-*	Client	I have tried to alleviate my withdrawal symptoms with medication, but it feels like a constant struggle.	Impact factors → Analysis and clarification of problems → Feedback on previous attempts

Table 2: Example utterances from counselors and clients and the respective category

egory at each level is assigned such a short code (values in the parentheses in the example above). The complete code of a category at lower level in the hierarchy is the concatenation of codes along the whole path, i.e. for *Social relationships* it is *CO-IF-AC-RF-RLS-SR*. These codes uniquely identify the categories at the finest level of detail and form the labels we trained the classifiers on.

Our categories can be seen as a domain-specific extension of broader dialogue-act tagging frameworks (e.g., ISO 24617-2). At a coarse level, categories such as *Formalities* correspond to *Social Obligations Management* (e.g., greeting, closing), *Moderation* corresponds to *Interaction Management* (e.g., turn/grounding management), and parts of *Impact factors* map to *Task* and *Feedback* functions (e.g., information-seeking, advising, checking understanding), while refining them into counseling-specific strategies (e.g., resource activation, motivation building). This positioning facilitates comparison to prior work using general dialogue-act standards while preserving the counseling-specific granularity of OnCoCo.

5. The OnCoCo 1.0 Dataset

To enable a fine-grained analysis of interaction patterns between counselors and clients, we created a new bilingual dataset (German with English translations).

5.1. Curation and Annotation

The creation of the dataset is a critical step in ensuring the validity and applicability of the fine-grained message classification system for online counseling conversations. Particular care was taken to ensure privacy, ethical integrity, and data quality.

Origin of Data: The research team deliberately refrained from anonymization existing real-world counseling datasets, as these either lacked explicit consent for research use or involved ethical concerns regarding the use of publicly available counseling content (Ghanem et al., 2022; Bruckman, 2002). Instead, the social scientists took inspiration from real data in public discussion forums and educational role-play sessions. They also drew on their extensive prior research experience to develop synthetic interactions that closely mirror the linguistic patterns, emotional dynamics, and structural characteristics of real counseling exchanges. In multiple iterative steps, the generated sample texts were evaluated for authenticity and compared to real data in order to ensure a high degree of realism and representativeness. In addition, the data was carefully curated and reviewed to minimize potential biases and avoid the reinforcement of stereotypes.

Privacy and Ethics: As the data were generated by experts, no personal data are contained in the dataset. Thus, no further ethical approval was required to publish the data.

Annotation Process: (Scheuerman et al., 2020) rightly point out that the annotation of datasets can introduce unconscious biases into classification processes. This effect can never be entirely ruled out. The annotation and category development in this study were carried out by individuals with the following identities and backgrounds:

The annotation team consisted of six trained online counselors and five student assistants, including students of social work and computer science. All annotators were born in Germany and held a university entrance qualification. The group included 9 women and 2 men.

Thus, on the one hand, it cannot be ruled out that

the lack of diversity among the annotators may have introduced implicit bias into the dataset. On the other hand, the bachelor’s program in social work in Germany includes targeted training designed to help students recognize and reflect on their own biases and engage in their professional work with cultural humility and awareness.

The annotators were instructed in detail in the underlying methodology of the category system. To ensure high data quality in this complex social domain, the labels were established through expert consensus. Annotators discussed disagreements in regular meetings until a shared interpretation of the 66 categories was reached. Final labels and category definitions were reviewed by a lead expert in online counseling to ensure theoretical validity and consistent use across the hierarchy. All student assistants had previous experience with the annotation of counseling data within the field of psychosocial counseling. This whole group was professionally guided by the project coordinator and the project management.

Translation: All conversations were originally written in German and then automatically translated into English. The translation was done using GPT-4o. A review based on random samples confirmed the expected high quality of the translations.

5.2. Description of the Dataset

The actual dataset comprises a total of 2,778 original messages. The data distribution is skewed with regard to the categories (Table 1). This is due to the naturally heterogeneous distribution of messages in counseling dialogues. Most of the client and counselor data belongs to the respective top-level category *Impact factors*, as this is the focus of methodological analysis. Table 2 gives an impression of what the text examples look like.

6. Classification

The dataset in its current form is created for classification. The related work described in Section 3 showed that transformer-based models are generally well suited for this purpose. However, existing data sets use just a few classes while our category system consists of 38 classes for counselor messages and 28 for clients. Multi-class classification with a high number of classes is always a challenge, especially if some classes are semantically close like *"Generation of motivation"* and *"Resource activation"*. The challenge becomes even more difficult if the number of records per category is low.

6.1. Experimental Setup

To demonstrate the usefulness of our data for classification, we chose a set of available multi-lingual encoder models and performed full fine-tuning for classification on our dataset (see Table 3). Our first choice fell on XLM RoBERTa which has shown high performance on numerous classification tasks. In addition, we included the two more recently published model families ModernBERT (Warner et al., 2024) and EuroBERT (Boizard et al., 2025). For each of these, we fine-tuned a smaller base model and a larger variant.

We experimented with separate models for counselors and clients, since the author of a message is known a priori and does not require classification. It turned out, however, that a single model trained on both types of messages is on par with separate models if the messages are simply prefixed with either "Counselor:" or "Client:". One drawback of the unified model is the potential for confusion between counselor and client categories during inference. To address this, we applied output masking by suppressing client-specific categories for counselor messages and vice versa at the level of the prediction logits. We also experimented with applying this masking during training, but observed no measurable improvement in performance.

6.2. Experimental Results

In order to report robust results, we performed 5-fold cross-validation with each model. Table 3 gives an overview of the results. We report the mean and sample standard deviation of the evaluation metrics across the validation data of the five folds. Our primary evaluation metric was the F1 Macro score, i.e. the unweighted average of the F1 score per class because the mean accuracy gives little indication of possibly low recall of classes with little support. As some categories are semantically very close and even for humans hard to distinguish, we also evaluated the second-best choice of the model (Accuracy@2 and F1@2).

The best performing model, XLM RoBERTa, reaches an accuracy of 0.79 and an F1 Macro of 0.72, increasing by over 10 points to 0.88 and 0.83 by taking also the second best prediction into account. Most categories could be detected with high recall and even categories with very little training data are recognized relatively well by this model, as can be seen in Table 4. ModernBERT and EuroBERT also perform well, but lag behind by a few points in both accuracy and F1 score. There is also a measurable performance degradation if smaller base models are used.

Beyond quantitative metrics, a qualitative examination of the outputs of the best performing model reveals that our OnCoCo dataset successfully cap-

Model	Parameters	Accuracy		F1 Macro	
		@1	@2	@1	@2
FacebookAI/xlm-roberta-large	561M	0.79 \pm 0.02	0.88 \pm 0.01	0.72 \pm 0.02	0.83 \pm 0.02
FacebookAI/xlm-roberta-base	279M	0.76 \pm 0.02	0.86 \pm 0.01	0.68 \pm 0.03	0.79 \pm 0.02
answerdotai/ModernBERT-large	396M	0.77 \pm 0.01	0.86 \pm 0.01	0.69 \pm 0.01	0.80 \pm 0.02
answerdotai/ModernBERT-base	150M	0.68 \pm 0.02	0.78 \pm 0.00	0.57 \pm 0.02	0.69 \pm 0.01
EuroBERT/EuroBERT-610m	610M	0.76 \pm 0.01	0.84 \pm 0.01	0.69 \pm 0.02	0.79 \pm 0.02
EuroBERT/EuroBERT-210m	210M	0.75 \pm 0.01	0.84 \pm 0.01	0.67 \pm 0.02	0.79 \pm 0.03

Table 3: Classification results for selected fine-tuned models on cross validation data, showing mean and sample standard deviation across five folds and second best choice (@2).

tures the core interaction structures of professional online counseling. The models perform strongly in identifying categories related to problem clarification, empathy, and resource activation—key mechanisms known from counseling research to contribute to effective psychosocial support (Grawe, 2000; Eichenberg and Küsel, 2016). The classification behavior generally aligns with theoretical expectations: counselor messages emphasizing clarification, motivational support, and structured counseling display high internal coherence and distinct linguistic markers, while categories that rely on subtle affective or contextual cues remain more challenging. This suggests that the OnCoCo category system and trained models are capable of reproducing the interpretive logic of human coders and, therefore, offer a bridge between qualitative content analysis and computational modeling in social work research.

Categories with low F1 scores in Table 4 mainly reflect semantic overlap and data sparsity. For example, the client categories *Problem statement* and *Problem definition* are often confused because both describe aspects of clients’ difficulties, differing only in abstraction level. Similarly, the counselor categories *Targeted*, *precise request* and *Request for change and development* show pragmatic similarity, leading to misclassification. In these cases, the second best (@2) choice of the model often matches the label. Rare types such as *Complex reflection* or *Future forecast* suffer from too few examples for stable learning. Motivational categories like *Eliciting/Evoking Change-Talk* also remain challenging, as such utterances are short and context-dependent. Overall, explicit or formulaic acts (e.g., greetings, farewells) are classified reliably, while categories requiring pragmatic inference or emotional interpretation yield lower performance—mirroring known limits of text-only models in capturing counseling subtleties.

6.3. Intercoder Reliability

Human-to-human: To assess the reliability of the annotation process, we conducted a systematic

inter-annotator agreement (IAA) study. All items from the dataset were independently coded by two human annotators from the trained annotation team, covering all categories of both message roles. The Macro-averaged Cohen’s κ across all items is 0.84 (0.85 for counselor and 0.80 for client utterances) and can be considered very strong. As expected in fine-grained counseling dialogue, agreement was higher for structural categories (e.g., *Greetings*, *Formalities*) and more moderate for semantically dense *Impact Factors* such as *Resource Activation*. However, the overall Macro-averaged κ of 0.84 indicates that the coding scheme remains robust and learnable across these varying levels of complexity. Gold-standard labels were finally established through expert consensus: disagreements were discussed between the annotators and a lead expert in psychosocial online counseling until a shared interpretation was reached.

Human-to-model: We additionally assessed the agreement between the best-performing model (XLM-RoBERTa-large) and codings from human experts. A stratified sample of 20% of the original German utterances (N=556), evaluated against the same gold-standard labels as the classification task, yielded a Cohen’s κ of 0.88. This high value strongly indicates that the trained model’s predictions are conceptually coherent with human expert interpretation. As summarized in Table 4, while the cross-validated Macro F1 (0.72) is numerically lower, largely due to the metric’s high sensitivity to rare categories in a 66-class set, the proximity between the model’s performance and the initial human inter-annotator agreement ($\kappa = 0.84$) demonstrates that the automated classification is approaching the performance ceiling for this granular task.

7. Summary and Further Research

This paper introduced OnCoCo 1.0, a bilingual, fine-grained dataset and category system for the automated analysis of psychosocial online counseling conversations. Integrating diverse counsel-

Code	Level 5	Precision	Recall	F1 Score	Support
CO-FA-**-**	Formalities at the beginning/Opening of the conversation	0.90	0.96	0.93	46
CO-FC-**-F*	Farewell	1.00	0.95	0.97	20
CO-FC-**-OPR*	Offer to use the counselor's professional resources	0.67	1.00	0.80	4
CO-IF-AC-RE-RCR-*	Complex Reflection	0.50	0.67	0.57	6
CO-IF-AC-RE-RES-*	Request for emotional state	0.68	0.72	0.70	18
CO-IF-AC-RF-RC-*	Request for concerns	0.63	0.79	0.70	24
CO-IF-AC-RF-RCD-*	Request for change and development	0.86	0.40	0.55	30
CO-IF-AC-RF-RLS-ES	Employment/economic situation	0.42	0.42	0.42	12
CO-IF-AC-RF-RLS-H	Health (mental and physical)	0.83	0.67	0.74	30
CO-IF-AC-RF-RLS-L	Leisure	0.82	0.82	0.82	34
CO-IF-AC-RF-RLS-PS	Professional situation	0.70	0.82	0.76	34
CO-IF-AC-RF-RLS-SR	Social relationships	0.72	0.91	0.81	34
CO-IF-AC-RF-RPA-*	Request for previous attempts at solutions	0.88	0.91	0.89	32
CO-IF-AC-RF-RPD-*	Request for personal data	0.89	0.61	0.72	28
CO-IF-AC-RF-RTP-*	Targeted, precise request	0.39	0.45	0.42	20
CO-IF-AC-RF-SRx-*	Simple reflection	0.94	0.96	0.95	52
CO-IF-AO-**-ICO-*	Definition of counseling objectives	0.64	0.58	0.61	12
CO-IF-AO-**-ROW-*	Request for objectives/wishes	0.37	0.70	0.48	10
CO-IF-HP-**-ICO-*	Calming	1.00	1.00	1.00	2
CO-IF-HP-**-IEA	Evaluation, Interpretation	0.60	1.00	0.75	6
CO-IF-HP-**-IF-*	Future forecast	0.00	0.00	0.00	2
CO-IF-HP-**-IPFR-*	Professional, formal, institutional recommendation	1.00	1.00	1.00	2
CO-IF-HP-**-ITFE-*	Technical or factual explanations	1.00	1.00	1.00	6
CO-IF-HP-**-IW-*	Warning	1.00	1.00	1.00	4
CO-IF-HP-**-PP-IA	Advice	0.67	1.00	0.80	2
CO-IF-HP-**-PP-IW	Wish	1.00	1.00	1.00	4
CO-IF-Mot-**-IAC-*	Articulation of the perceived ability to change in the client (MI)	1.00	0.70	0.82	10
CO-IF-Mot-**-IEM-*	Encouragement, motivation of the client	0.82	1.00	0.90	14
CO-IF-Mot-**-ITA-*	Thanks and appreciation to the client	1.00	0.86	0.92	14
CO-IF-Mot-**-RFC-*	Eliciting/Evoking "Change-Talk" (MI)	0.80	0.57	0.67	14
CO-IF-Mot-**-RS-*	Question about possible support resources	0.64	0.64	0.64	14
CO-IF-RA-**-N-RAFa	Suggestion for resource activation at the family level	1.00	1.00	1.00	10
CO-IF-RA-**-N-RAFr	Suggestion for resource activation at the friendship level	1.00	1.00	1.00	2
CO-IF-RA-**-RAP-*	Suggestion for resource activation at professional level	1.00	1.00	1.00	14
CO-IF-RA-**-RP-*	Request for problem statement	1.00	0.67	0.80	6
CO-Mod-**-**	Moderation	0.81	0.81	0.81	36
CO-O-**-O-*	Other statements	0.00	0.00	0.00	2
CO-O-**-UCO-*	Inappropriate remark	1.00	0.50	0.67	6
CL-E-**-ECC-*	Compassion for others (EC)	0.83	0.71	0.77	14
CL-E-**-ECP-*	Concern for another person (EC)	0.61	0.92	0.73	12
CL-E-**-PT-*	Empathy for third parties or related to the present situation (PT)	0.78	0.58	0.67	12
CL-FB-**-**	Formalities at the beginning and greeting	1.00	1.00	1.00	28
CL-FC-**-F-*	Formalities for conclusion	1.00	1.00	1.00	8
CL-FC-**-UPR-*	Further use of professional resources of counselor	0.83	0.95	0.88	20
CL-IF-ACP-**-Cons-*	Consent	0.78	1.00	0.88	14
CL-IF-ACP-**-DPD-*	Disclosure of personal data	1.00	0.95	0.97	20
CL-IF-ACP-**-FPA-*	Feedback on previous attempts at solutions	0.88	0.85	0.86	26
CL-IF-ACP-**-OE-*	Own emotional expression	0.95	0.95	0.95	22
CL-IF-ACP-**-PD-*	Problem definition	0.44	0.79	0.56	14
CL-IF-ACP-**-PS-*	Problem statement	0.72	0.70	0.71	30
CL-IF-ACP-**-Rej-*	Rejection	0.85	0.85	0.85	20
CL-IF-ACP-**-Req-*	General request	1.00	0.93	0.96	28
CL-IF-AO-**-Ext-*	Extension of the assignment	0.83	0.36	0.50	14
CL-IF-AO-**-Obj-*	Objective of the assignment	0.67	0.83	0.74	24
CL-IF-HP-**-Fail-*	Final failure	0.93	0.72	0.81	18
CL-IF-HP-**-NegFR-*	Negative feedback on specific action recommendation	0.68	0.77	0.72	22
CL-IF-HP-**-PosF-*	General positive feedback	0.83	0.94	0.88	16
CL-IF-HP-**-PosFR-*	Positive feedback on specific recommendations for action	1.00	0.88	0.94	26
CL-IF-HP-**-RepRA-*	Report on the implementation of recommendations for action	0.91	0.83	0.87	12
CL-IF-HP-**-Succ-*	Final success	1.00	0.89	0.94	18
CL-IF-Mot-**-FC-*	Eliciting/Evoking "Change-Talk" (MI)	0.53	0.67	0.59	12
CL-IF-Mot-**-RC-*	Articulation of reasons for a change in the client	0.80	0.67	0.73	18
CL-IF-RA-**-RF-*	Considering resource activation at the level of friends and family	0.89	1.00	0.94	8
CL-IF-RA-**-RP-*	Considering resource activation at a professional level	0.73	0.50	0.59	16
CL-O-**-O-*	Other statements	1.00	0.50	0.67	8
CL-O-**-UCO-*	Inappropriate remark	0.88	0.94	0.91	16

Table 4: Classification report for the best performing model xlm-roberta-large on the 1,112 test samples

ing approaches, the system enables the detailed classification of counselor-client interactions and supports scalable, data-driven evaluation of counseling practices. The created OnCoCo dataset, together with fine-tuned classification models, offers significant potential for improving the quality and efficiency of psychosocial support services.

While the models already achieve satisfactory results for most categories, certain fine-grained types remain difficult to distinguish due to semantic overlap, subtle contextual cues, and limited

data support. Future research should focus on enhancing the modeling of pragmatic and affective dimensions, for instance by integrating discourse-level features or multi-task learning objectives, to strengthen the classification of conceptually overlapping categories. Further improvements could also result from expanding the dataset with additional annotated examples for underrepresented categories or by leveraging large language models to generate high-quality synthetic samples.

8. Limitations

Our dataset and the models come with several limitations:

Skewed Data Distribution: The dataset has an uneven distribution of categories, which may lead to challenges in detecting underrepresented categories. While it is planned to expand the dataset, the current imbalance affects model performance.

Potential Generalization Issues: Although the models achieve high accuracy and F1, the performance may drop on data coming from types of online counseling which were not covered by the training data.

Limited Cultural Adaptability: The dataset, though bilingual (German and English), does not explicitly address cultural differences in counseling or how these might affect the applicability of the models across diverse populations and languages.

Narrow Scope of Application: While the dataset and models are focused on online counseling, their application is limited to this domain. Expanding the models to broader contexts, such as peer support or general social work, would enhance their societal impact.

Sociolinguistic and Ethical Limitations: From a qualitative and sociological perspective, the dataset represents online counseling dialogues under idealized professional conditions. This focus ensures ethical soundness and methodological consistency, but also constrains the variety of linguistic and relational phenomena captured. Informal, ambiguous, or boundary-crossing communication (e.g. such as humor expressions, self-disclosure, or other indirect emotional cues) is rare or even absent. While this reflects the professional standards of counseling practice, it can reduce the ability of the model to generalize to less formal contexts, where such elements are relevant to building rapport. In addition, the synthetic, but expert-generated nature of the dialogues means that spontaneous and situational nuances of real counseling interactions are only partially comprised.

9. Bibliographical References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. [Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health](#). *Transactions of the*

Association for Computational Linguistics, 4:463–476.

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M. Alves, et al. 2025. [EuroBERT: Scaling Multilingual Encoders for European Languages](#).

Amy Bruckman. 2002. [Studying the amateur artist: A perspective on disguising data collected in human subjects research on the Internet](#). *Ethics and Information Technology*, 4(3):217–231.

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot. 2020. [The ISO standard for dialogue act annotation, second edition](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 549–558, Marseille, France. European Language Resources Association.

Jie Cao, Michael Tanana, Zac E. Imel, Eric Poitras, David C. Atkins, and Vivek Srikumar. 2019. [Observing Dialogue in Therapy: Categorizing and Forecasting Behavioral Codes](#).

Ben Cohen, Moreah Zisquit, Stav Yosef, Doron Friedman, and Kfir Bar. 2024. [Motivational Interviewing Transcripts Annotated with Global Scores](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11642–11657. ELRA and ICCL.

Christiane Eichenberg and Cornelia Küsel. 2016. [Zur Wirksamkeit von Online-Beratung und Online-Psychotherapie](#). *E-Journal für biopsychosoziale Dialoge in Psychotherapie, Supervision und Beratung*, 4(2):93–107.

Christian Ghanem, Markus Eckl, and Robert Lehmann. 2022. [Big data und forschungsethik in der sozialen arbeit](#). *EthikJournal*, 8(1).

Philipp Grandeit, Carolyn Haberkern, Maximiliane Lang, Jens Albrecht, and Robert Lehmann. 2020. [Using BERT for Qualitative Content Analysis in Psychosocial Online Counseling](#). In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 11–23. Association for Computational Linguistics.

Klaus Grawe. 2000. *Psychologische Therapie, Göttingen, 2.*, korrigierte auflage edition. Hogrefe.

Sabrina Amanda Hancken. 2023. [Beziehungsgestaltung in der Sozialen Arbeit, 2.](#), durchgesehene auflage edition. Number 6015 in utb. Vandenhoeck & Ruprecht.

- S. Gabe Hatch, Zachary T. Goodman, Laura Vowels, et al. 2025. [When ELIZA meets therapists: A Turing test for the heart and mind](#). *PLOS Mental Health*, 2(2):e0000145.
- Ron Kraus, George Stricker, and Cedric Speyer. 2010. *Online Counseling: A Handbook for Mental Health Professionals*, 2nd ed edition. Practical Resources for the Mental Health Professional. Elsevier Academic Press.
- Kevin M. Laska, Alan S. Gurman, and Bruce E. Wampold. 2014. [Expanding the lens of evidence-based practice in psychotherapy: A common factors perspective](#). *Psychotherapy*, 51(4):467–481.
- Siyang Liu, Chujie Zheng, Orianna Demasi, et al. 2021. [Towards Emotional Support Dialog Systems](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483. Association for Computational Linguistics.
- Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. [Speaker and Time-aware Joint Contextual Learning for Dialogue-act Classification in Counselling Conversations](#). In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 735–745. ACM.
- Philipp Mayring. 2015. [Qualitative Content Analysis: Theoretical Background and Procedures](#). In Angelika Bikner-Ahsbals, Christine Knipping, and Norma Presmeg, editors, *Approaches to Qualitative Research in Mathematics Education*, pages 365–380. Springer Netherlands.
- William R. Miller, Theresa B. Moyers, et al. 2008. *Manual for the Motivational Interviewing Skill Code (MISC), Version 2.1*.
- William R. Miller and Stephen Rollnick. 2002. *Motivational Interviewing: Preparing People for Change*, 2nd ed edition. Guilford Press.
- William R. Miller and Stephen Rollnick. 2012. *Motivational Interviewing: Helping People Change*. Guilford press.
- Theresa B. Moyers, Lauren N. Rowell, Jennifer K. Manuel, Denise Ernst, and Jon M. Houck. 2016. [The Motivational Interviewing Treatment Integrity Code \(MITI 4\): Rationale, Preliminary Reliability and Validity](#). *Journal of Substance Abuse Treatment*, 65:36–42.
- NHS. 2024. Record numbers of women accessing perinatal mental health support.
- Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. [Understanding and Predicting Empathic Behavior in Counseling Therapy](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1435. Association for Computational Linguistics.
- Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, et al. 2016. [Building a Motivational Interviewing Dataset](#). In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 42–51. Association for Computational Linguistics.
- Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. [What Makes a Good Counselor? Learning to Distinguish between High-quality and Low-quality Counseling Conversations](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935. Association for Computational Linguistics.
- Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. 2020. [How We’ve Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis](#). *Proc. ACM Hum.-Comput. Interact.*, 4:58:1–58:35.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, et al. 2024. [Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference](#).
- WHO. 2021. *Mental Health Atlas 2020*, 1st ed edition. World Health Organization.
- Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023. [Creation, Analysis and Evaluation of AnnoMI, a Dataset of Expert-Annotated Counselling Dialogues](#). *Future Internet*, 15(3):110.
- Zixiu Wu, Simone Balloccu, Vivek Kumar, et al. 2022. [Anno-MI: A Dataset of Expert-Annotated Counselling Dialogues](#). In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181. IEEE.
- Bo Xiao, Doğan Can, James Gibson, et al. 2016. [Behavioral Coding of Therapist Language in Addiction Counseling Using Recurrent Neural Networks](#). In *Interspeech 2016*, pages 908–912. ISCA.

Predicting Social Media User Actions: A Hybrid Approach for Common and Rare Behavior Prediction on Bluesky

Benjamin White, Anastasia Shimorina

Orange Research, Lannion, France

{firstname.lastname}@orange.com

Abstract

Understanding and predicting user behavior on social media platforms is crucial for content recommendation and platform design. While existing approaches focus primarily on common actions like retweeting and liking, the prediction of rare but significant behaviors remains largely unexplored. This paper presents a hybrid methodology for social media user behavior prediction that addresses both frequent and infrequent actions across a diverse action vocabulary. We evaluate our approach on a large-scale Bluesky dataset containing 6.4 million conversation threads spanning 12 distinct user actions across 25 persona clusters. Our methodology combines four complementary approaches: (i) a lookup database system based on historical response patterns; (ii) persona-specific LightGBM models with engineered temporal and semantic features for common actions; (iii) a specialized hybrid neural architecture fusing textual and temporal representations for rare action classification; and (iv) generation of text replies. Our persona-specific models achieve an average macro F1-score of 0.64 for common action prediction, while our rare action classifier achieves 0.56 macro F1-score across 10 rare actions. These results demonstrate that effective social media behavior prediction requires tailored modeling strategies recognizing fundamental differences between action types. Our approach achieved first place in the SocialSim: Social-Media Based Personas challenge organized at the Social Simulation with LLMs workshop at the Conference on Language Modeling (COLM 2025).

Keywords: social media, Bluesky, user behaviour prediction, hybrid approaches, persona-based modelling, classification

1. Introduction

Social media platforms have become central to modern communication, generating vast amounts of user interaction data that offer unprecedented insights into human behavior patterns. Understanding and predicting how users engage with content—whether they like, share, comment, or take other actions—has significant implications for both theoretical research in computational social science and practical applications in content recommendation, platform design, and user experience optimization.

The prediction of user behavior on social media platforms presents a complex challenge. Traditional approaches have primarily focused on the most frequent user actions, such as retweeting and liking, often employing hand-crafted features or network-based representations (Suh et al., 2010; Yang et al., 2010; Peng et al., 2011). However, the emergence of large language models (LLMs) has opened new possibilities for more sophisticated user behavior modeling through persona-based simulation and generation (Rossetti et al., 2024; Törnberg et al., 2023).

Despite these advances, several key challenges remain unresolved. First, most existing work has concentrated on binary prediction tasks or a limited set of common actions, leaving the prediction of rare but potentially important behaviors (such as

blocking, unfollowing, or content deletion) relatively unexplored (Wu et al., 2020). Second, recent research suggests that smaller, specialized models may outperform large language models on action classification tasks (Qiu et al., 2025). Third, the integration of temporal dynamics, user personas, and content semantics into a unified prediction framework remains an open challenge.

This work addresses these limitations by developing a comprehensive approach to social media user behavior prediction that handles both common and rare actions while incorporating rich temporal and semantic features. We present a hybrid methodology that combines lookup-based prediction for high-confidence cases, specialized tabular models for common actions, transformer-based architectures for rare action classification, and LLM-based generation of replies.

Our key contributions are threefold: (1) We demonstrate that a portion of user behavior can be predicted through historical pattern matching; (2) We show that persona-specific tabular models using carefully engineered temporal and semantic features outperform transformer-based approaches for common action prediction; and (3) We develop a specialized hybrid neural architecture that effectively handles the class imbalance inherent in rare action prediction by fusing textual and temporal representations.

Our approach achieved first place in the So-

cialSim: Social-Media Based Personas challenge 2025¹, and it demonstrates how the strengths of several NLP approaches to social media text analysis can be combined to more accurately predict and simulate user behavior on a modern social media platform.

2. Background

Social Media User Behavior Prediction User behavior on social media platforms has been extensively studied, given its theoretical interest and industrial importance. With Twitter as a primary source of data, previous work has focused extensively on understanding and predicting retweeting behaviors (Suh et al., 2010; Yang et al., 2010) and using features based on the text message content (Peng et al., 2011) or graph and network properties (Sotiropoulos et al., 2019).

To improve behavior prediction, approaches which incorporate richer user-specific information have been developed. For example, Firdaus et al. (2021) combine users' tweeting and retweeting historical data to construct topic preferences and psychological profiles, and use these enriched persona features to improve retweet prediction F1 scores.

Clustering the potentially very large number of distinct users into a smaller number of persona groups has also been explored; in these approaches, users are grouped by behavior, demographic, or other similarity conditions, and then the user's membership in a particular persona cluster is incorporated in downstream prediction tasks. For example, Sun et al. (2023) used persona information to predict user written replies to Twitter news headlines, with appropriate sentiment intensity and polarity.

To go beyond manual feature engineering, approaches using end-to-end deep learning have attracted much attention: an early approach (Zhang et al., 2016) learnt joint representations of the tweet author, text, and the responding user to predict retweet behavior. Later approaches combine further actions and relationships obtained from the social media platform, such as network properties and message metadata, to improve prediction performance (Meštrović et al., 2022).

While most research has focused on the more frequent actions of retweeting and liking, there has been comparatively less examination of other rarer specific actions. In this direction, Wu et al. (2020) focused entirely on the unfollow action on

the Weibo platform, finding the decision to unfollow another user to be more complex – and thus requiring more sophisticated user and persona representations – than e.g. the decision to follow another user.

Personas, LLMs, and Agentic Simulation With recent improvements in large language models (LLMs), research has explored the potential for using these systems for emulating social media platforms and their human users.

Rossetti et al. (2024) developed a LLM-powered digital twin of a social media platform, using agents characterized by age, interests, and personality traits, and based their modeling of agent activity on real data obtained from Bluesky.

Törnberg et al. (2023) combined LLMs and agent-based modeling by creating realistic personas based on real demographic data, and demonstrate its applicability to studying complex user interaction scenarios.

Going further, Touzel et al. (2024) built a sophisticated environment based on the Mastodon messaging app, enabling persona agents to take full control of their user accounts and emulate human interactions, and implemented a realistic longitudinal survey of the agents' political opinions and voting patterns.

Zhou et al. (2024) simulates user reactions on social media with an agentic approach, and incorporate persona information and grounding by implementing a dedicated persona module that informs the agent's planning and action decisions.

However, recent research has also identified several limitations of LLMs for simulating human behavior in general, outlining the beginning of a more rigorous science of persona generation (Li et al., 2025).

Using LLMs to simulate social media engagement including both subtasks of action prediction and written response generation, has been explored recently by Qiu et al. (2025). Interestingly, the authors find that smaller fine-tuned BERT models outperform larger frontier LLMs on action classification tasks and furthermore that the LLMs exhibit particular sensitivity to the specific prompt context used for the classification task. However they find that LLMs perform well on the text generation subtask, and in particular that few-shot prompting with user information and historical examples improves the semantic alignment with reference tweets. Taken together, especially since their work explored a dataset with only 3 available actions (quote, rewrite, retweet), this recent work suggests that hybrid approaches to social media user simulation may still outperform purely LLM-based approaches, with the strength of LLM components lying in their ability to generate faithful text exam-

¹<https://sites.google.com/view/social-sims-with-llms/shared-task>,
<https://www.kaggle.com/competitions/social-sim-challenge-social-media-based-personas/leaderboard>

ples when augmented with relevant persona data obtained from other feature extraction approaches.

3. Data and Task Description

We describe our submission to the SocialSim challenge 2025, which distributed the persona-based social media dataset from Bluesky (Bück-Kaeffer et al., 2025). The dataset description paper was released only after the challenge had concluded. Therefore, below we present our exploratory data analysis as conducted for the challenge.

We use BLUEPRINT, the persona-based social media dataset from Bluesky (Bück-Kaeffer et al., 2025). Bluesky is a modern microblogging social media platform with a similar interface and user experience to platforms such as X, allowing users to post messages and interact with messages and threads posted by other users. The public training/development dataset version available for the SocialSim 2025 challenge² consists of 6,435,348/213,556 samples, and all results herein refer to evaluations performed on this development split (our results on the competition holdout private test set were similar). Each sample is a conversation thread taken from Bluesky, containing one or more messages in English written by the users who participated in the conversation. Each of the messages in a given thread contains the anonymized user ID, the content of the message, and a relative timestamp indicating when the message was posted to Bluesky. In addition, the users in BLUEPRINT are clustered into 25 different personas: user embeddings were obtained by pooling all individual posts, quotes, and replies authored by each user, and then these user embeddings were clustered to form 25 persona clusters. Inspection and TF-IDF analysis (visible in Table 10 of Bück-Kaeffer et al. (2025)) of the 25 obtained clusters reveals that they correspond to different behavioral and topic groups, with e.g. cluster #22 corresponding mainly to creative workers who discuss art related topics. The final message in each thread of the BLUEPRINT dataset contains the final user’s persona cluster which allows for the analysis of persona specific social behavior in subsequent modeling and analysis.

The modeling task is to predict how a real user, knowing which of the 25 *persona* clusters that user belonged to, would respond to the conversation i.e. firstly to predict the *action* that the user took, and secondly to generate the *text* that they wrote if applicable. Table 1 shows an example from the dataset, and Table 2 provides the complete list of

all 12 possible Bluesky user actions that appear in the training data.

Field	Value
first message time	3,885,851 sec
first message text	Trump was filmed yesterday clearly ...
first user id	dd4724
second message time	8,012,737 sec
second user id	4ffd33
second user cluster	17
second user action	FOLLOW
second message text	None

Table 1: A dataset example where there are only 2 turns in the thread, so that the first user is the original poster and the second user from the persona cluster 17 is the one who is interacting with the post—in this case by deciding to follow the first user. Message times are measured in seconds, and are given relative to an unspecified start time—in this case the second user interacted 47 days after the first message was posted. We show in blue the fields that are to be predicted.

Persona Clusters In total, 25 different persona clusters exist in the dataset. From inspection of the text content of their messages, they mainly represent coherent behavioral groupings, such as scientific communities, sports fandoms, and political affiliations (Bück-Kaeffer et al., 2025). Our manual reviewing of TF-IDF analyses on 25 clusters found that the primary distinction was whether clusters focused on political topics, with political clusters differing mainly in the intensity of political sentiment and country of interest, while non-political clusters were primarily centered on technology, art, or adult content.

Social Media Context Public access to Bluesky opened in February 2024, leading to a rapid increase in users and the inclusion of very recent content, so content in the training data is potentially more up-to-date than that found in frontier LLMs. This influx likely consisted of entire communities with shared interests, resulting in an overrepresentation of certain viewpoints and a higher frequency of positive actions (e.g., LIKE; see Table 2) compared to negative ones. Additionally, since most accounts are new, the high number of FOLLOW actions likely reflects users actively building their networks during this period.

Training Data Simplification For the action prediction subtask, we retained only the conversations of length 2 because we noted that in the training dataset the actions for conversations of length 3

²<https://www.kaggle.com/competitions/social-sim-challenge-social-media-based-personas/data>

were always REPLY. Therefore we decided to systematically predict REPLY if the given sample contained more than 2 messages. This simplification reduced the number of REPLY actions from 47,798 (Table 2) to 39,349.

Task Evaluation The task performance is evaluated using F1-scores (both weighted and macro) for the prediction of the user’s action, and a cosine similarity score for the text generation (i.e. how closely the model’s predicted user text was to the text in the actual reply). Macro F1 was used as the primary metric in the SocialSim 2025 challenge, given the highly imbalanced dataset, to encourage performance across all actions including the rare actions. The cosine similarity is calculated using a text embedding model³.

Action	Count	Percentage
follow	4,386,038	68.16%
like	1,846,842	28.70%
unfollow	97,804	1.52%
reply	47,798	0.74%
quote	41,401	0.64%
unlike	13,209	0.21%
post_update	1,014	0.02%
repost	744	0.01%
block	479	0.01%
post_delete	14	0.00%
unblock	3	0.00%
unrepost	2	0.00%

Table 2: Train dataset actions with counts and percentages. The REPLY action highlighted in green is generative—the action is accompanied by the user also inputting text.

Task Simplification We note here an important aspect of the dataset that simplifies the prediction task and may bias results or prevent generalization to all users of social media platforms: by the nature of the dataset construction, all the samples are associated with a definite user action. However, in the real setting of Bluesky, the choice of not taking any action is itself a possibility—indeed it may be the most frequent “action” across the entire platform as most users presumably do not interact with most messages. Therefore, the subset of messages that appear in the dataset are those which were for one reason or another able to elicit an interaction by a social media user, and may not be representative of the nature of generic social media messages.

³<https://huggingface.co/intfloat/multilingual-e5-large>

4. Methodology

4.1. Prediction Pipeline Summary

In this section, we outline our prediction pipeline and detail each step in subsequent sections.

Each query in the dataset consists of a conversation with 1 or more messages with their text content, and raw relative timestamp data for each message. The query also contains the persona cluster of the user whose action we must predict in response to reading this conversation.

We first measure how many messages there are: if there are 3 or more messages, we predict the user action REPLY, based on our observation of this pattern in the training data as described in Section 3.

For remaining samples with exactly 2 messages, we lookup the exact string of the first user’s message in a lookup database of Bluesky messages. We determine whether we have sufficient information about all *other users*’ interactions with this message to perform a “majority vote” prediction: if many users from the same cluster have all responded in a very similar way, we take this most frequent action as our current prediction. Alternatively, if we do not have data from the same cluster, we also examine the entire user base but with a more selective threshold criterion (Section 4.2).

Next, if the lookup database does not allow us to confidently make a prediction, we use a cluster-specific trained LightGBM model corresponding to the current user’s persona cluster. This model will predict either FOLLOW, LIKE, or OTHER. If this model predicts FOLLOW or LIKE, we take this as the prediction. If the model predicts OTHER, it means that we believe that the user action was one of the “rare actions” (Section 4.3).

If the previous step resulted in the LightGBM model predicting OTHER, we send the query to a specialized rare action classification transformer model, which will determine which of the 10 specific rare actions (BLOCK, UNFOLLOW, REPOST, etc.) to predict for this query (Section 4.4).

Finally, after generating the action label predictions for the entire dataset we filter all those samples where the predicted action is REPLY. We send these conversations to an LLM in order to generate a prediction for the text that the user actually wrote (Section 4.5).

4.2. Repeated Messages Lookup Database

We constructed a database of all first messages that appeared in our dataset and then aggregated, for each message, all of the observed action responses to that message on a cluster-by-cluster level.

Message	Cluster	Like	Follow	Block	Unfollow
"Here is a cute picture of a puppy"	0	9172	23	0	0
	1	254	0	0	0
	2	469	11	0	0
"I like pineapple on pizza"	0	0	0	710	2848
	1	946	671	20	178
	2	7894	541	0	0

Table 3: Artificial examples for the message lookup database step. Here we show 2 artificial examples and only 3 clusters and 4 actions, for clarity. We show different types of unanimity patterns as encountered in the real dataset: here the first message leads to essentially unanimous responses both within each cluster, and across the entire dataset. The second message (more divisive or controversial) leads to unanimous “negative” reactions in Cluster 0, unanimous “positive” reactions in Cluster 2, and heterogeneous reactions within Cluster 1.

We show in Table 3 some artificial (for pedagogical purposes) examples of entries in this lookup database. As with the illustrative examples in the table, real messages also vary in how unanimous the Bluesky users’ responses are. Remarkably however, we found that many messages lead to quasi-unanimous response patterns within persona clusters or even in some cases at the entire global dataset level (i.e. when combining all responses from 25 persona clusters). This therefore suggested that it would be possible use a database lookup and majority-vote approach to generate predictions for such samples.

We show in Figure 1 a sample of 10,000 messages from our lookup database. We refer to the number of times that different users have taken an action in response to a given message as the action “votes” for that message. For a given “voting pattern” (i.e. distribution of different actions) we refer to the most commonly taken action as the “winner” and its total vote frequency as the “winner percentage”. Thus for example a message that 20 users have interacted with, leading to 18 LIKES and 2 BLOCKS, would have the LIKE action as the winner, with a winner percentage of 90%.

When examining the winner percentage for a given message, the total number of votes is an important variable: we consider that it is more significant if 200 users all respond identically to a message rather than if only 2 users respond identically. In the right plot of the figure, the top-right region of the plot shows the messages with both a high total number of votes, and where the votes all tend to agree with one another leading to a high winner percentage; this is where we could be confident in making a prediction for a new user’s action for a message based solely on other users’ responses to that same message.

Since we recorded each message’s action distribution as a function of each of the 25 clusters, we could decide to use solely the votes cast by other users from within the same shared persona cluster, or alternatively consider all the votes across all the

clusters in the global database. We defined 3 possible lookup strategies when using this message database for predictions: Cluster Specific (only take into account votes from other users in the same persona cluster), Global Fallback (if there aren’t sufficiently many data points from within the same cluster, examine the global vote pattern across all users), and finally No Lookup Match (cases where even the Global Fallback strategy does not lead to identifying a “winner action” with a high winner percentage).

The specific thresholds of winner percentage and total votes to use for this lookup-based prediction are adjustable hyperparameters. We fixed a minimum number of total votes of 10 or more, and a “winner percentage” threshold of 85% for the Cluster Specific strategy and of 90% for the Global Fallback strategy. These thresholds were both lowered to 70% for all messages where the “winner action” was one of the rare actions: this is because, after the subsequent main classifier model training steps (see later), we did not obtain any classifiers that achieved better than 70% accuracy for the rare actions. Therefore we found it would always improve overall performance to just take majority-votes based on the lookup database rather than use the trained models.

4.3. Tabular Model Development with Grouped Rare Actions

We selected LightGBM (Ke et al., 2017) as our modeling approach for the action classification task, after finding during initial exploration that deep learning approaches failed to attain good performances on this task: encoder language models (see Table 5 for our baselines) attained macro-F1 scores of only 0.48 even after finetuning specialized social media models, while generative LLM approaches using in-context examples of the available actions on Bluesky produced even lower scores. Bück-Kaeffer et al. (2025, Table 5) demonstrated similar results with LLM approaches,

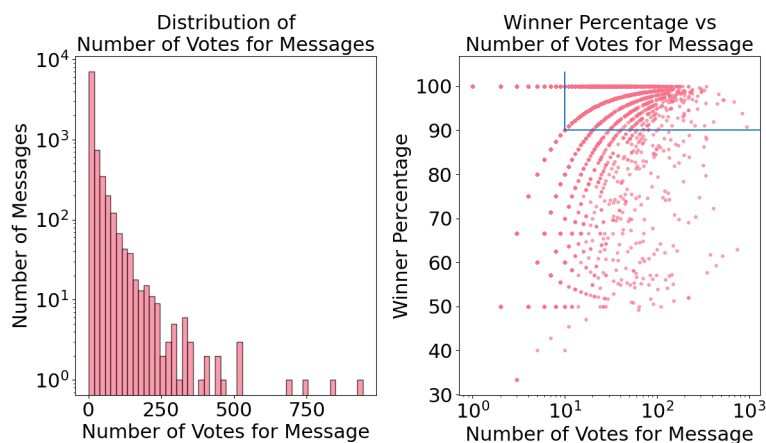


Figure 1: Sample of 10,000 messages from lookup database. In the right plot, the highlighted top-right rectangular region of interest contains messages with both a large number of votes (more than 10) and where the most frequent action has a high winner percentage (here more than 90%) across all clusters.

achieving F1 scores of 0.33 and showing only moderate gains from finetuning.

We developed a distinct LightGBM model for each of the 25 persona clusters: this approach allowed us subsequently to more clearly adjust model thresholds for classifier predictions, to study feature importance for each persona cluster.

In order to reduce the class imbalance, we replaced the 10 individual rare action labels (see Table 2) with a synthetic grouped label containing all of these rare actions (label OTHER). This therefore simplified the LightGBM training task to focus on only 3 possible classification labels: the 2 real and common actions (LIKE and FOLLOW) and 1 synthetic group (OTHER).

For LightGBM training⁴ we used class weights for the 3 action labels to handle imbalanced learning, weighting according to the inverse occurrence frequency of the 3 actions across the given cluster. We performed stratified 5-fold cross validation with hyperparameter tuning (number of estimators, learning rate, maximum depth, etc.), and for each of the 25 models we implemented threshold optimization using precision-recall curves.

4.3.1. Feature Engineering

We developed keyword, textual, and temporal features for LightGBM modelling.

Keyword Features Based on our exploratory data analysis, most Bluesky messages were found to be well-written, thus enabling direct word matching rather than fuzzy string matching for any keywords.

We built an extensive keyword database, organized according to 11 primary topics (those identi-

Category	Subcategory	Keyword Features
Politics	Candidates	biden, trump, harris
	Issues	healthcare, taxes, border
	Process	ballot, campaign, primary
Bluesky	Twitter	elon, leaving, bird app
	Onboarding	new here, how do i
	Community	invite codes, cozy
Gaming	General	steam, xbox, switch
	Streaming	twitch, vtuber

Table 4: Selected examples of keyword features.

fied during the persona cluster analysis, including politics, gaming, tech, etc.) and then specialized secondary subtopics. We show in Table 4 the organization of this keyword database.

We applied all of these keyword features to our dataset: we recorded for each the total count of each keyword in order to have a measure of the “intensity” of each keyword - for example some messages contain the word “Trump” a dozen times, which we view as a different signal compared to a message where this word only appears once.

Textual Features We constructed an additional set of various text features for basic text characteristics and metrics: total character count, word count, question mark count, hashtag count, etc. We also recorded the occurrence of various anonymization tags that occur in the dataset, such as <USERNAME> or <URL>. Some of these textual features can be seen in Table 6.

⁴<https://github.com/microsoft/LightGBM>

Temporal Features For temporal features, we used the following information:

- the time of the first message;
- the time of the second message;
- the time difference between the two messages;
- 7 categorical encodings derived from this time difference (e.g. `is_immediate_reply` if the time difference is less than 1 minute, etc.).

4.4. Specialized Fine-grained Rare Action Classification

We next built a dedicated model for handling samples that had been predicted as OTHER by LightGBM. This model’s training objective was therefore to predict the specific individual rare action.

For this stage, due to the low frequency of the 10 rare actions across the entire dataset let alone within individual persona clusters, we trained a single rare action classification model using all available data rather than training 25 cluster-specific models.

Given that messages associated with rare actions would probably require more sophisticated semantic understanding and temporal features, we implemented a custom hybrid neural architecture to fuse textual representations with temporal behavior pattern representations.

We built a dual-branch neural network (Baltrušaitis et al., 2018), containing both a text branch that takes an encoder language model to obtain a dense embedding for a given sample’s text content as well as a temporal branch that uses dense layers to produce dense temporal embeddings. We then added an early fusion layer that combines both modalities, before sending this fused representation to a final classifier head.

For the text branch of our neural network, we used a model fine-tuned on Twitter⁵, using the CLS token (a 768-dimensional embedding) as our text representation. This particular model has been trained on 154 million tweets from Twitter/X up until a December 2022 cut-off date (Antypas et al., 2023), suggesting that the text content, modern writing style, and named entities might be particularly close to those from our Bluesky dataset.

For the temporal branch of our neural network, we augmented our existing timestamp features with several other features. Since the raw data contained relative timestamps (the actual message time and date was not specified), we created cyclical encodings corresponding to times within a day

⁵<https://huggingface.co/cardiffnlp/twitter-roberta-base-emotion>, with the trained classifier head being ignored

(even if the absolute value, i.e. 08:15am, was not possible to determine, we could still assign a consistent relative timestamp from 0 to 86,400 seconds, i.e. 24h, to all samples in a cluster) and days within a week.

Our temporal branch module consists of a simple sequential module that scales all 12 temporal features as input, then passes through a 256-dimensional linear hidden layer trained with dropout and ReLU activation, followed by a second 128-dimensional linear hidden layer trained with dropout and ReLU activation to produce the temporal embeddings.

We experimented with different fusion architectures: early fusion by simply concatenating the 2 embedding types, late fusion by creating separate classifiers and using a learned weighting scheme, and finally a cross-attention fusion approach allowing temporal queries to attend to the text and using this attended text for classification. Across all hyperparameter searches we found that while attention fusion approaches could in some cases lead to higher performance, the early fusion approach was the most stable with consistently good results, so we retained it for our final model.

To handle class imbalance we implemented a focal loss (Lin et al., 2017) for our training criterion, using the class weights according to inverse occurrence frequency for the 10 rare actions across the training dataset, and a gamma value of 2.0 as its hyperparameter.

We found during development that a two-phase training strategy consisting of a warmup phase, where we froze the text encoder and trained the temporal/fusion layers for 2 epochs, followed by a fine-tuning phase, where we trained the architecture end-to-end for a further 3 epochs, improved overall model performance.

4.5. Text Generation

We used OpenAI GPT-4.1-mini⁶ to generate reply messages by providing a conversation thread and requesting a suitable next response.

We implemented a characteristic Bluesky persona with a system prompt (“You are a politically liberal human social media user.”), since the site is used predominantly by liberal users; in the user prompt we described contextual information and dataset constraints (300 character limit, possibility of anonymized data in messages). See Appendix A.1 for the whole prompt.

5. Results

Our prediction pipeline routed 0.96% of samples using the simple rule-based approach on number

⁶accessed: 28 August 2025

of messages, 22.09% of samples using the lookup database approach, 70.48% of samples were classified as LIKE or FOLLOW directly by the sample’s cluster-specific LightGBM model, with the remaining 6.47% classified as OTHER by the LightGBM model and thus being sent to the custom trained rare action classifier. In total 6.60% of all predictions were REPLY actions, that were therefore subsequently sent to the LLM stage to generate a plausible message text.

	Min	Max	Avg
Majority Class Macro	0.23	0.31	0.28
Majority Class Weighted	0.36	0.78	0.59
RoBERTA-base Macro	-	-	0.48
RoBERTA-base Weighted	-	-	0.75
RoBERTA-Twitter Macro	-	-	0.48
RoBERTA-Twitter Weighted	-	-	0.75
LightGBM Macro	0.52	0.75	0.64
LightGBM Weighted	0.81	0.88	0.83
LightGBM Follow	0.79	0.92	0.88
LightGBM Like	0.57	0.86	0.73
LightGBM Other	0.00	0.58	0.30

Table 5: F1 scores for trained models on the simplified 3-label prediction task (FOLLOW, LIKE, OTHER). Baseline: majority class within each of the 25 clusters. Results include RoBERTa-base and RoBERTa-base-Twitter finetuned on 100,000 samples, and our LightGBM models per cluster, with min, max, and average scores across clusters.

5.1. Lookup Database

We recall here that we had selected as hyperparameters the requirement that a message appear with at least 10 votes in this database, and set the winner percentage to be over 85% (or 90% if using the Global Fallback voting strategy). As a result, predictions made with this route had an accuracy of 85-90% respectively. These highly accurate predictions were mostly for the most common actions (LIKE, FOLLOW, but also some REPLY and REPOST) due to the fact that the minimum number of 10 votes was difficult to achieve for the rare actions. This lookup route therefore contributed mainly to improving overall macro-F1 performance via the most frequently represented action categories rather the rare actions.

5.2. LightGBM Models

The performance for the 25 trained LightGBM models on training data is summarized in Table 5.

Remarkably we see that, compared to using finetuned language models, our manual text (and temporal) feature engineering approach with simpler LightGBM models outperformed by 16 macro-F1

points and 8 weighted-F1 points. Examining the variability in class-level LightGBM performance, we see that the FOLLOW action obtained consistently good performance, the LIKE action obtained decent performance in general, and that the OTHER action was highly variable (the single model which produced an F1 score of 0.00 was the model corresponding to cluster 9 which had only 15,891 samples and very few rare actions in total).

We analyzed the feature importance of all the trained LightGBM models; we share in Table 6 the most important features for our model trained on persona cluster 0.

We noted similar feature importance trends across all 25 models: the top 10 features we always dominated by the main timestamp features, followed by several general text features such as character count or word count, and then topic-specific feature counts such as `political_canadian_politics_count`.

5.3. Rare Action Classifier

The results for the rare action classifier are shown in Table 7.

Consistent with the LightGBM feature importance results in Table 6, we found that incorporating temporal features into the rare action classifier also significantly improved performance.

The most difficult classes were, predictably, those with very low support: UNREPOST, UNBLOCK, POST_DELETE — we consistently obtained F1 scores of 0 for these three actions.

5.4. Text Generation

Due to the large dataset size, we randomly sampled 1,000 conversations across the 25 clusters where the user action was REPLY, and sent these messages to GPT-4.1-mini to test the ability to generate a plausible next message in the conversation. After our prompt engineering approaches, we achieved an average cosine similarity of 0.83 across the 1,000 conversations (min 0.70, max 0.95).

6. Discussion

Our results using lookup strategies alongside persona- and topic-based keywords contributed significantly to the overall classification performance of our competition submission. However, given the importance of the temporal features in improving the trained model predictions, we discuss a methodological point about the dataset and social media prediction in general.

Indeed, presumably, by far the most common reaction by most users to most social media messages is to take no action whatsoever, i.e. to just

Feature	Importance	Feature	Importance
second_relative_integer_time	1049	political_government_count	96
first_relative_integer_time	793	question_count	95
avg_word_length	737	is_same_user	93
uppercase_ratio	633	tag_url_count	84
char_count	624	trump_specific_total_count	84
time_diff	615	political_candidates_count	81
word_count	330	political_election_2024_count	78
political_canadian_politics_count	268	social_citation_patterns_count	69
political_total_count	174	political_political_parties_count	68
digit_count	173	social_keywords_total	64
sentence_count	144	trump_specific_trump_names_count	59
exclamation_count	143	profanity_intensity_total_count	57
hashtag_count	117	social_engagement_count	48

Table 6: Top 26 feature importance scores for final LightGBM model: here we show the specific values for the Cluster 0 model, other models show similar patterns. Time features are highlighted in blue, and topic-specific keyword features are in red.

	F1
Majority Class Macro	0.07
Majority Class Weighted	0.34
Text-only Model Macro	0.42
Text-only Model Weighted	0.41
Text + Temporal Model Macro	0.56
Text + Temporal Model Weighted	0.87

Table 7: Trained Rare Action Classifier performance. Baselines: majority class guess (UNFOLLOW, 50.4% of 10 rare actions) and results for finetuning only the text component, i.e. ablating temporal features.

scroll past after briefly reading, but by the design of the dataset all messages were known to contain a user action. Therefore the dataset samples are a very specific subset of the entire platform and its user base—they are messages which are known to have caused at least some users to want to interact with them in some way.

In concrete terms, since the timestamp data of the messages was available and since the most common actions are LIKE and FOLLOW, it is reasonable to expect that if a user interacts with a message in under 30 seconds (for example) then their action will be a LIKE, because the very fact that they were able to respond in such a short time suggests that they are likely to be already following the message author (otherwise they would not have seen the message).

In practice, outside of the competition setting where this historical timestamp data would not be available, we believe that this notion of interaction temporal dynamics could be studied with indirect methods: for example taking into account whether or not a user already follows another, or whether they have recently refreshed their front page, could

all be proxies for how quickly a user interacts with a message. We leave this as a direction for future work.

Finally, we studied user behavior in the context of Bluesky by using the persona cluster of the responding user but it would be interesting to extend the persona labeling and annotation to the user creating the original message also. This information was not available in the existing version of BLUEPRINT, but it would allow further research into inter-persona interaction dynamics.

7. Conclusion

We presented a comprehensive hybrid approach to social media user behavior prediction that addresses the challenge of predicting both common and rare user actions on social platforms. Our methodology demonstrates that different types of user behaviors and personas require specialized modeling approaches: lookup-based prediction for historically consistent patterns (achieving 85-90% accuracy for 22% of samples), persona-specific tabular models for common actions (outperforming transformer models by 16 macro-F1 points), and hybrid neural architectures for rare action classification (achieving 0.56 macro-F1). The performance on extremely rare actions suggests challenges in predicting very low-frequency behaviors that may require alternative approaches such as synthetic data augmentation.

8. Ethics Statement

The dataset used in this paper has been made public for the SocialSim: Social-Media Based Personas shared task. The dataset creators used publicly available Bluesky data collected in compliance

with the platform’s Terms of Service (Bück-Kaeffer et al., 2025). To protect user privacy, data was anonymised by removing identifiable information, using relative timestamps, pseudonymizing usernames, and analyzing aggregated behaviors rather than individual actions. Users can request data removal at any time. The dataset will be shared on HuggingFace under a license restricting use to research purposes and prohibiting unethical applications. Only anonymized records are provided, supporting research on NLP techniques in social media contexts while acknowledging potential dual-use risks and emphasizing responsible AI governance.

We note that generative language models were used solely for text refinement in preparing this manuscript.

9. Acknowledgments

We thank the organizers of the Social Simulation with LLMs workshop at COLM 2025 for designing this very interesting challenge, and for their work in curating and publishing the Bluesky dataset.

10. Bibliographical References

- Dimosthenis Antypas, Asahi Ushio, Francesco Barbieri, Leonardo Neves, Kiamehr Rezaee, Luis Espinosa-Anke, Jiaxin Pei, and Jose Camacho-Collados. 2023. [SuperTweetEval: A Challenging, Unified and Heterogeneous Benchmark for Social Media NLP Research](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12590–12607, Singapore. Association for Computational Linguistics.
- Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.
- Aurélien Bück-Kaeffer, Je Qin Chooi, Dan Zhao, Maximilian Puelma Touzel, Kellin Pelrine, Jean-François Godbout, Reihaneh Rabbany, and Zachary Yang. 2025. [BluePrint: A Social Media User Dataset for LLM Persona Evaluation and Training](#).
- Syeda Nadia Firdaus, Chen Ding, and Alireza Sadeghian. 2021. [Retweet Prediction based on Topic, Emotion and Personality](#). *Online Social Networks and Media*, 25:100165.
- Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: a highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 3149–3157.
- Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. 2025. [LLM Generated Persona is a Promise with a Catch](#).
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.
- Ana Meštrović, Milan Petrović, and Slobodan Beliga. 2022. [Retweet Prediction Based on Heterogeneous Data Sources: The Combination of Text and Multilayer Network Features](#). *Applied Sciences*, 12(21).
- Huan-Kai Peng, Jiang Zhu, Dongzhen Piao, Rong Yan, and Ying Zhang. 2011. [Retweet Modeling Using Conditional Random Fields](#). In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 336–343.
- Zhongyi Qiu, Hanjia Lyu, Wei Xiong, and Jiebo Luo. 2025. [Can LLMs Simulate Social Media Engagement? A Study on Action-guided Response Generation](#).
- Giulio Rossetti, Massimo Stella, Rémy Cazabet, Katherine Abramski, Erica Cau, Salvatore Citraro, Andrea Failla, Riccardo Improta, Virginia Morini, and Valentina Pansanella. 2024. [Y Social: an LLM-powered Social Media Digital Twin](#). *CoRR*, abs/2408.00818.
- Konstantinos Sotiropoulos, John W Byers, Polyvios Pratikakis, and Charalampos E Tsourakakis. 2019. Twittermancer: predicting user interactions on twitter. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 973–980. IEEE.
- Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi. 2010. [Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network](#). In *Proceedings of the 2010 IEEE Second International Conference on Social Computing, SOCIALCOM ’10*, page 177–184, USA. IEEE Computer Society.
- Chenkai Sun, Jinning Li, Hou Pong Chan, ChengXiang Zhai, and Heng Ji. 2023. [Measuring the effect of influential messages on varying personas](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 554–562, Toronto,

Canada. Association for Computational Linguistics.

Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. 2023. [Simulating Social Media Using Large Language Models to Evaluate Alternative News Feed Algorithms](#). *CoRR*, abs/2310.05984.

Maximilian Puelma Touzel, Sneheel Sarangi, Austin Welch, Gayatri Krishnakumar, Dan Zhao, Zachary Yang, Hao Yu, Ethan Kosak-Hine, Tom Gibbs, Andreea Musulan, Camille Thibault, Busra Tugce Gurbuz, Reihaneh Rabbany, Jean-François Godbout, and Kellin Pelrine. 2024. [A Simulation System Towards Solving Societal-scale Manipulation](#). *CoRR*, abs/2410.13915.

Haozhe Wu, Zhiyuan Hu, Jia Jia, Yaohua Bu, Xiangan He, and Tat-Seng Chua. 2020. Mining unfollow behavior in large-scale online social networks via spatial-temporal interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 254–261.

Zi Yang, Jingyi Guo, Keke Cai, Jie Tang, Juan-Zi Li, Li Zhang, and Zhong Su. 2010. [Understanding retweeting behaviors in social networks](#). *Proceedings of the 19th ACM international conference on Information and knowledge management*.

Qi Zhang, Yeyun Gong, Jindou Wu, Haoran Huang, and Xuanjing Huang. 2016. [Retweet Prediction with Attention-based Deep Neural Network](#). In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, pages 75–84. ACM.

Junkai Zhou, Liang Pang, Ya Jing, Jia Gu, Huawei Shen, and Xueqi Cheng. 2024. [Knowledge Boundary and Persona Dynamic Shape A Better Social Media Agent](#).

A. Appendix

A.1. Prompts

GPT-4.1-mini system prompt

You are a politically liberal human social media user.

GPT-4.1-mini user prompt

Below is a series of one or more messages from the social network BlueSky, which is a more liberal variant of Twitter or X.

Read the entire conversation and then generate what you think is a suitable next reply message. There is a 300 character limit, so don't write long paragraphs.

The majority of users are politically liberal, from the USA or Canada or Western Europe, and the conversation took place in 2024.

If you see any words such as @<USER-NAME> it is because the data has been anonymized.

In rare cases the conversation might consist of a single user posting a multi-message thread, so if so you should try to continue in their style.

Conversation to reply to
{conversation}

The Data Acquisition Framework: Bridging Psychometrics and NLP for Personality Dataset Construction

Lorenz Dumanski, Michael Spranger, Melanie Siegel

University of Applied Sciences Mittweida, University of Applied Sciences Darmstadt
Technikumplatz 17 09648 Mittweida, Schöfferstr. 3 64295 Darmstadt
{dumanski, spranger}@hs-mittweida.de, melanie.siegel@h-da.de

Abstract

Existing datasets for personality recognition in Natural Language Processing (NLP) suffer from documented quality problems: self-reported labels lacking psychometric validation, limited domain diversity and lack of context. Despite these known limitations, state-of-the-art approaches continue relying on the same datasets due to absence of alternatives. We present the Data Acquisition Framework (DAF), which addresses this gap by systematically translating psychometric questionnaire items into controlled communication scenarios through expert-community validation. DAF-items, validated scenario descriptions with contextual parameters, are deployed via the Automatic Data Acquisition and Annotation Tool (ADAAT). Participants complete personality surveys and engage in scenario-based text interactions with LLM personas configured to the DAF-Item context. This yields communication data with direct, item-level psychometric annotations.

Keywords: personality recognition, psychometric validation, dataset construction, scenario-based assessment, item-level annotation

1. Introduction

Personality recognition has become increasingly relevant for human-centered natural language processing (NLP) applications. From adaptive dialogue systems to mental health screening, understanding individual differences in personality enables more effective and personalized human-computer interaction (Sirasapalli and Malla, 2023; Majumder et al., 2017; Mohammad and Turney, 2013).

Existing personality recognition approaches predominantly rely on data from social media platforms or essay corpora (Mushtaq and Kumar, 2023; Singh and Singh, 2024; Kelvin and Utomo, 2024). Such data presents critical limitations: personality labels are often self-reported without standardization, and the data is domain-specific, limiting generalizability across communication contexts (Section 2).

This lack of psychometric rigor manifests in several ways: inconsistent use of personality assessments across studies, weak coupling between test administration and data collection, and limited attention to fundamental psychometric properties such as construct validity and test-retest reliability (Stachl et al., 2020). Consequently, the labels used to train personality recognition models may themselves be unreliable.

Despite over a century of methodological development in psychometrics, from classical test theory to modern item response theory, personality recognition in NLP has largely evolved independently from these established practices. While

psychology employs rigorous, theory-driven approaches to personality assessment, NLP research has favored opportunistic data collection driven by availability rather than psychometric validity (Stachl et al., 2020; Naz et al., 2025; Yang et al., 2021). This disconnect represents a significant gap: there exists no systematic framework for controlled, psychometrically grounded data acquisition in personality NLP.

We present the Data Acquisition Framework (DAF), a methodological approach that bridges psychometric test construction and NLP-oriented data collection. DAF applies rational-empirical validation principles to translate psychometric items into communication scenarios (DAF-items). The Automatic Data Acquisition and Annotation Tool (ADAAT) deploys these scenarios through text-based interactions with Large Language Model (LLM) personas, enabling scenario-based data collection at scale. Participants complete standardized personality assessments, then engage with scenario-driven chat interactions, yielding communication data with direct, item-level annotations.

2. Background: Core Datasets in Personality Recognition

Personality recognition research has relied on a limited set of datasets documented across multiple surveys (Mushtaq and Kumar, 2023; Naz et al., 2025; Singh and Singh, 2024):

The *MyPersonality* dataset (Stillwell and Kosin-

ski, 2015) comprised, at its peak, Facebook¹ data from over 6 million users with Big-Five (Goldberg, 1992) personality labels (Kosinski et al., 2015). Following its withdrawal in 2018 due to ethical concerns, a reduced sample of approximately 9,900 status updates from 250 users remains publicly available², though the authors discourage its use.

The *Stream-of-Consciousness* (SoC) dataset (Pennebaker and King, 1999) contains 2,468 student essays with Big Five annotations using dichotomous high/low classification.

The *Kaggle MBTI* dataset (Mitchell, 2018) comprises approximately 8,600 forum posts from PersonalityCafe³, labeled with self-reported Myers-Briggs Type Indicator (MBTI) (Myers, 1998) categories.

The *PANDORA* dataset (Gjurković et al., 2021) is a multilingual corpus of Reddit⁴ posts annotated with Big Five and MBTI traits across multiple languages.

These datasets represent popular data infrastructure employed in personality recognition research over the past decade.

2.1. Data Quality Concerns

Multiple limitations of these datasets have been documented across recent literature:

Dichotomous Classification. The SoC dataset employs binary high/low classification for Big Five dimensions, which are inherently continuous scales (Saeteros et al., 2025). This artificial dichotomization renders the dataset potentially unsuitable for trait-level classification, as individuals near decision boundaries are treated identically to those at distribution extremes (Stachl et al., 2020). Such discretization fundamentally misrepresents the dimensional nature of Big Five personality constructs.

Weak Psychometric Grounding. The MBTI itself is critiqued within personality psychology (Gjurković et al., 2021; Goldberg, 1992) and lacks the empirical foundation of the Big Five (Pittenger, 2005). Beyond framework concerns, datasets lack measurement and reporting standardization. While the Big Five can be measured using instruments ranging from brief scales (BFI-10 (Rammstedt et al., 2014)) to comprehensive inventories (NEO-PI-R (Costa and McCrae, 1992)), researchers often cannot determine which was used or whether administration followed validated protocols (Gjurković et al., 2021). This undermines construct validity and cross-study comparability, par-

ticularly in *PANDORA* and *Kaggle MBTI* datasets (Saeteros et al., 2025; Yang et al., 2021; Gjurković et al., 2021).

Domain Specificity and Lack of Contextual Control. All major datasets derive from specific domain contexts (Facebook status updates, forum posts, academic essays), limiting cross-domain generalization (Singh and Singh, 2024). Additionally, social media texts lack situational metadata. Researchers cannot determine the communicative context (Tang et al., 2025) (e.g., responding to conflict, casual conversation) that may modulate trait expression (Mischel, 1973; Zayas et al., 2008).

Class Imbalance and Size. Kelvin and Utomo (2024) document severe class imbalance in the MBTI dataset, while Singh and Singh (2024) highlight noise, limited size, and language constraints across available datasets.

The withdrawal of MyPersonality has exacerbated data scarcity issues.

Table 1 summarizes core datasets and their limitations.

2.2. Continued Reliance on Problematic Data

Despite documented limitations, recent work towards text-based personality recognition continues relying on these datasets (Tang et al., 2025; Saeteros et al., 2025; Li et al., 2025; Fatahian and Ravanmehr, 2025; Kelvin and Utomo, 2024). This persistent reliance reflects not researcher oversight but the absence of alternatives, as documented by Naz et al. (2025) across virtually all contemporary approaches.

Research is focused on Big Five (Goldberg, 1992) and MBTI (Myers, 1998) personality models. Although more than these two approaches to personality exist, the lack of datasets employing alternative frameworks has implicitly favoured NLP personality recognition to these two models.

2.3. Calls for New Data Infrastructure

The limitations above have prompted explicit calls for improved data resources. Mushtaq and Kumar (2023) identify critical needs emerging from their systematic review:

- **Large-scale shared datasets** with psychometrically validated labels
- **Contextually rich data** capturing situational variability in personality expression
- **Communication-focused corpora** beyond social media monologues, including interactive dialogue
- **Cross-domain diversity** enabling model generalization testing

¹<https://www.facebook.com>

²<https://github.com/nlp-psych/personality>

³<https://www.personalitycafe.com>

⁴<https://www.reddit.com>

Dataset	Source	Labels	Key Limitations
<i>MyPersonality</i> (Stillwell and Kosinski, 2015)	Facebook posts	Big Five	Withdrawn 2018; domain specific (Singh and Singh, 2024)
<i>SoC Essays</i> (Pennebaker and King, 1999)	2,468 student essays	Big Five (binary)	Binary classification misrepresents continuous trait dimensions (Saeteros et al., 2025; Stachl et al., 2020); domain specific
<i>Kaggle MBTI</i> (Mitchell, 2018)	8,600 forum posts	MBTI (self-report)	Self-reported without validation; MBTI itself is in critique (Pitenger, 2005; Goldberg, 1992); class imbalance (Kelvin and Utomo, 2024)
<i>PANDORA</i> (Gjurković et al., 2021)	Reddit	Big Five, MBTI	Self-reported labels + inferred labels; platform-specific communication; lacks situational metadata (Gjurković et al., 2021)

Table 1: Core personality recognition datasets and documented limitations

Singh and Singh (2024) explicitly state that "limited data availability and few shared datasets" represent primary obstacles to advancing personality recognition research. Saeteros et al. (2025) emphasize the need for interdisciplinary collaboration to avoid applying NLP methods to psychologically inadequate data.

These converging calls underscore a fundamental gap: while NLP has advanced methodologically (transformer architectures, large language models, multi-task learning), the field remains constrained by data infrastructure. Our work addresses this gap through a systematic framework for generating psychometrically grounded communication data.

3. The Data Acquisition Framework

Personality recognition in NLP suffers from two critical data challenges: scarcity of psychometrically annotated communication data and limited domain diversity in existing datasets (see Section 2.1). We present DAF, a methodological approach that addresses both challenges through controlled, theory-driven data generation directly coupled with standardized personality assessments.

3.1. DAF-Item Development

The DAF method translates psychometric questionnaire items into communication scenarios through iterative expert-community validation. Big Five personality measurement instruments like the NEO-PI-R (Costa and McCrae, 1992) or the IPIP-NEO-120 (Johnson, 2014) consist of brief self-statements rated on Likert scales. These items

provide no concrete information about setting, social context, or circumstances. Domain experts with experience in personality assessment or psychometric test construction draft candidate scenarios that operationalize these abstract items into concrete communicative situations.

For instance, the IPIP-NEO-120 item "I feel comfortable around people" might be transformed into a scenario such as: "A friend invites you to their place" accompanied by contextual specifications like "It's his/her birthday", "20–30 expected guests", "party at home/club/bar" or "whether you know other attendees". These contextual elements aim to provide necessary framing for consistent scenario deployment while enabling investigation of how situational factors modulate trait expression in textual communication.

We propose both experts and non-expert (community) validators evaluating scenarios against three criteria:

1. **Construct validity:** Does the scenario elicit behavior reflecting the source item?
2. **Comprehensibility:** Is the scenario description clear and unambiguous?
3. **Realism:** Is this a realistic, familiar situation? Especially regarding everyday online chat communication?

While experts may identify theoretically relevant contextual parameters based on psychological theory, we propose these require validation through community input. Additionally, community validators would assess two parameter aspects:

1. **Behavioral impact:** "Would this parameter change how you communicate in this situation?" and

Component	Example Content
Source Item	"I feel comfortable around people." Participant response: 5 (Strongly Agree).
DAF Scenario	A friend invites you to their place.
DAF Scenario Parameters	Birthday, Party, 20-30 Guests, Participant only knows host
Communication Sample	Friend (LLM): "Hey, next week is my birthday and I would be very happy if you came to my party" Participant: "Wow, that sounds very nice! Thank you for the invitation. Who is coming besides me?" [Conversation continues.]

Table 2: DAF-Item description example of the IPIP-NEO-120 (Johnson, 2014) item "I feel comfortable around people".

2. **Realistic variation:** "How commonly do you encounter different values of this parameter?"

Parameters demonstrating high behavioral impact and real-world variation would advance to final DAF-item specifications. Scenarios achieving consensus across both groups would become validated DAF-items.

This iterative expert-community validation parallels established approaches in psychometric test development (Thomas et al., 1992; DeVellis and Thorpe, 2021) to ensure both construct validity and practical applicability.

Table 2 illustrates how the DAF-ADAAT pipeline will transform a psychometric item into annotated communication data. This example represents a use case; actual implementations will be refined based on expert input, pilot testing, and empirical validation.

3.2. ADAAT: Automated Data Collection Platform

ADAAT is designed to operationalize DAF-items through a controlled data collection platform. The ADAAT workflow consists of three phases:

1. **Psychometric Assessment:** Participants complete a standardized personality inventory (e.g., NEO-PI-R (Costa and McCrae, 1992)), providing validated labels for each questionnaire item. These responses are stored for subsequent annotation.
2. **Scenario-Based Communication:** Participants engage with DAF-item scenarios via text-based chat interaction. An LLM persona, configured to match the scenario context, serves as conversational partner and gradually introduces and maintains contextual parameters from the DAF-item into the conversation. This operationalizes situational conditions under which trait-relevant behavior can be observed (Mischel, 1973).
3. **Annotation:** Communication data generated in Phase 2 is automatically annotated with the

participant's item-level responses from Phase 1, creating direct coupling between psychometric labels and observable communication behavior. Each conversational interaction is linked to the source item, participant response value, and scenario context.

This design aims to yield communication data with psychometrically grounded annotations, addressing the weak labeling problem prevalent in existing personality NLP datasets (see Section 2.1).

LLMs enable scalable deployment without requiring human confederates for each scenario, though persona consistency requires careful validation (see Section 5).

While designed for personality assessment, DAF generalizes to questionnaire-based constructs that manifest behaviorally in communication. Potential applications include constructs where self-reports may be linked to observable interaction patterns, like emotion regulation, social anxiety, or attachment styles.

4. Conclusion

DAF addresses the lack of psychological rigor in current state-of-the-art datasets through its methodological approach to item development. Coupled with ADAAT and its LLM-persona-based scenario initialization, it provides a cost-effective alternative to manual labeling, which remains prohibitively expensive for most research contexts (Mushtaq and Kumar, 2023). Since DAF-items are context-enriched scenario descriptions directly annotated with questionnaire responses, they may enable novel approaches to personality classification. For instance, topic modeling strategies could identify communication fragments that yield information regarding response tendencies for specific items. Yang et al. (2021) demonstrated that answering personality questionnaire items based on behavioral cues in social media posts can be a valid strategy. A DAF-ADAAT dataset could advance interdisciplinary research across data science, NLP, and psychology. Such datasets may

enable improvements in both classification accuracy and interpretability of personality expression in communication.

5. Limitations

Several design questions require empirical investigation: which personality inventory to use as foundation (e.g., NEO-PI-R vs. IPIP-NEO-120), the required number of expert and community validators, essential contextual parameters, formalization approaches, optimal scenario granularity, and evaluation thresholds. Additionally, mapping granularity, whether scenarios target individual items or personality facets remains open. Item-level mapping maximizes precision but requires numerous scenarios; facet-level approaches reduce scenario count but need validation. Expert consultation and pilot studies will address these questions.

Constructing LLM personas can be challenging and come with several caveats. To ensure the persona does reflect the scenario context and stays in character, their implementation will be rather complex. This is addressed in Section A.2.

ADAAT is currently in the conceptual design phase. Key technical questions, including optimal LLM persona initialization strategies, appropriate interaction duration, conversational turn structure, and user interface design, will be resolved through iterative prototyping and pilot testing (see Section A). The deployment of the first prototype is planned in the near future.

Furthermore, the use of human-LLM interactions and fictitious scenarios may limit external validity, particularly if the aim is to eventually infer personality from naturalistic human-human dialogue. Addressing this gap represents an important direction for future work.

6. Ethical Considerations

Participants will provide informed consent for personality assessment and LLM interaction, with clear disclosure of artificial conversational partners. ADAAT employs pseudonymization without collecting personally identifiable information; participants retain withdrawal rights. Data handling complies with institutional ethics approvals and data protection regulations.

Personality recognition systems risk discriminatory application in high-stakes contexts. Deployment should follow responsible AI frameworks (UNICRI and INTERPOL, 2026), emphasizing transparency, human oversight, and safeguards against misuse.

7. Acknowledgements

This project is co-funded by the European Union and the tax revenues on the basis of the budget adopted by the Saxon State Parliament.



Co-funded by
the European Union



This project is co-financed from tax revenues on the basis of the budget adopted by the Saxon State Parliament.

8. Bibliographical References

- Paul T. Costa and Robert R. McCrae. 1992. *Revised NEO Personality Inventory (NEO PI-R) and NEP Five-Factor Inventory (NEO-FFI) : Professional Manual*. NEO PI-R. Psychological Assessment Resources, Odessa, Fla. (P.O. Box 998, Odessa 33556).
- Lee J. Cronbach. 1951. *Coefficient Alpha and the Internal Structure of Tests*. *Psychometrika*, 16(3):297–334.
- Paul G. Curran. 2016. *Methods for the detection of carelessly invalid responses in survey data*. *Journal of Experimental Social Psychology*, 66:4–19.
- Robert F DeVellis and Carolyn T Thorpe. 2021. *Scale Development: Theory and Applications*. Sage publications. ISBN: 978-1-5443-7934-0.
- Mohammad Fatahian and Reza Ravanmehr. 2025. *Personality Recognition in Social Media using Sentence Embeddings Based on Transformer Networks*. *SN Computer Science*, 6(7):797.
- Matej Gjurković, Vanja Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Snajder. 2021. *PANDORA Talks: Personality and Demographics on Reddit*. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 138–152, Online. Association for Computational Linguistics.
- Lewis R. Goldberg. 1992. *The development of markers for the Big-Five factor structure*. *Psychological Assessment*, 4(1):26–42.

- John A. Johnson. 2014. [Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120](#). *Journal of Research in Personality*, 51:78–89.
- Kelvin Kelvin and Yesun Utomo. 2024. [Overview of Text Based Personality Prediction Using Deep Learning](#). *Engineering, MATHematics and Computer Science Journal (EMACS)*, 6(2):93–100.
- Michal Kosinski, Sandra C. Matz, Samuel D. Gosling, Vesselin Popov, and David Stillwell. 2015. [Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines](#). *American Psychologist*, 70(6):543–556.
- Zheng Li, Sujian Li, Dawei Zhu, Qilong Ma, and Weimin Xiong. 2025. [EERPDP: Leveraging Emotion and Emotion Regulation for Improving Personality Detection](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7721–7734, Abu Dhabi, UAE. Association for Computational Linguistics.
- Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. 2017. [Deep Learning-Based Document Modeling for Personality Detection from Text](#). *IEEE Intelligent Systems*, 32(2):74–79.
- Roderick P. McDonald. 1999. *Test Theory: A Unified Treatment*. Test Theory: A Unified Treatment. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US. ISBN: 978-0-8058-3075-0.
- Walter Mischel. 1973. [Toward a cognitive social learning reconceptualization of personality](#). *Psychological Review*, 80(4):252–283.
- J Mitchell. 2018. [\(MBTI\) Myers-Briggs Personality Type Dataset](#). Accessed: 2026-02-12.
- Saif M. Mohammad and Peter D. Turney. 2013. [Crowdsourcing a Word–Emotion Association Lexicon](#). *Computational Intelligence*, 29(3):436–465.
- Sumiya Mushtaq and Neerendra Kumar. 2023. [Text-Based Automatic Personality Recognition: Recent Developments](#). In *Proceedings of Third International Conference on Computing, Communications, and Cyber-Security*, pages 537–549, Singapore. Springer Nature. ISBN: 978-981-19-1142-2.
- Isabel Briggs Myers. 1998. *MBTI Manual : A Guide to the Development and Use of the Myers-Briggs Type Indicator*. Palo Alto, Calif. : Consulting Psychologists Press. ISBN: 978-0-89106-130-4.
- Anam Naz, Hikmat Ullah Khan, Amal Bukhari, Bader Alshemaimri, Ali Daud, and Muhammad Ramzan. 2025. [Machine and deep learning for personality traits detection: A comprehensive survey and open research challenges](#). *Artificial Intelligence Review*, 58(8):239.
- James W. Pennebaker and Laura A. King. 1999. [Linguistic styles: Language use as an individual difference](#). *Journal of Personality and Social Psychology*, 77(6):1296–1312.
- David J. Pittenger. 2005. [Cautionary comments regarding the Myers-Briggs Type Indicator](#). *Consulting Psychology Journal: Practice and Research*, 57(3):210–221.
- B. Rammstedt, C. J. Kemper, M. C. Klein, C. Beierlein, and A. Kovaleva. 2014. [Big Five Inventory \(BFI-10\)](#). *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*.
- David Saeteros, David Gallardo-Pujol, and Daniel Ortiz-Martínez. 2025. [Text speaks louder: Insights into personality from natural language processing](#). *PLOS ONE*, 20(6):e0323096.
- Simarpreet Singh and Williamjeet Singh. 2024. [AI-based personality prediction for human well-being from text data: A systematic review](#). *Multimedia Tools and Applications*, 83(15):46325–46368.
- Joshua Johnson Sirasapalli and Ramakrishna Murty Malla. 2023. [A deep learning approach to text-based personality prediction using multiple data sources mapping](#). *Neural Computing and Applications*, 35(28):20619–20630.
- Clemens Stachl, Quay Au, Ramona Schoedel, Samuel D. Gosling, Gabriella M. Harari, Daniel Buschek, Sarah Theres Völkel, Tobias Schuerk, Michelle Oldemeier, Theresa Ullmann, Heinrich Hussmann, Bernd Bischl, and Markus Bühner. 2020. [Predicting personality from patterns of behavior collected with smartphones](#). *Proceedings of the National Academy of Sciences*, 117(30):17680–17687.
- DJ Stillwell and M Kosinski. 2015. [myPersonality project website](#). Accessed: 2026-02-16.
- Qirui Tang, Wenkang Jiang, Xinlong Pan, Lei Lin, Jizhao Zhu, Yihua Du, and Donghong Sun. 2025. [Using Psycholinguistic Clues to Index Deep Semantic Evidences: Personality Detection in Social Media Texts](#). *Chinese Journal of Information Fusion*, 2(2):112–126.

Suzanne D. Thomas, Donna K. Hathaway, and Kristopher L. Arheart. 1992. [Face Validity](#). *Western Journal of Nursing Research*, 14(1):109–112.

UNICRI and INTERPOL. 2026. [An overview of the AI Toolkit | AI Toolkit](#). Accessed: 2026-02-19.

Feifan Yang, Tao Yang, Xiaojun Quan, and Qinliang Su. 2021. [Learning to Answer Psychological Questionnaire for Personality Detection](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1131–1142, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vivian Zayas, Donna Whittsett, Jenna Lee, Nicole Wilson, and Yuichi Shoda. 2008. [From Situation Assessment to Personality: Building a Social-Cognitive Model of a Person](#). In G. J. Boyle, Gerald Matthews, and D. H. Saklofske, editors, *The SAGE Handbook of Personality Theory and Assessment: Volume 2 — Personality Measurement and Testing*, pages 377–401. SAGE Publications Ltd.

A. Pilot Study Design for ADAAT Validation

The following outlines planned pilot study procedures to address key empirical uncertainties in the ADAAT deployment. These studies are designed to systematically validate the technical and methodological components of the framework prior to full-scale data collection.

A.1. LLM Selection

LLM selection represents a critical design decision for ADAAT, as response quality and latency directly affect both data quality and participant experience. Candidate models must be evaluated against two practical constraints.

Hardware and Deployment Resources. Research facilities typically operate under hardware limitations that preclude deployment of the largest available models. Local deployment via frameworks such as Ollama offers cost-effective alternatives to cloud-based inference, though at the expense of model scale. Cloud deployment, while enabling access to state-of-the-art models, introduces per-participant costs that may be prohibitive at scale and raise data privacy concerns. Response latency must remain within acceptable bounds to maintain a realistic communication experience; excessively delayed responses disrupt conversational flow and may systematically affect participant behavior.

Response Quality Evaluation. While response latency can be assessed automatically, response quality requires human evaluation. We propose a benchmark procedure in which candidate LLMs are instantiated with a set of pilot DAF-item scenarios. Participants engage in scenario-based chat interactions and subsequently complete a post-chat survey assessing two criteria: (1) the naturalness of the interaction, and (2) whether the LLM appropriately introduced and maintained the DAF-item’s contextual parameters throughout the conversation. Aggregated survey responses serve as a benchmark across candidate models, informing final model selection.

A.2. LLM Persona Consistency

If the LLM drifts off-character or introduces behavior inconsistent with the scenario context, the resulting communication data cannot be reliably linked to the intended situational conditions. Persona behavior is sensitive to both model choice and prompt design, and must therefore be validated empirically. At this early stage, the following steps for persona/prompt evaluation are planned:

1. Design a complex pilot scenario and iteratively refine the system prompt, identifying prompt features that reliably elicit on-character behavior.
2. Generalize validated prompt-engineering features into a reusable template that can be instantiated with arbitrary DAF-item specifications.
3. Following each pilot interaction, participants complete a post-chat survey item indicating whether the conversational partner remained consistent with the scenario context.
4. Interactions flagged as off-character are investigated and utilized for prompt/persona redesign.

Since ADAAT’s development is work in progress, best practices for prompt and persona engineering will likely advance and thus have to be monitored alongside ADAAT’s development.

A.3. Questionnaire Data Validation

Self-report personality inventories are subject to response validity concerns, including careless or inattentive responding (Curran, 2016). To address this, ADAAT will incorporate **Instructed Response Items** into the survey flow: items that explicitly direct the participant to select a specific response option (e.g., “For this item, please select ‘Strongly Agree.’”) (Curran, 2016). In addition

ADAAT will record response times per Item as suggested by Curran (2016).

Subsequently, scale reliability will be assessed post-collection for each Big Five domain and facet using Cronbach's α (Cronbach, 1951) or McDonald's ω (McDonald, 1999) to verify that the selected inventory performs as expected within the collected sample. Results will be reported transparently to inform the interpretation of downstream personality recognition results.

Language Ideologies in a Multilingual Society: An LLM-based Analysis of Luxembourgish News Comments

Emilia Milano¹, Alistair Plum¹, Yves Scherrer², Christoph Purschke¹

¹University of Luxembourg, Esch-sur-Alzette, Luxembourg,

²University of Oslo, Oslo, Norway

{emilia.milano, alistair.plum, christoph.purschke}@uni.lu

yves.scherrer@ifi.uio.no

Abstract

Detecting language ideologies is a valuable yet complex task for understanding how identities are constructed through discourse. In Luxembourg’s multicultural and multilingual society, language ideologies reflect more than simple preferences: they carry deep cultural and social meanings, shaping identities and social belonging. Following recent developments in applying Natural Language Processing tools to linguistics and social science, this paper explores the potential of large language models to assist in the detection of language ideologies. We manually annotate a corpus of user comments in Luxembourgish with predefined ideological categories and then evaluate the performance of large language models under varying prompt conditions to assess their ability to replicate these human annotations. Since Luxembourgish is a small language and poorly represented in the LLMs’ training data, we also investigate whether machine-translating the data to high-resource languages increases performance on the ideology detection task. Our findings suggest that, while LLMs are not yet fully optimized for a multi-class ideological annotation task, they are practical tools to identify language ideological content.

Keywords: Language Ideologies, Large Language Models, Computational Sociolinguistics

1. Introduction

Automatic detection of language ideologies is a promising but challenging task. Language ideologies, or shared beliefs about language, play a central role in reinforcing and establishing identities and power imbalances (Blommaert, 2010). Although culturally tied to the speakers’ community, similar language ideologies have been found in different contexts. Often, language ideologies are linked to matters of social belonging and discrimination (Grondelaers and van Hout, 2011; Lippi-Green, 2012). This is also found in our case study, Luxembourg. The country is characterized by a complex and dynamic societal multilingualism, in which the choice of language in interaction carries rich social meaning aside from practical requirements or personal preference (Horner and Weber, 2008). This is evidenced in public discourses on the state of multilingualism and the role of Luxembourgish, where language choices and ideological stances are often closely linked. For example, the phrase *En français, s’il vous plaît* (‘In French, please’), used by (French-speaking) service personal as a reaction to requests in Luxembourgish, has become a prominent figure in discourse (Purschke, 2025): It is interpreted (and quoted) by speakers of Luxembourgish as a sign of the unwillingness to integrate in society, linguistically and socially, and linked to questions of demographic development and language vitality (Purschke, 2023).

Hence, being able to automatically and systematically detect language ideologies in large amounts

of text would be beneficial for obtaining a deeper understanding of cultural dynamics and language-related discourse at scale.

Against this backdrop, the increasing use of Large Language Models (LLMs) for linguistic (Manning et al., 2020; Wu et al., 2025) and social science research (Ma et al., 2024) presents a promising opportunity to systematically investigate the automatic detection and classification of language ideologies.

For this study, we base ourselves on a corpus of written Luxembourgish comments provided by the national broadcaster RTL¹ (Ranasinghe et al., 2023). We annotate the comments at two levels: (1) whether they are related to language ideology or not, and (2) which different types of language ideologies they contain (see Section 2.2). We assess a range of open-weights and closed-weights LLMs on their ability to predict these annotations, testing several prompts in both zero-shot and few-shot settings.

Since Luxembourgish is a small language, we presume that it is comparatively poorly represented in the LLMs’ training data and that this results in low classification performance. We therefore follow the *translate-test* paradigm and investigate whether machine translating the Luxembourgish comments into a higher-resourced language (specifically, English, French or German) increases classification accuracy.

¹Our agreement with the data provider prevents us from publicly releasing the corpus. However, we can provide upon request the comment IDs so that the corpus can be recreated.

More specifically, our research questions are:

1. Is language ideology classification a suitable task for the current generation of LLMs?
2. Do the languages of the data and the quality of their translations have an impact on LLM performance on language ideology classification?
3. Can LLMs contribute to inform sociolinguistic research without requiring high technical expertise or task-specific model adaptation?

To explore the feasibility and challenges of the task, we provide LLMs with language ideology categories, descriptions and examples, and instructions to annotate accordingly. Our contributions offer an evaluation of LLMs for language ideologies identification, assessing both the impact of the source language and translations on the same task. In doing so, we provide an overview of how LLMs work with Luxembourgish on cultural tasks, compared to more established languages in NLP. We find that fine-grained language ideology detection is a complex task for LLMs, and that translations do not lead to substantial performance improvement. Integrating NLP tools into sociolinguistic research, we acknowledge limitations and possible improvements of the task setup.

2. Related work

In this section, we discuss work related to Luxembourgish NLP (Section 2.1), as well as language ideologies (Section 2.2). After that, we discuss work on ideologies and LLMs (Section 2.3), as well as cross-lingual transfer in LLMs (Section 2.4).

2.1. Luxembourgish

Luxembourgish is a West Germanic language with about 400.000 speakers (Gilles, 2019). It is one of three official languages in Luxembourg in addition to German and French, but it was only recognized as such in 1984. Historically, Luxembourgish mainly developed as an oral language, acquiring more written domains only recently with the rise of digital and social media. As a multilingual country with an extremely dynamic language regime, Luxembourg is a particularly interesting case for sociolinguistic studies (Purschke and Gilles, 2023). However, in NLP, Luxembourgish should be viewed as a small language, although various resources have been made available in recent years (Plum et al., 2025; Lutgen et al., 2025; Philippy et al., 2024; Plum et al., 2024).

2.2. Language ideologies

We follow Blume (2003)'s description of ideologies and understand them as descriptions of collective, depersonalized normative interpretations used by individuals and groups to justify and evaluate their own and others' actions. As such, ideologies towards and about language (Blommaert, 2010) play a central role in creating and ordering social meaning (Rhodes, 2023), in shaping language policies and perception, in structuring social relations and societal belonging (Woolard, 2020), and in contributing to individual and collective identity constructions (Kroskrity, 2022).

From a methodological perspective, (language) ideologies have also been studied with quantitative tools and methods. Baker and Levon (2015) compares qualitative and corpus-based approaches to the analysis of newspaper articles on masculinity, showing that the two approaches can complement each other, and Vessey (2017) investigates language ideologies in Canada through corpus linguistics approaches. Combining different approaches to corpus study and language ideologies, Purschke (2025) analyzes language ideologies on the RTL.lu corpus as part of public discourse formation.

Building on this foundation, we search for the following five language-ideological motifs about Luxembourgish and multilingualism in the country in the RTL.lu comments.

Identity: statements about national identity established by a common language, culture, and history, or personal preference towards the national language;

Example ID 636: *'1. For us Luxembourgers (250,000 people), Luxembourgish is our first language.'*²

Vitality: statements about the state of a language use, e.g., language endangerment and decay, reasons for these developments and/or means to prevent them;

Example ID 3631: *'Good text, but I'm afraid our language would die out.'*

Belonging: statements about requirements for speakers to integrate in the local society;

Example ID 410165: *'it worked: all residents of Luxembourg have linguistically integrated themselves.'*

Responsibility: statements delegating to an actor or a group of actors responsibility for the linguistic situation in the country (e.g. language decay or support);

Example ID 666: *'And if our cross-border commuters don't speak Luxembourgish, it's our own fault.'*

²Original texts are provided in the appendix.

Recognition: statements about language policies, language status, and how people perceive the languages spoken in the country.

Example ID 636: ‘2. *Luxembourgish is FIRST OF ALL a spoken language.*’

2.3. Ideologies and LLMs

There is abundant work showing how human stereotypes, biases, and ideologies influence the performance of (large) language models (Lin and Li, 2025). They can be introduced in the (pre-)training data (Hovy and Prabhumoye, 2021), optimization stages (Dahlgren Lindström et al., 2025), and data exposure level (Chen et al., 2024), they can change according to the tested model (Lin et al., 2024) and language (Buyl et al., 2025), and can be shaped by language ideologies. For instance, Gururangan et al. (2022) investigate the role of language ideologies in selecting text data and the subsequent inequalities. LLMs have also been shown to perpetuate in-group solidarity and out-group hostility (Hu et al., 2024) – where in-group refers to the community speakers socially and emotionally identify with, and out-group refers to communities perceived as outsiders. The language material generated by LLMs has been shown to reproduce social bias (Dhamala et al., 2021) as well as language ideologies, both for the standard language ideology in the US (Smith et al., 2025) and for gendered language ideology in English (Watson et al., 2025) and English dialects (Duncan, 2024).

2.4. Cross-lingual transfer in LLMs

Luxembourgish can be considered a small language in a multilingual context characterized by higher resourced languages. Almost all Luxembourgers are fluent in the other official languages, French or German (Fehlen et al., 2023), and despite Luxembourg’s membership in the European Union, Luxembourgish is not an official EU language. Additionally, Luxembourgish is not the language of education, and its domain expansion and societal uptake in writing are relatively recent (Gilles, 2015). For all these reasons, the amount of textual data to train large-scale NLP systems is limited. In such scenarios, it is common to resort to cross-lingual transfer techniques, ensuring that a model’s high performance on a specific task in a high-resource language can be transferred to a low-resource language. A simple cross-lingual transfer technique that can be easily applied in the context of generative LLMs is called *translate-test* (Ponti et al., 2021; Artetxe et al., 2023): the test data (in our case, the user comments written in Luxembourgish) are machine-translated to a high-resource language (in our case, English, French or German) and then fed to the model. While this setup is prone to error

propagation (the translations might contain errors), it is assumed that the performance gains obtained through the model’s better “understanding” of the translated data outweigh the errors introduced in the translation step.

3. Data

We frame ideology detection as a multi-class sentence annotation task over Luxembourgish user comments from the RTL.lu corpus. We operate at sentence level to reduce topic drift within long comments. Our corpus contains 300 comments with 1524 total sentences. Comments were split into sentences using GPT, and then manually corrected where the sentence splitting did not work correctly. The main reason for using GPT for this task is that, for social media data with irregular sentence constructions and punctuation patterns (e.g., including Emoji as sentence boundaries) other sentence splitters usually struggle and GPT, compared to these, includes semantic context parsing for better accuracy. We retained the punctuation for each user and did not carry out any other pre-processing steps such as lexical normalization, as these could blur ideology markers. In a setting without a fully implemented standard, the use of incorrect or correct orthography could be an expression of specific language ideologies.

The corpus is extracted from a larger set of user comments (1,422,759 comments at the time of the study³). It contains two subsets, one subset of 150 comments on language use in the country (from now on, we will refer to this subset as *Lang*), and one subset of 150 comments not related to this topic (referred to as *NotLang*). For the *Lang* subset, the selection of comments was made through keyword matching, using a total of 19 Luxembourgish variants of *Luxembourgish, our language, mother tongue*. Of the comments retrieved (16,015), 150 are then randomly selected and manually annotated.

For the *NotLang* subset, 75 comments were selected following the same list of keywords as for the *Lang* subset, and were carefully read to make sure that they refer to topics other than language use in the country. The remaining 75 comments are randomly selected from the complete set of comments (1,422,759), and do not contain language-related keywords. By diversifying the selection criteria across subsets, we aim to prevent the models from overfitting to surface-level lexical cues and to assess their performance in identifying ideologically charged discourse about language.

The *Lang* subset is annotated by two expert annotators (the first and last authors of this study) with

³For this study, we used the 2025-04 build of the RTL corpus.

the five language ideology categories presented in Section 2.2: IDENTITY, VITALITY, BELONGING, RESPONSIBILITY and RECOGNITION. We use a sixth label NONE for sentences that cannot be assigned one of these specific labels. The observed agreement between the two annotators is 0.78 and Cohen’s Kappa is 0.66. Disagreement cases are discussed by annotators and are labeled in common agreement. The category counts across the dataset are shown in Table 1.

Category	Lang	NotLang
Identity	49	0
Vitality	72	0
Belonging	63	0
Responsibility	121	0
Recognition	125	0
None	273	821
Total	703	821

Table 1: Sentence based gold standard annotation

4. Prompt Engineering

In this first experiment, we focus on evaluating different prompts in both zero- and few-shot scenarios.

Prompts We design four prompt configurations that vary the amount of information that is passed to the model.

- **Prompt 1** lists the labels, instructs the LLM to annotate each sentence with one of the five ideology labels or NONE and to provide a justification for the annotation. Text in quotation marks should not be annotated.
- **Prompt 2** adds three positive and three negative examples per category, and six additional negative examples to prompt 1. The examples are chosen because they are clear illustrations for the categories and are consistent throughout the experiment.
- **Prompt 3** adds detailed descriptions per category to prompt 1.
- **Prompt 4** adds both the examples of prompt 2 and the detailed descriptions of prompt 3 to prompt 1.

The structure of prompt 4, the most complex one, is illustrated in Figure 1. All prompts are written in English, and all comments remain in Luxembourgish. All prompts instruct models to return a single JSON object with fixed keys and value constraints, see Figure 2 below.

Experimental Setup We treat the task as multi-class annotation with sentence level ground truth derived from the human annotation. We report weighted F1, macro F1, and per label F1. We implement quality control to improve robustness without over-correcting model behavior. We validate every response against the JSON schema and re-issue malformed outputs up to 5 retries per sentence.

We test three models of the GPT family: GPT-4o (OpenAI et al., 2023), GPT-4o-mini (OpenAI, 2024), o3 (OpenAI, 2025). We standardize decoding to reduce variance induced by sampling. Temperature is set to 0.2, maximum tokens to 3000, however, we keep defaults for safety-related flags that affect refusal behavior. We batch requests with one comment (5 sentences on average) per call. Also, we enforce per thread rate limits of rps/5 to avoid provider throttling.

Results We observe minimal differences between GPT-4o, GPT-4o-mini, and o3 under prompt 4. This prompt also consistently outperforms other configurations across models, as F1 weighted score improves up to around 0.3 from prompt 1 to prompt 4 – see Table 2. Therefore, we adopt prompt 4 as the final prompting template for all further experiments.

Prompt	o3		GPT-4o		GPT-4o-mini	
	F1w	F1m	F1w	F1m	F1w	F1m
Prompt 1	0.734	0.431	0.491	0.248	0.563	0.280
Prompt 2	0.787	0.510	0.652	0.397	0.459	0.389
Prompt 3	0.806	0.567	0.765	0.427	0.468	0.391
Prompt 4	0.806	0.559	0.783	0.441	0.734	0.427

Table 2: Results per model and prompting conditions (Weighted F1 and Macro F1)

5. Evaluating LLMs and cross-lingual transfer settings

In this main experiment, we evaluate several widely used generative LLMs on the ideology annotation task with Prompt 4. We also vary the language of the comments, comparing annotation performance on the original Luxembourgish comments with their translations to English, French and German.

Machine translation To answer research question 2, we translate the original material into German, French and English via the Google Translate API. The automatic translations receive targeted manual checks⁴ that correct mistranslated country and language mentions. Additionally, 700 sentences evenly split between Lang and NotLang are

⁴Automatic translation quality for Luxembourgish can be uneven, therefore, subtle stance cues may shift.

Prompt Text
Detect one or more of the following categories in the sentences of the comment below, after the instructions and schemas. The categories are: Identity, Vitality, Belonging, Responsibility, Recognition. These categories concern language ideologies and only these exact categories can be used for the annotation. No overlap between categories is allowed, each sentence can have 0 or 1 category. If you cannot identify one of the category assign N/A. Do not take any quoted parts (i.e. text in quotation marks) into account for the labelling. For each sentence, explain why you have assigned that label. Here are the explanations of the five categories: [...]
Recognition: The language related ideology 'recognition' groups opinion on Luxembourgish language and other languages spoken in the country. In this category, different ways of acknowledging languages are involved: [...]
Examples: [...]
Examples where this label is not applicable: [...]

Figure 1: Prompt 4 comprehensive of instructions, definitions, and examples

```
"sentence_id": "<string>",
"text": "<original sentence>",
"labels": ["Identity" | "Vitality"
| "Belonging" | "Responsibility" |
"Recognition" | "None"],
"rationale": "<justification>"
```

Figure 2: JSON representation of the required schema.

manually reviewed to improve translation quality; correcting mistakes due to orthographic variation and lack of capitalization, and adapting ideology bearing content to the target languages. The distribution of categories within the two translation sets is described in Table 3. We run the same prompting and inference pipeline on each language set. Therefore, any observed changes can be attributed to language rather than prompt drift.

Labels	Post-Edited	Automatic
Identity	18	31
Vitality	22	50
Belonging	16	47
Responsibility	64	57
Recognition	56	69
None	525	569

Table 3: Translation sets

Models We evaluate widely used frontier and open generative LLMs for applied annotation workflows. The frontier group includes GPT-5 (OpenAI, 2025), while the open group includes Aya Expansive-32B (Aya; Dang et al. 2024), Llama 4 Scout 17B Instruct (Llama; AI 2025a), Mistral Magistral Small 24B (Magistral; AI 2025b), Qwen3-Next-80B-A3B-Thinking (Qwen; Yang et al. 2025), DeepSeek-V3 (DeepSeek; Liu et al. 2024), and gpt-oss-20b (GPT-OSS; OpenAI et al. 2025). All models follow the experimental setup described in Section 4.

Lang.	DeepSeek		GPT-5		Magistral		Qwen	
	F1w	F1m	F1w	F1m	F1w	F1m	F1w	F1m
Luxemb.	0.537	0.478	0.558	0.497	0.436	0.351	0.500	0.380
German	0.518	0.469	0.551	0.500	0.450	0.381	0.529	0.399
French	0.507	0.474	0.585	0.525	0.461	0.406	0.532	0.404
English	0.490	0.446	0.596	0.542	0.462	0.346	0.514	0.392

Table 4: Comparison of models per language (Weighted F1 and Macro F1)

Results Of the models tested, GPT-OSS is discarded because it is unable to produce consistent output. Aya and Llama are also discarded, as they return up to 10 and 9 labels respectively even when explicitly stating the number of labels in the prompt – probably because of their limited reasoning capabilities compared to the other tested LLMs. Although Qwen and Magistral also sporadically return additional categories, we keep them in our experiments as the categories are justified by the model-generated explanations as a subset of the main ones (e.g., PERSONAL IDENTITY is mapped to IDENTITY).

Table 4 shows the performance of the remaining models broken down by language. While DeepSeek gets the highest scores with the original Luxembourgish data, the remaining models work slightly better with the translations. GPT-5 outperforms the other models in all languages, and the best overall performance is obtained with the English translations.

5.1. Translation effects

Zooming in on the two different sets of translations, we see that the quality of the translations does not affect model performance. For both translation sets, the best performing model is GPT-5 in all languages. As shown in Table 5, the GPT-5 performance gap between English and Luxembourgish is larger with automatic translations than with post-edited ones. For DeepSeek, Luxembourgish seems to work better than English both for automatic translations and post-edited ones.

Lang.	DeepSeek		GPT-5		Magistral		Qwen	
	Auto	PE	Auto	PE	Auto	PE	Auto	PE
Luxemb.	0.762	0.771	0.797	0.780	0.699	0.737	0.755	0.661
German	0.747	0.783	0.803	0.767	0.728	0.738	0.774	0.753
French	0.731	0.781	0.806	0.800	0.721	0.744	0.768	0.766
English	0.743	0.750	0.829	0.783	0.724	0.746	0.767	0.740

Table 5: Comparison of models on Automatic (Auto) and Post-Edited (PE) translations across languages and on the original Luxembourgish corpus with the same split (Weighted F1)

Discussion To answer research question 2, translating to more-represented languages in NLP produces a limited performance improvement, whereas we find no improvement in the task between using automatic and post-edited translation sets. We consider two potential explanations to this: cross-lingual transfer in LLMs and the challenge of transferring cultural and social meanings from the source text to the target languages.

On the one hand, Luxembourgish shares some grammatical and lexical features with high-resource languages – mainly German and French – which are represented in the models’ training data and officially supported by them. Therefore, it is quite likely that the tested models generalize from shared multilingual representations (Zhang et al., 2024).

On the other hand, the way sociocultural meanings are treated in the translation process may also negatively impact the classification performance. For the post-edited set, we tried to find a middle ground between preserving the same language ideologies, reporting appropriate linguistic cues to each target language, and avoiding introducing noise in the translations. For instance, capitalization for emphasis is kept, while lack of orthographic capitalization is corrected in the translations. In the non-post-edited set, only mistranslated country and language mentions are corrected. However, our results do not show an increase in annotation performance that could be ascribed to the different translation conditions. Similarly, there is no significant performance improvement between the original Luxembourgish comments and the translations. This is attributable to the failure to eliminate noise and reproduce ideology in target languages.

Due to these challenges, and to the LLMs’ relatively even performances across languages, keeping the original text seems more appropriate.

5.2. Label-specific results

Table 6 shows F1 scores averaged on the four languages per category and model. All analyzed models perform better for discriminating between items with and without language ideologies than for distinguishing between individual categories. In fact, the absence of language ideologies – the label NONE – is the category with higher F1 scores com-

pared to the others (around 0.9). Compared to our gold standard, RESPONSIBILITY and RECOGNITION are under-detected, while VITALITY, IDENTITY, and BELONGING are over-predicted.

DeepSeek and GPT-5 achieve the highest per-label F1 scores among the evaluated models. Table 7 shows per-label F1 scores for GPT-5 and DeepSeek on the original data, compared to the highest F1 score overall. GPT-5 has the highest scores in English for BELONGING (0.571) and VITALITY (0.555), and in French for RECOGNITION (0.565). DeepSeek has highest scores in French for IDENTITY (0.475) and in Luxembourgish for RESPONSIBILITY (0.398). These values are not significantly higher than the F1 scores obtained for Luxembourgish for DeepSeek and GPT-5, except for RECOGNITION for DeepSeek (0.330).

Analysis Here, we examine the annotation categories in detail, focusing on the original Luxembourgish data and the Lang subset. We discuss consistent overlaps among categories, as well as potential causes of mislabeling. To understand the cause of systematic overlaps, we refer both to the content of the sentence and to the models-generated explanations. The benefit of models-generated explanations is discussed in the following section.

False negatives are observed for the category RECOGNITION in all other categories. Reading the explanations, the models seem to consider the hierarchy among languages – which we subsume under RECOGNITION – as a personal or national preference towards one or more languages. In the Luxembourg ideological landscape, the use of the pronoun *our* (*eis*) can indicate a reference to the national in-group. This information is provided as part of the IDENTITY category description. Therefore, also when not carrying an IDENTITY ideology, first-person plural pronouns and mentions of the ingroup are used as markers of national identity (Magistral and DeepSeek). Overlap with VITALITY arises when sentences describe the limited use of Luxembourgish in the country (Magistral, GPT-5, DeepSeek, Qwen). RECOGNITION is then mislabeled as (social) BELONGING when Luxembourgish is stated as the main language of the country (DeepSeek and Qwen). Furthermore, comments perceived as blaming the current language situation on actors mentioned in the text are labeled as RESPONSIBILITY (Qwen).⁵

For RESPONSIBILITY, it is mislabeled as BELONGING when the responsible actor is an out-group member (Qwen, GPT-5), and as RECOGNITION when it is an official institution or a private person (Qwen). It is then mislabeled as VITALITY when the implications of language decay are considered

⁵Further examples and statistics are provided in the appendix.

Label	DeepSeek			GPT-5			Magistral			Qwen		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Belonging	0.468	0.603	0.527	0.541	0.492	0.516	0.466	0.329	0.386	0.365	0.468	0.410
Identity	0.338	0.526	0.411	0.504	0.362	0.421	0.216	0.612	0.319	0.341	0.459	0.391
Recognition	0.502	0.218	0.304	0.479	0.608	0.536	0.447	0.092	0.153	0.458	0.348	0.395
Responsibility	0.603	0.242	0.345	0.689	0.211	0.323	0.667	0.190	0.296	0.368	0.295	0.328
Vitality	0.554	0.375	0.447	0.481	0.611	0.538	0.435	0.382	0.407	0.278	0.625	0.385
None	0.858	0.952	0.903	0.908	0.950	0.929	0.854	0.950	0.899	0.926	0.864	0.894

Table 6: Comparison of models per category

Label	DeepSeek	GPT-5	Highest
Belong.	0.487	0.512	0.571 (EN)
Identity	0.376	0.386	0.475 (FR)
Recogn.	0.330	0.520	0.565 (FR)
Respons.	0.398	0.299	0.398 (LU)
Vitality	0.448	0.500	0.555 (EN)
None	0.910	0.927	0.934 (EN)

Table 7: Per-class F1-scores of DeepSeek and GPT-5. The first two columns show F1-scores for DeepSeek and GPT-5 respectively, on the **Luxembourgish** dataset. The third column reports the absolute best F1 scores per label, regardless of the language of the dataset.

to be the main ideology of the item (Qwen, GPT-5, Magistral). Additionally, the use of first-person pronouns is reported in the model-generated explanations to justify the `IDENTITY` label (Magistral, DeepSeek).

`VITALITY` is mislabeled as `IDENTITY` when language decay is associated with national decay (DeepSeek), and through the use of first-person plural pronouns (Qwen). According to the model-generated explanations, items encouraging the use of the national language as a preservation mechanism are annotated as `RECOGNITION` (GPT-5).

`BELONGING` is mislabeled as `IDENTITY` when Luxembourgish is stated as the language of society (Magistral, Qwen). The presence of the in-group in these utterances makes it difficult for the models to discriminate between the national language as a characteristic of the nation self-determination – which we subsume under `IDENTITY` – and as a means of social exclusion – which we subsume under `BELONGING`.

Finally, `IDENTITY` is mislabeled as `RECOGNITION` when the model (GPT-5) fails to disambiguate between the use of language to create and reinforce a national feeling – which we subsume under `IDENTITY` – and the description of language use in the country – which we subsume under `RECOGNITION`.

5.3. Examining model-generated explanations

The language ideology definitions developed by the two annotators were given to the LLMs as part of the prompt. However, these guidelines were not always efficient, showing that defining language ideologies for LLMs introduces different challenges compared to human annotators. For instance, since language ideologies might share both topics and linguistic features to express them, the annotators already faced (and solved) cases of overlapping categories. Despite providing specific guidelines, these cases are still error sources for LLMs – together with unexpected ones. In this context, the explanations provided by the LLMs are particularly valuable, as they help us to understand how definitions might cause mislabeling, and how clearly delineated the categories are.

Table 8 provides an example of a model-generated explanation. The annotated item is shown in its English translation, and is followed by the labels assigned by the evaluated models and the explanations. The `IDENTITY` label is justified by the use of the first-person plural pronoun *we* (*mir*) as a reference to the national group, and by the use of Luxembourgish in opposition to German and French. The `VITALITY` explanations show that inciting people to use more Luxembourgish words is considered a strategy against language decay. This example was annotated as `RECOGNITION` as it shows a desired hierarchy of language use opposed to the perceived status quo. However, ambiguity is induced by the exhortation of using Luxembourgish *more*, that adds a hint to another ideological layer – `VITALITY` – to the language hierarchy. Similar examples caused label overlaps also for human annotators, but are already disentangled in the prompt given to the models.

Language Ideology Definitions In general, clues which seem helpful for human annotators resulted in confusion between categories for the models. For instance, the use of specific pronouns and determiners is a useful hint for human annotators to discriminate between `IDENTITY` and other categories, but it causes an over-prediction of this category in the models – as shown in the example

Dear – We should use more Luxembourgish words, and not German words, nor French words.
Gold standard: RECOGNITION
Magistral : IDENTITY 'The sentence refers to the national group using the first-person plural pronoun ('mir') and emphasizes the use of Luxembourgish language in opposition to other languages (German and French). This aligns with the national identity category, as it highlights the importance of Luxembourgish language for Luxembourgers.'
DeepSeek: VITALITY 'The sentence emphasizes the need to use Luxembourgish words instead of German or French words, which aligns with the 'vitality' category as it addresses the preservation of Luxembourgish language.'
Qwen: VITALITY 'The sentence expresses concern about the use of German and French words instead of Luxembourgish words, which is a typical indicator of language decay and threat to the vitality of Luxembourgish. The advocacy for using more Luxembourgish vocabulary over foreign alternatives directly relates to preserving the language's integrity and resisting linguistic erosion.'
GPT-5: VITALITY 'The speaker urges using more Luxembourgish words instead of German or French ones, proposing a way to safeguard the language against lexical replacement; this fits concerns and remedies typical of vitality.'

Table 8: Examples of models-generated explanations. The first line reports the annotated sentence, followed by labels and explanations by each model

above. Then, for RESPONSIBILITY, human annotators identified a subject of responsibility, as well as an object. Three different subjects of responsibility have been found – in-group, politician, out-group – and given to the model together with the description of the category. From the explanations, it seems that the presence of the three different responsibility agents leads the tested models to mislabel the category as RECOGNITION – official institutions as subject of responsibility – and BELONGING – the out-group as subject of responsibility.⁶

Language Ideology Labels The analysis in the previous paragraphs suggests that the five language ideology labels might be too fine-grained for reliable prediction with simple in-context learning. Therefore, to answer research question 3: LLMs require expert support or task-specific model adaptation to adequately inform sociolinguistic research. This raises the question whether a more coarse-grained classification scheme would be beneficial. At the same time, the example also demonstrates how LLMs could serve as a valuable tool to improve human annotations and improve them, especially in cases of doubt.

For instance, both IDENTITY and BELONGING, although referring to two different social groups, convey feelings of social belonging obtained and constructed through language use. However, these ideologies highlight a power imbalance that will be lost if the categories are merged, i.e., the dominant

⁶See Gal and Irvine (2019) for more information about language ideologies definitions.

group constructs its own identity – IDENTITY – and shapes the role of other people in society – BELONGING. Similarly, statements about language preference could be grouped under one label, but it will cause losing implications between IDENTITY – personal preference – and RECOGNITION – reference to language status and hierarchy among languages. Therefore, restructuring the categorization scheme could mean losing important aspects of societal power dynamics for sociolinguistic analyses.

6. Conclusions

This paper offers an evaluation of LLMs for language ideology detection. It shows that Luxembourgish does not need to be translated to higher-resourced languages in order to obtain satisfactory classification performance with generative LLMs. It highlights the crucial role of model-generated explanations for this task and that the evaluated models successfully discriminate items with language ideologies from those without. However, discriminating among the different categories of ideologies turns out to be more challenging. The first research question of the paper has thus both a positive and a negative answer: current generations of LLMs can perform a binary language ideology classification, but a fine-grained classification seems to be a more challenging task. Although it might not help with further understanding language ideologies and language discourse in sociolinguistic research, a binary classification between content with and without language ideologies already seems to be a quite useful task, mainly for selecting material to

be analyzed. Our approach could be repurposed for other less-represented languages in NLP, to test if translations of original material could benefit the automatic analysis of language ideologies on the source language. However, given the complexity of the ideology classification task, this would quite likely result in an overview of the presence/absence of language ideologies in a discourse and, hence, require an in-depth analysis from trained (socio)linguists to understand how the motifs are implemented and whether additional language ideologies characterize the discourse in question.

As for the nature of the task, language ideologies do not have a unique linguistic encoding: they may share topical vehicles, the context of utterance, as well as grammatical and lexical features, and are quite challenging for human annotators to discriminate. As including human annotators guidelines in the prompt design turned out to be of limited usefulness for LLMs, new paths should be considered. One option, discussed above, could be the simplification of the annotation scheme. Another option could consist in collecting more manually annotated data to enable fine-tuning. The challenge will be to find a balance between sociolinguistic interests, LLM capabilities and human annotation efforts.

Although LLMs show promise as tools for sociolinguistic inquiry, their potential for the study of language ideologies remains limited without the guidance of human experts. Mainly, language ideologies cannot be studied only using LLMs with no previous understanding of the discourse in analysis – as demonstrated in the trial phase. Mixed methods seem to be a promising way to enrich sociolinguistic research with quantitative methods and NLP advancements (Kircher and Hawkey, 2022; Nguyen et al., 2016; Purschke, 2025).

In future experiments, we will focus on providing more general descriptions instead of detailed, example-based ones. We will improve them to highlight both core and specific elements per category, to give space to infrequent ideological expressions, and to reframe features frequently leading the models to mislabel certain categories.

Limitations

As the project aimed to understand how two disciplines – sociolinguistics and natural language processing – can be mutually informative for each other, we are aware of limitations from both fields. As per the goal of the project, keyword searching should be adjusted to include language ideologies concerning Luxembourgish. Expanding the keyword search to include the different languages spoken in the country would help to get a more comprehensive view of language ideologies in the country. Additionally, we refer to the LLM-generated annota-

tions to understand what could be improved in the prompt design, but we are aware of their limitations in accurately describing the reasoning behind the classification (Randl et al., 2025; Di Bonaventura et al., 2024). Concerning the choice of models, two main limitations can be reported. First, models with limited reasoning capabilities proved to be not robust enough for this task and will be discarded for future experiments. Second, since we only had access to a small amount of annotated data, we were unable to train or fine-tune models on the task. Future work should explore this direction, as improved model performance could support language discourse understanding, as well as its societal implications. However, although a fine-tuned model could achieve high performances, human annotators and linguistic experts remain essential to understand the consistency of model outputs.

We noticed a grammatical mistake in the prompt used to preform this classification: for IDENTITY we state that it is expressed by the use of the first-person plural pronoun *eis*. However, *eis* is the dative form of the first-person plural pronoun *mir*, and its possessive determiner. As this could be the cause of some labeling errors, we will implement the task further with the correct version of the prompt.

Ethical Statement

Data used for the project are provided and anonymized by the data owner RTL.lu.

Acknowledgments

This research was supported by the Luxembourg National Research Fund (Project code: C22/SC/117225699). We would like to thank the members of the TRAVOLTA project and the Language and Technology Group of the University of Oslo for their invaluable advice on the project, and Raul Ian Sosa for his precious feedback.

Bibliographical References

- Meta AI. 2025a. [Llama 4 scout 17b instruct](#).
- Mistral AI. 2025b. [Magistral](#). *arXiv preprint arXiv:2506.10910*.
- Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. [Revisiting machine translation for cross-lingual classification](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6489–6499, Singapore. Association for Computational Linguistics.

- Paul Baker and Erez Levon. 2015. [Picking the right cherries? A comparison of corpus-based and qualitative analyses of news articles about masculinity](#). *Discourse & Communication*, 9(2):221–236.
- Jan Blommaert, editor. 2010. *Language Ideological Debates*, originally published 1999 edition. Number 2 in Language, Power and Social Process. De Gruyter, Berlin.
- Thomas Blume. 2003. *Ideologie*. Vandenhoeck & Ruprecht.
- Maarten Buyl, Alexander Rogiers, Sander Noels, Guillaume Bied, Iris Dominguez-Catena, Edith Heiter, Iman Johary, Alexandru-Cristian Mara, Raphaël Romero, Jeffrey Lijffijt, and Tijn De Bie. 2025. [Large Language Models Reflect the Ideology of their Creators](#). ArXiv:2410.18417 [cs].
- Kai Chen, Zihao He, Jun Yan, Taiwei Shi, and Kristina Lerman. 2024. [How Susceptible are Large Language Models to Ideological Manipulation?](#)
- Adam Dahlgren Lindström, Leila Methnani, Lea Krause, Petter Ericson, Íñigo Martínez de Rituerto de Troya, Dimitri Coelho Mollo, and Roel Dobbe. 2025. [Helpful, harmless, honest? Sociotechnical limits of AI alignment and safety through reinforcement learning from human feedback](#). *Ethics and Information Technology*, 27(2):28.
- John Dang, Shivalika Singh, Daniel D'souza, et al. 2024. [Aya Expanse: Combining research breakthroughs for a new multilingual frontier](#). *arXiv preprint arXiv:2412.04261*.
- Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. [BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation](#). In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 862–872. ArXiv:2101.11718 [cs].
- Chiara Di Bonaventura, Lucia Siciliani, Pierpaolo Basile, Albert Meroño-Peñuela, and Barbara McGillivray. 2024. Is explanation all you need? an expert survey on llm-generated explanations for abusive language detection. In *Proceedings of the 10th italian conference on computational linguistics (CLiC-it 2024)*, pages 280–288.
- John Duncan. 2024. [Does ChatGPT have sociolinguistic competence?](#) *Journal of Computer-Assisted Linguistic Research*, 58:123–145.
- Fernand Fehlen, Peter Gilles, Louis Chauvel, Isabelle Pigeron-Piroth, Yann Ferro, and Etienne Le Bihan. 2023. *Une diversité linguistique en forte hausse*. STATEC, Luxembourg, Luxembourg.
- Susan Gal and Judith T. Irvine. 2019. *Signs of Difference: Language and Ideology in Social Life*. Cambridge University Press.
- Peter Gilles. 2015. *From Status to Corpus: Codification and Implementation of Spelling Norms in Luxembourgish*. Palgrave Macmillan UK.
- Peter Gilles. 2019. [39. Komplexe Überdachung II: Luxemburg. Die Genese einer neuen Nationalsprache](#). In Joachim Herrgen and Jürgen Erich Schmidt, editors, *Sprache und Raum - Ein internationales Handbuch der Sprachvariation. Volume 4 Deutsch*, pages 1039–1060. De Gruyter Mouton, Berlin, Boston.
- Stefan Grondelaers and Roeland van Hout. 2011. The standard language situation in the Netherlands. In Tore Kristiansen & Nik Coupland, editor, *Standard languages and language standards in a changing Europe*, page 113–118. Novus.
- Suchin Gururangan, Dallas Card, Sarah K. Dreier, Emily K. Gade, Leroy Z. Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. 2022. [Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection](#). ArXiv:2201.10474 [cs].
- Kristine Horner and Jean Jacques Weber. 2008. [The language situation in luxembourg](#). *Current Issues in Language Planning*, 9:69–128.
- Dirk Hovy and Shrimai Prabhumoye. 2021. [Five sources of bias in natural language processing](#). *Lang. Linguistics Compass*, 15(8).
- Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander Van Der Linden, and Jon Roozenbeek. 2024. [Generative language models exhibit social identity biases](#). *Nature Computational Science*, 5(1):65–75.
- Ruth Kircher and James W. Hawkey. 2022. [Mixed-methods approaches to the study of language attitudes](#). In Ruth Kircher and Lenore A. Zipp, editors, *Research Methods in Language Attitudes*, pages 330–345. Cambridge University Press.
- Paul V. Kroskrity. 2022. [5 Language ideologies and social identities](#). In Svenja Völkel and Nico Nassenstein, editors, *Approaches to Language and Culture*, pages 101–126. De Gruyter.
- Luyang Lin, Lingzhi Wang, Jinsong Guo, and Kam-Fai Wong. 2024. [Investigating Bias in LLM-Based](#)

- Bias Detection: Disparities between LLMs and Human Perception. ArXiv:2403.14896 [cs].
- Xinru Lin and Luyang Li. 2025. Implicit bias in LLMs: A survey. *arXiv preprint arXiv:2503.02776*.
- Rosina Lippi-Green. 2012. *English with an accent: Language, ideology and discrimination in the United States*. Routledge.
- Aixin Liu, Bei Feng, Bing Xue, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Anne-Marie Lutgen, Alistair Plum, Christoph Purschke, and Barbara Plank. 2025. Neural text normalization for Luxembourgish using real-life variation data. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 115–127, Abu Dhabi, UAE. Association for Computational Linguistics.
- Bolei Ma, Xinpeng Wang, Tiancheng Hu, Anna-Carolina Haensch, Michael A. Hedderich, Barbara Plank, and Frauke Kreuter. 2024. The Potential and Challenges of Evaluating Attitudes, Opinions, and Values in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8783–8805, Miami, Florida, USA. Association for Computational Linguistics.
- Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.
- Dong Nguyen, A. Seza Dođruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Computational Linguistics*, 42(3):537–593.
- OpenAI. 2024. Gpt-4o mini explained: Fast, affordable ai model with 128k context.
- OpenAI. 2025. Gpt-5 system card. Technical report, OpenAI. Accessed: 2025-10-20.
- OpenAI. 2025. Gpt-o3 model overview.
- OpenAI et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- OpenAI et al. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.
- Fred Philippy, Shohreh Haddadan, and Siwen Guo. 2024. Forget NLI, use a dictionary: Zero-shot topic classification for low-resource languages with application to Luxembourgish. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 97–104, Torino, Italia. ELRA and ICCL.
- Alistair Plum, Caroline Döhmer, Emilia Milano, Anne-Marie Lutgen, and Christoph Purschke. 2024. LuxBank: The first Universal Dependency treebank for Luxembourgish. In *Proceedings of the 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024)*, pages 30–39, Hamburg, Germany. Association for Computational Linguistics.
- Alistair Plum, Tharindu Ranasinghe, and Christoph Purschke. 2025. Text generation models for Luxembourgish with limited data: A balanced multilingual strategy. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 93–104, Abu Dhabi, UAE. Association for Computational Linguistics.
- Edoardo Maria Ponti, Julia Kreutzer, Ivan Vulić, and Siva Reddy. 2021. Modelling latent translations for cross-lingual transfer.
- Christoph Purschke. 2023. Diskurs-Figuren. Wie Politik und Öffentlichkeit in Luxemburg über Sprache sprechen. *Hemecht: Zeitschrift für Luxemburger Geschichte*, (2023/3):313–330.
- Christoph Purschke. 2025. Discourse figures in the luxembourg language debate (2015–2020). *Journal of Language Variation and Sociolinguistics*, 1(2):37–53.
- Christoph Purschke and Peter Gilles. 2023. Sociolinguistics in Luxembourg. In *The Routledge handbook of sociolinguistics around the world*, pages 542–549. Routledge.
- Tharindu Ranasinghe, Alistair Plum, Christoph Purschke, and Marcos Zampieri. 2023. Publish or hold? Automatic comment moderation in Luxembourgish news articles. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 968–978, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- Korbinian Randl, John Pavlopoulos, Aron Henriksson, and Tony Lindgren. 2025. Mind the gap: from plausible to valid self-explanations in large language models. *Machine Learning*, 114(10):220.
- Catherine R Rhodes. 2023. Language ideologies. In *Oxford Research Encyclopedia of Anthropology*. Oxford University Press.
- Genevieve Smith, Eve Fleisig, Madeline Bossi, Ishita Rustagi, and Xavier Yin. 2025. Standard Language Ideology in AI-Generated Language. ArXiv:2406.08726 [cs].

Rachelle Vessey. 2017. [Corpus approaches to language ideology](#). *Applied Linguistics*, 38(3):277–296.

Julia Watson, Sophia S. Lee, Barend Beekhuizen, and Suzanne Stevenson. 2025. [Do language models practice what they preach? Examining language ideologies about gendered language reform encoded in LLMs](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1201–1223, Abu Dhabi, UAE. Association for Computational Linguistics.

Kathryn A. Woolard. 2020. [Language Ideology](#). In James Stanlaw, editor, *The International Encyclopedia of Linguistic Anthropology*, 1 edition, pages 1–21. Wiley.

A Wu, Ryan Liu, Kerem Oktar, T Sumers, and T Griffiths. 2025. Are large language models sensitive to the motives behind communication? *Advances in Neural Information Processing Systems (NeurIPS)*, 38.

An Yang, Anfeng Li, Baosong Yang, et al. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.

Zhihao Zhang, Jun Zhao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. [Unveiling linguistic regions in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6228–6247, Bangkok, Thailand. Association for Computational Linguistics.

Appendix A

In this appendix, we report the original Luxembourgish version of the examples in Section 2.2:

- **Example ID 636:** 1. Lëtzebuergesch ass fir eis Lëtzebuurger (250'000 Leit) eis éischt Sprooch.
- **Example ID 3631:** Gudden Text, mee ech färten, eis Sprooch wärt ausstiewen.
- **Example ID 410165:** et ass geschafft: All Awunner vu Lëtzebuerg hu sech sproochlech integréiert.
- **Example ID 666:** A wann eis Frontalieren kee Letzebuergesch schwaetzen dann ass daat eis eege schold.
- **Example ID 636:** 2. Lëtzebuergesch ass VIRUN ALLEM eng geschwaate Sprooch.

Appendix B

Keywords used to retrieve the comments:

'eis sprooch', 'eis sproch', 'ons sprooch', 'ons sproch', 'lëtzebuergesch', 'lëtzeboiesch', 'letzebuergesch', 'lëtzeburgesch', 'lëtzbuergesch', 'letzeburgesch', 'letzeboiesch', 'lëtzebuergesch', 'eiser sprooch', 'onser sprooch', 'eiser sproch', 'onser sproch', 'lëtzebuergesch', 'mammesprooch', 'mammesproch'.

Appendix C

In this appendix, we add the complete version of Prompt 4:

Detect one or more of the following categories in the sentences of the comment below, after the instructions and schemas. The categories are: Identity, Vitality, Belonging, Responsibility, Recognition. These categories concern language ideologies and only these exact categories can be used for the annotation. No overlap between categories is allowed, each sentence can have 0 or 1 category. If you cannot identify one of the category assign N/A. Do not take any quoted parts (i.e. text in quotation marks) into account for the labeling. For each sentence, explain why you have assigned that label. Here are the explanations of the five categories:

Identity: The language related ideology 'identity' groups both national and personal identity. National identity: it refers to Luxembourgish national identity and it only concerns Luxembourgish people, born in Luxembourg with Luxembourgish as native language. It is characterized by a strong emotional attachment to the language and culture of Luxembourg, in opposition to people with different languages and cultures. It is found in clauses or part of clauses underling:

- the tied connection between national identity and national language (example: Luxembourgishers unified by a common national language)
- the tied connection between Luxembourgish culture and Luxembourgish language (example: importance of Luxembourgish language and culture)
- the formation of the Luxembourgish national identity during the Second World War (Luxembourgishers explicitly chose Luxembourgish language as national language for the first time)
- the opposition between the national group (= born and raised in Luxembourg, by family of Luxembourgish origins, with Luxembourgish culture and Luxembourgish as native language) and the group seen as external to the

country (= not born in Luxembourgish families, or coming from abroad, or living across the borders but going to Luxembourg for work. These people are grouped by lack of knowledge of Luxembourgish language and refusal to learn it)

Grammatically, this category is expressed by the use of the first-person plural pronoun ('eis') when referring:

- to the national group
- to what belongs to the national group
- to what characterizes the national group.

'eis' is used in opposition to the third person plural pronoun ('si'). The third person plural pronoun refers to the group seen external to the country. Personal identity = personal preference towards a language with no reference to any group-shared ideology or feeling (examples = speaking Luxembourgish as personal preference and not as a national identity trait)

examples:

1. lëtzebuergesch ass fir eis lëtzebuerger (250'000 leit) eis éischt sprooch.
2. et gött leit de zu dachau agespaart ware wéint eiser sprooch an e vollek dat seng sprooch verléiert verléiert och seng identitéit. . .
3. ech perséinlech maan mer emmer en spaass draus vir mat de verschiddenen wierder ze spillen wann ech eppes schreiwen

examples where this label is not applicable:

1. tschüss!
2. mee mir kann et dee moment egal sinn, well ech da wahrscheinlech an däitschland wunnen... tschüss!
3. patricia courtois

Vitality: The language related ideology 'vitality' groups opinions about Luxembourgish language considered:

- endangered
- threatened
- in decay
- soon to be replaced by other languages

Threatening factors are:

- languages other than Luxembourgish spoken in the country

- impossibility of speaking Luxembourgish in everyday contexts (shops, hospitals, ...)
- languages other than Luxembourgish used at official levels
- lack of use or incorrect use of Luxembourgish orthographical and grammatical norms
- use of non-Luxembourgish words instead of available Luxembourgish words
- scarcity of Luxembourgish speaking people

Proposing ways to preserve and safeguard Luxembourgish is also grouped in this category. The focal element of this category is always the language situation. Even when proposing possible ways to save the language and reporting the role policy-makers have in this, the focus always stays on the language situation and does not shift on secondary elements. There are a few semantic fields typical of this category: decay, endangerment, violence (towards the language).

examples:

1. eng dagesmamm déi lëtzebuergesch schwätzt ze fannen, ass quasi onméiglech. awer eng babá fënnt ee séier
2. soss gi mer all frankophon gemaach
3. ech denken sou seier geht dat net verluer

examples where this label is not applicable:

1. bonjour a vill gléck a neie joer.
2. all guddes an deem sënn.
3. ech verstinn eent net, léif fra.

Belonging: The language related ideology 'belonging' groups answers to the following questions:

- Do people born outside Luxembourg belong in Luxembourg?
- Do people speaking Luxembourgish at a basic level belong to Luxembourg?
- Do people not learning Luxembourgish belong to Luxembourg?

This category is about the integration in Luxembourgish society of people coming from abroad or people of non-Luxembourgish descents. It always involves the presence of foreigners, or of people not speaking Luxembourgish or not being Luxembourgish native speakers. Two main factors are significant in this category: 1) Luxembourgish language, 2) foreigners. These two factors intertwine as follows:

- Luxembourgish language is seen as the only means of integration in Luxembourgish society for foreigners
- Foreigners are expected to learn Luxembourgish
- Foreigners refuse to learn Luxembourgish language
- Luxembourgish language has to be learned at school for foreigners not to be disadvantaged in society
- The required level of Luxembourgish to obtain the citizenship is too low
- Foreigners who learn or try to learn/speak Luxembourgish are welcomed in Luxembourgish society
- Foreigners adapting to Luxembourgish language and culture are part of Luxembourgish society

examples:

1. eischt offiziell sprooch awer onbedengt fir e kloert zeeche fir d'integratioun ze setzen, fir eisen auslännesche matbierger kloer ze verstoen ze ginn dass mir eng eegestänneg denkweis, sprooch a kultur hunn an si sech eis unzepassen hunn andeems si net just 1 joer lëtzebuergesch léieren mee bis si et sou kënnen dass si et och kënnen schwätzen a verstoen
2. ween lëtzebuenger wëll ginn muss och lëtzebuergesch schwätzen
3. ganz traureg daat do. wann et engem schlecht geht muss een sech kennen an senger mamesprooch ausdrecken an et huet een et nët néideg sech vun schlecht gelaanten franséischen infirmièren ungranzten ze loossen. déi däitsch maachen jo nach éischer en effort an probéieren lëtzebuergesch ze schwätzen

examples where this label is not applicable:

1. ierch all een schéinen owend
2. där meenung sin ech och.
3. wou soll daat dann hin feieren?

Responsibility: The language related ideology 'responsibility' groups opinions about who is considered responsible for Luxembourgish decay. This category has two main elements: 1) responsible agent, 2) subject of responsibility. The subject of responsibility is the Luxembourgish language decay (everything defined in 'vitality' category) Possible responsible agents are:

- Politicians and policymakers: because they don't defend Luxembourgish language through adequate language policies and requires a low Luxembourgish language level to acquire Luxembourgish citizenship
- Luxembourgers: because they never impose themselves to speak Luxembourgish with foreigners, they never ask foreigners to learn Luxembourgish, but adapt to foreigners with English, French, or German
- Foreigners: because they refuse to learn/speak Luxembourgish and impose the languages they can speak (mainly French and English)

These responsible agents are also seen as responsible for a possible solution to the subject of responsibility. Other expressions fitting in this category involve Luxembourgers defending themselves from being called being racist, like 'we are not racist, but...', 'they call us racist when we advocate for Luxembourgish.'

examples:

1. dann sollen dei ausländesch elteren emol ufänken eis sprooch ze leieren mee dofir sin mer ze faul
2. et ass wéi am palais, do gët och keen lëtzebuergesch geschwaat.sie hun problemer vir sech an eiser sprooch ze artikuléieren
3. et as leider net einfach ze verlaangen, dass soll letzebuergesch geschwaat soll gin, well soss get een nämlech ganz ganz ganz seier asl rassist ugesin. wann dir am geschäft engem soot e soll letzebuergesch schwetzen, da get een einfach lenks leihen gelos an färdeg

examples where this label is not applicable:

1. also muss daat och hei goen.
2. schummt iech!
3. as daat lo wierklech den sprengenden punkt vun desem artikel?

Recognition: The language related ideology 'recognition' groups opinion on Luxembourgish language and other languages spoken in the country. In this category, different ways of acknowledging languages are involved:

- Luxembourgish considered as an oral language or/and a dialect
- Luxembourgish considered as an official, standardized language, whose norms have to be followed by speakers/writers

- Luxembourgish has to be recognized at the European level as one of the official languages of the European Union
- Hierarchy among languages

examples:

1. mee ech verstinn haut nach ëmmer net, wéisou alles, awer och wierklech alles hei zu lëtzebuerg op franséisch leeft.
2. jo mir hun eng nationalsprooch, an déi solle mir och héich halen, mee ons stäerkt war ëmmer, dass mir dräisproocheg waren, an och houfreg drop waren
3. ass wuel net esou einfach, en dialekt forcement zu enger sprooch wëllen ze forméieren...

examples where this label is not applicable:

1. ganz richtig
2. gudd geschriwen
3. bai allem respekt awer dier sidd mengen ech am joer 1945 henken bliwen.

examples where none of the labels described are applicable:

- Ech kennen 2 lëtzebuergesch Mammen déi aleng mat hire Kanner do stin, déi kruten null Ennerstëtzung fir eng Wunneng ze fannen.
- Letzebuerg huet keen mobiliteitsproblem ausser moies an owes zu den Spetzenstonnen.
- Kommt mir lossen eis net mei manipuleiren!
- Do huet de Fotograf awer gutt oppgepass fir emmer Leit opt Bild ze krein ...
- gott sein dank dass mir keng vum "adr" an der spëtz hun soss hätten mir awer wirklech "FOLKLOR"
- Am EU parlament gouf gestemmt: 39 Jo- 478 neen-165 enthalung-72 abwesend nozekucken ob dem eu parlameter vun zdf. do kann een och kucken wien wei ofgestemmt hued.Letzebuergesch Parlamentarier hun all neen gestemmt ausser Mme Lulling.

Appendix D

In this appendix, we add further examples of annotations mislabeled by all the models per category. The example is first reported in Luxembourgish followed by the English translation, its gold standard, and the mislabeled annotation by

the models and model generated explanations.

Example 586

Mir sinn schon dat eenzegt Land wou d’Kanner mussen an enger anerer Sprooch alphabetiséiert ginn wéi hier Mammensprooch, fir den net franco-phonen=Lëtzebuenger et lo nach méi schwéier ze machen, sollen d’Kanner lo och nach op franséisch alphabetiséiert ginn??

We are already the only country where the children have to be alphabetized in a language other than their mother tongue, to make it even more difficult for non-francophone=Luxembourgers, should the children be alphabetized in French??

Gold standard: RESPONSIBILITY

Magistral : BELONGING

‘The sentence discusses the integration of children from different linguistic backgrounds into Luxembourgish society, focusing on the role of Luxembourgish language in this process. It involves the presence of foreigners and the expectation to learn Luxembourgish.’

DeepSeek: VITALITY

‘The sentence emphasizes the need to use Luxembourgish words instead of German or French words, which aligns with the ‘vitality’ category as it addresses the preservation of Luxembourgish language.’

Qwen: VITALITY

‘The sentence expresses concern that children in Luxembourg are being alphabetized in a language other than their mother tongue (Luxembourgish), and further that non-French-speaking Luxembourgish children are being forced to learn in French, which is portrayed as making the situation even harder. This reflects a perception of Luxembourgish being threatened by the dominance of other languages in education — a core concern of the ‘vitality’ category, as it highlights endangerment through institutional language use.’

GPT-5: RECOGNITION

‘The sentence critiques alphabetizing children in a language different from their mother tongue and questions doing it in French, framing Luxembourgers as non-Francophone; this is an opinion about which languages should be used in official/educational contexts and the dominance of French.’

Example 339

Si sollen mol zu alleréischt all Lëtzebuergesch léieren.

First of all, they should all learn Luxembourgish.

Gold standard: BELONGING

Magistral : IDENTITY

‘The sentence emphasizes the importance of learning Luxembourgish first, which is a personal identity trait.’

DeepSeek: RESPONSIBILITY

‘The sentence emphasizes the responsibility of

teaching Luxembourgish first, highlighting the importance of prioritizing Luxembourgish language education.'

Qwen: IDENTITY

'The sentence emphasizes that children should learn Luxembourgish 'at first' ('zu alleréischt'), framing it as the foundational and primary language of the national group. The use of 'si sollen' (they should) implies a collective norm tied to national identity Luxembourgish as the essential, first language of Luxembourgers, reinforcing the tied connection between national identity and national language.'

GPT-5: VITALITY

'Advocates that all children first learn Luxembourgish, emphasizing increased acquisition and priority of Luxembourgish as a way to strengthen and preserve the language.'

Example 356

Jo mir hun eng Nationalsprooch, an déi solle mir och héich halen, mee ons Stäerkt war ëmmer, dass mir dräisproocheg waren, an och houfreg drop waren.

Yes, we have a national language, and we should hold it in high esteem, but our strength has always been that we were trilingual, and we were also proud of it.

Gold standard: IDENTITY

Magistral : RECOGNITION

'The sentence acknowledges Luxembourgish as a national language and discusses the strength of being multilingual, which falls under the recognition category.'

DeepSeek: RECOGNITION

'The sentence acknowledges Luxembourgish as the national language and emphasizes the importance of maintaining it, fitting the 'Recognition' category.'

Qwen: RECOGNITION

'The sentence acknowledges Luxembourgish as a 'Nationalsprooch' that should be 'héich gehalten' (held high), while also recognizing the value of trilingualism ('dräisproocheg'). This reflects a hierarchical acknowledgment of languages — affirming Luxembourgish as official and worthy of prestige, while accepting coexistence with other languages. This fits the 'recognition' category, which deals with status, standardization, and hierarchy of languages.'

GPT-5: RECOGNITION

'It explicitly discusses the status of Luxembourgish as a national language and values Luxembourg's trilingualism, i.e., views on languages and their place in the country.'

Example 678

Gleichzaiteg gehéiert Lëtzebuergesch zu onser Kultur a muss eis erhaale bleiwen!

GLikewise Luxembourgish belongs to our culture and must be preserved!

Gold standard: VITALITY

Magistral : IDENTITY

'The sentence emphasizes the importance of Luxembourgish to the national culture and identity, which is characteristic of the 'Identity' category.'

DeepSeek: IDENTITY

'The sentence emphasizes the importance of Luxembourgish to Luxembourg's culture, fitting the 'Identity' category as it highlights a connection between language and national identity.'

Qwen: IDENTITY

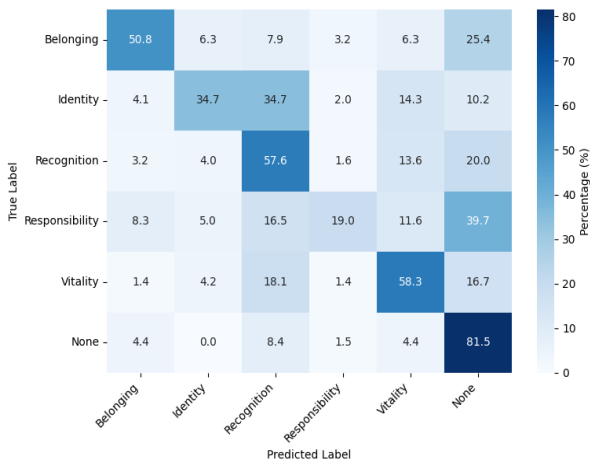
'The sentence asserts that Luxembourgish belongs to 'our' culture and must be preserved, using the first-person plural 'eis' (our) to express collective emotional attachment to the language as a core element of national identity. This reflects the tied connection between national identity and national language, fitting the 'identity' category.'

GPT-5: IDENTITY

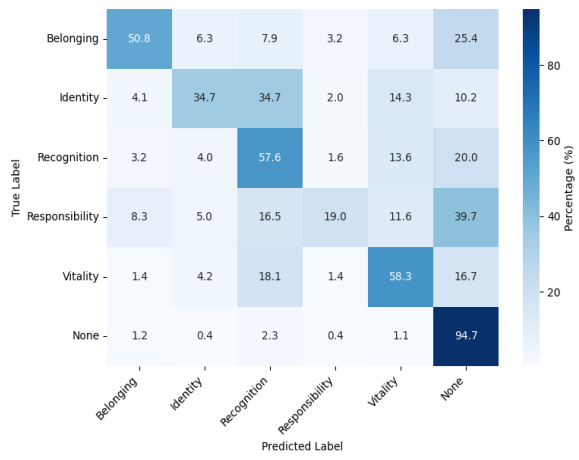
'It explicitly links Luxembourgish to 'our culture' and uses first-person plural to assert a cultural-national attachment, a hallmark of the identity category.'

Appendix E

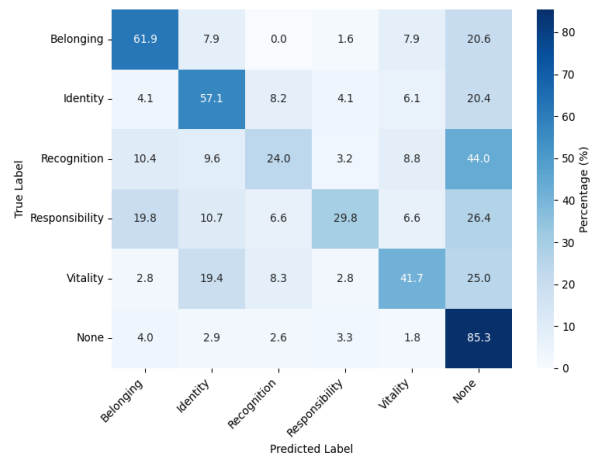
In this appendix, we add row normalized confusion matrices for the Lang subset and for the whole dataset in analysis (Lang and NotLang) in its original language (Luxembourgish) for the four analyzed models: GPT-5, DeepSeek, Magistral, and Qwen.



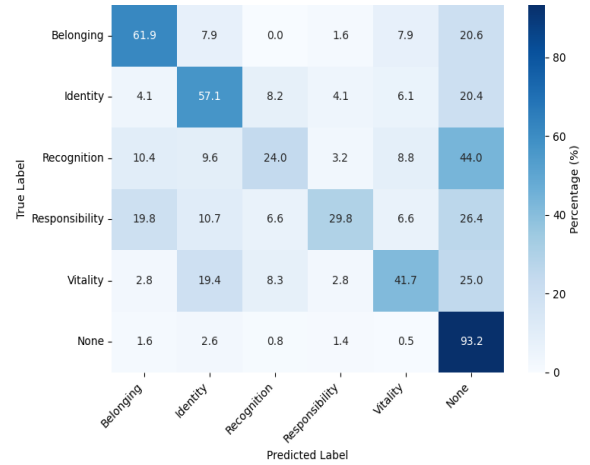
(a) Lang subset



(b) Lang and NotLang subsets



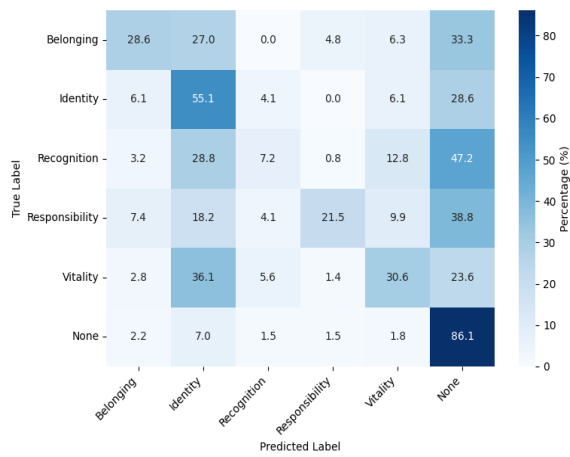
(a) Lang subset



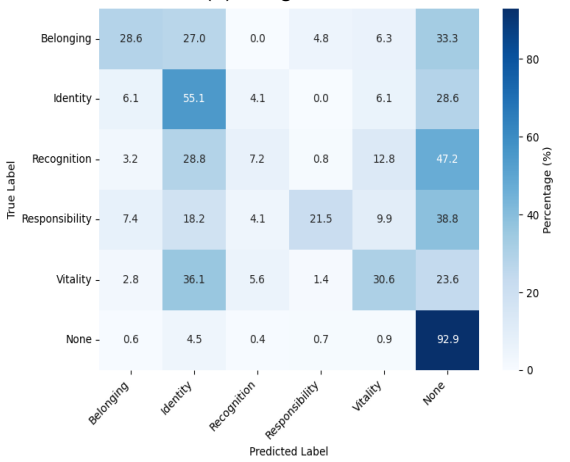
(b) Lang and NotLang subsets

Figure 3: Confusion matrix per annotation with GPT-5

Figure 4: Confusion matrix per annotation with DeepSeek



(a) Lang subset

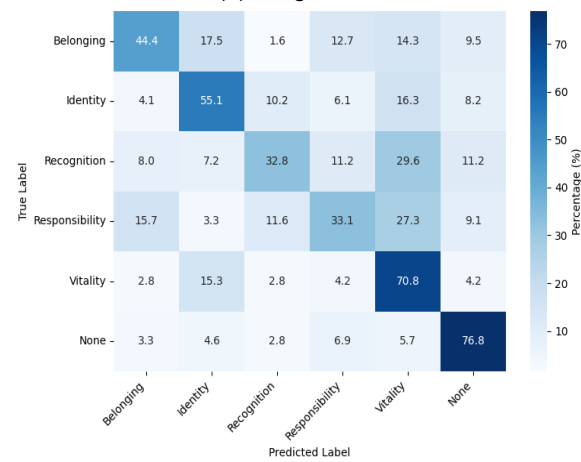


(b) Lang and NotLang subsets

Figure 5: Confusion matrix per annotation with Magistral



(a) Lang subset



(b) Lang and NotLang subsets

Figure 6: Confusion matrix per annotation with Qwen

Personality Anchoring for Social Simulation: Linking Personality, Social Behavior, and Interaction Success with LLM Agents

Vahid Sadiri Javadi [♣], Aksa Aksa [♣], Fryderyk Róg [♣]
Lucie Flek [♣], Johanne R. Trippas [♣]

University of Bonn, Conversational AI and Social Analytics (CAISA) Lab, Bonn, Germany [♣]
RMIT University, School of Computing Technologies, Melbourne, Australia [♣]

Abstract

Social interactions are shaped by the interplay of dispositional traits and situational context, yet systematically investigating how personality configurations between individuals jointly influence social behavior across diverse social contexts remains methodologically challenging. We address this gap by introducing a simulation pipeline adapted from the CHARISMA framework, which employs well-known movie characters and public figures as psychologically grounded agents for multi-LLM social simulation using a method we term *personality anchoring*. We present a large-scale empirical study examining how dyadic Agreeableness composition influences social interaction outcomes across 1,010 simulated conversations. Our results reveal a monotonic relationship between dyadic Agreeableness composition and shared goal achievement, with Homogeneous-Agreeable pairs achieving success 10 times the rate of Homogeneous-Disagreeable pairs (62% vs. 6%). Behavioral mediation analysis reveals that Agreeableness shapes goal achievement partially through cooperative strategy selection, though it continues to predict outcomes within the same dominant strategy, indicating pathways beyond observable conversational behavior. Robustness analyses confirm high consistency of results across repeated simulations (ICC = 0.89) and stable personality expression across diverse scenarios, validating personality anchoring as a viable operationalization strategy.

Keywords: Simulation, Social Psychology, Large Language Models

1. Introduction

Understanding how dispositional traits and situational context jointly shape social interaction outcomes is central to social psychology. Attribution theory provides a foundational framework for this inquiry, explaining how individuals infer the causes of behavior by distinguishing between dispositional and situational factors (Heider, 1958; Kelley, 1967; Weiner, 1986). Among dispositional factors, individual differences are commonly operationalized through the Big Five personality framework, within which Agreeableness has been identified as the dimension most closely tied to interpersonal conflict processes and outcomes (Jensen-Campbell and Graziano, 2001), with highly agreeable individuals preferring negotiation and compromise, while those low in Agreeableness tend toward competitive or coercive strategies (Graziano et al., 1996; Wilmot and Ones, 2022). Yet investigating how different personality configurations between interacting individuals jointly shape social outcomes across diverse contexts remains methodologically challenging. Traditional experimental methods, while yielding important insights (Aronson et al., 1990), face limitations in scalability, reproducibility, and the systematic manipulation of complex social variables (Nosek et al., 2022; Open Science Collaboration, 2015; Wicherts et al., 2016).

Recent advances in large language models (LLMs) have created new opportunities for com-

putational social psychology by enabling the simulation of open-ended social interactions at unprecedented scale (Park et al., 2023; Zhou et al., 2024). Prior work has demonstrated that LLMs can effectively simulate Big Five personality traits with behaviors that human observers rate as believable (Jiang et al., 2024; Serapio-García et al., 2025; Javadi et al., 2025), and frameworks such as SOTOPIA (Zhou et al., 2024) have enabled systematic evaluation of social intelligence in LLM-based agents. Studies on personality-conditioned agents have further explored how traits influence cooperation in games (Qiu, 2025) and social media behavior in large-scale simulations (Yang et al., 2024b).

However, several important gaps remain. First, most existing simulations examine personality within relatively narrow settings, such as bargaining tasks or cooperative games, without grounding scenarios in a validated taxonomy of everyday human goals. Second, personality is typically introduced through explicit trait prompting (e.g., “you are highly agreeable”), which frames personality as an instruction rather than a naturally occurring behavioral tendency. Third, there is limited focus on *dyadic personality composition*, i.e., how different configurations of personality traits between two interacting individuals shape joint outcomes across varied social contexts. Fourth, existing evaluations tend to emphasize overall task outcomes or persona fidelity, often overlooking the *behavior strategies* through which personality influences so-

cial outcomes, i.e., the conversational mechanisms that mediate the personality–outcome relationship.

In this paper, we address these gaps by adapting the CHARISMA framework (Sadiri Javadi et al., 2026), which employs well-known movie characters and public figures as psychologically grounded agents for multi-LLM social simulation. Rather than assigning abstract trait scores, CHARISMA leverages LLMs’ embedded knowledge of characters’ backstories and behavioral tendencies to simulate personality-consistent behavior. We term this approach *personality anchoring*. We present a comprehensive empirical study examining how the dispositional trait Agreeableness influences social interaction outcomes. Agreeableness is operationalized through systematic dyadic pairings of characters with crowd-sourced Big Five profiles from the Personality Database (PDB)¹ across large-scale simulated conversations spanning seven social goal categories derived from a taxonomy of 135 human goals (Chulef et al., 2001). Our paper makes the following contributions:

1. We introduce a personality-driven simulation methodology that integrates personality through character-based anchoring, a structured taxonomy of human goals, and behavior strategies in conversational interaction, enabling systematic analysis of how dispositional traits and situational factors jointly shape social behavior in simulated interactions.
2. We conduct a large-scale empirical analysis of how dyadic Agreeableness composition shapes social behavior across seven social goal categories, two difficulty levels, and multiple interaction models.
3. We provide a behavioral mediation analysis examining whether and how conversational strategies mediate the relationship between personality composition and interaction outcomes, distinguishing between direct and indirect pathways of personality influence.
4. We evaluate robustness along two dimensions: (i) result consistency across repeated simulations and (ii) personality expression stability across diverse scenarios, assessing the reliability of character-based anchoring as a personality operationalization strategy.

The code, dataset, full list of behavior strategies and characters, and behavioral analysis scripts are publicly available.²

¹<https://www.personality-database.com/>

²<https://github.com/vahidsj/PersonalityAnchoring>

2. Related Work

2.1. LLM-Based Social Simulation

LLM-powered social simulation has scaled rapidly since the introduction of Generative Agents (Park et al., 2023), which demonstrated that 25 LLM agents could sustain coherent social behavior, including relationship formation and activity coordination, over multiple simulated days using memory, reflection, and planning components. SOTOPIA (Zhou et al., 2024) shifted focus to systematic evaluation, introducing 90 social scenarios and a 7-dimensional evaluation framework assessing goal completion, relationship maintenance, and social norm adherence, with GPT-4 as an LLM-based evaluator. Follow-up work has extended this ecosystem: SOTOPIA- Ω (Zhang et al., 2025) injects negotiation strategies enabling 7B models to surpass GPT-4 on social goals, while Sotopia-RL (Yu et al., 2025) introduces utterance-level multi-dimensional rewards for training socially intelligent agents. At larger scales, OASIS (Yang et al., 2024b) simulates up to one million agents on social media platforms, replicating information spreading and group polarization dynamics. AgentSociety (Piao et al., 2025) integrates Maslow’s hierarchy of needs and the Theory of Planned Behavior into 10,000+ agents, successfully reproducing real-world social experiments including polarization dynamics and universal basic income effects. GenSim (Tang et al., 2025) provides a general-purpose platform supporting 100K+ agents with error-correction mechanisms.

Alongside simulation environments, role-playing language agents have been extensively studied. RoleLLM (Wang et al., 2024a) benchmarks persona consistency across 100 roles, SimsChat (Yang et al., 2024a) generates multi-turn dialogues for 68 characters defined by traits and aspirations, and SocialBench (Chen et al., 2024) evaluates agents at both individual and group levels, finding that individual-level proficiency does not imply group-level competence. These systems demonstrate that LLM agents can participate in coherent social interactions, but most do not systematically ground scenarios in validated psychological taxonomies or examine how personality configurations between interacting agents shape joint outcomes.

2.2. Personality Expression and Operationalization in LLMs

Research on personality in LLMs has progressed along three methodological lines: prompting, training, and activation steering. The prompting approach is most established. PersonaLLM (Jiang et al., 2024) assigned Big Five configurations to 320 personas and found large effect sizes in self-

reported BFI scores, with human evaluators identifying traits at up to 80% accuracy. Serapio-García et al. (Serapio-García et al., 2025) tested 18 LLMs with psychometric instruments (IPIP-NEO, BFI), demonstrating that personality can be reliably measured and shaped under specific prompting configurations. Additional evidence shows that LLMs form stable, interpretable Big Five patterns across repeated trials (Sorokovikova et al., 2024).

More recent work has moved beyond prompting. BIG5-CHAT (Li et al., 2025) uses supervised fine-tuning and DPO on a 100K-dialogue dataset grounded in real human personality expressions, outperforming prompt-based methods on psychometric measures. Activation-steering approaches use representation engineering to directly manipulate personality-related internal representations (Ong et al., 2025), finding that higher Agreeableness improves cooperation but increases exploitation vulnerability. However, psychometric measurement remains challenging: PERSIST (Tosato et al., 2026) tests 25 models across 2M+ responses and finds that even 400B+ parameter models show substantial measurement instability under question re-ordering. Most studies use explicit trait prompting (e.g., “you are highly agreeable”) (Jiang et al., 2024; Serapio-García et al., 2025; Sorokovikova et al., 2024), which frames personality as an instruction rather than a naturally occurring behavioral tendency. Character-based approaches such as In-Character (Wang et al., 2024b) uses psychological interviews of 32 fictional characters and achieves 80.7% personality alignment with human-perceived types from the Personality Database. Moon (Moon, 2025) develops narrative backstory conditioning that reproduces population-level cooperative behaviors in social dilemmas without explicit trait labels. Our work follows this character-based line, leveraging LLMs’ embedded knowledge of well-known movie characters and public figures’ behavioral tendencies rather than explicit trait descriptors.

When personality-conditioned agents interact in social tasks, studies show trait effects on cooperative behavior. Huang and Hadfi (Huang and Hadfi, 2024) find that Big Five profiles influence negotiation outcomes and strategy use. NetworkGames (Qiu, 2025) assigns MBTI types to agents in iterated Prisoner’s Dilemma on network topologies, showing that macro-level cooperation depends on both dyadic personality pairings and network structure. Zeng et al. (Zeng et al., 2025) model dynamic personality evolution across evolutionary generations. However, most of these studies examine personality within narrow task domains (e.g., cooperative games) and focus on individual trait expression rather than systematic dyadic composition across diverse social contexts, which is the central focus of our work.

2.3. Agreeableness and Interpersonal Conflict

Among the Big Five dimensions, agreeableness has the strongest theoretical and empirical connection to interpersonal conflict and cooperation. Graziano et al. (Graziano et al., 1996) demonstrated through multi-method designs that low-agreeableness individuals rate power assertion significantly more favorably during conflict than their high-agreeableness counterparts. Jensen-Campbell and Graziano (Jensen-Campbell and Graziano, 2001) established through diary studies that agreeableness is the Big Five dimension most closely associated with conflict processes and outcomes, with low-agreeableness individuals using more destructive tactics that predict poorer adjustment. The most comprehensive quantitative review to date, by Wilmot and Ones (Wilmot and Ones, 2022), synthesizes 142 meta-analyses across 275 variables, confirming that agreeableness produces desirable effects for 93% of variables examined. Thielmann et al. (Thielmann et al., 2020) provide a complementary theoretical framework identifying situational affordances that moderate personality-prosociality links across economic games.

Recent computational studies converge with these psychological findings. Sakai et al. (Sakai et al., 2025) test personality steering in repeated Prisoner’s Dilemma and find agreeableness is the dominant factor promoting cooperation across multiple LLM models. Noh and Chang (Noh and Chang, 2024) report across 1,500 multi-issue negotiation simulations that agreeableness is the most important personality trait for negotiation outcomes.

Our work extends this body of research in three ways. First, we examine the agreeableness effects across diverse social goal categories rather than in a single task domain. Second, we operationalize personality through character-based anchoring rather than explicit trait prompting. Third, we analyze the *behavior strategies* through which agreeableness influences outcomes, i.e., the conversational mechanisms that mediate the personality-outcome relationship.

3. Methodology

We introduce a simulation pipeline adapted from the CHARISMA framework (Sadiri Javadi et al., 2026) for a large-scale empirical study of how dispositional traits and situational factors jointly shape social behavior in social interactions. As shown in Figure 1, the simulation pipeline consists of five stages: (1) social scenario setup, (2) character pairing curation, (3) scenario generation and curation, (4) interaction generation with behavior strategy, and (5) simulation evaluation.

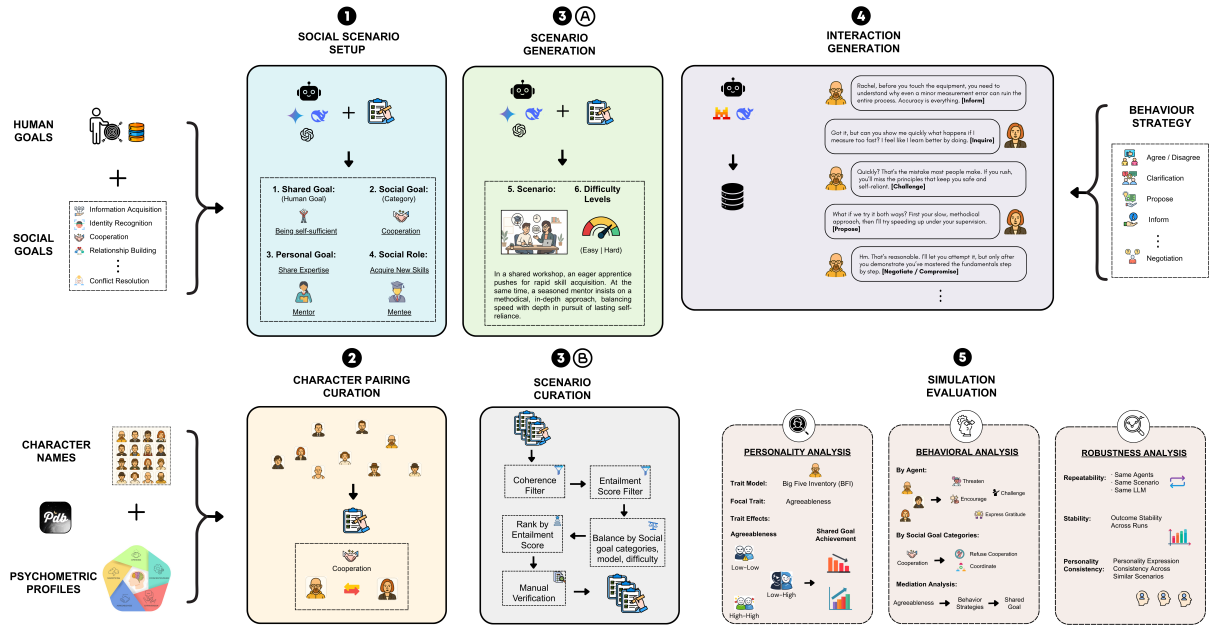


Figure 1: Overview of the simulation pipeline adapted from CHARISMA. **Stage 1:** Social scenario setup derives shared goals, social goal categories, personal goals, and social roles from a 135-goal taxonomy (See Section 3.1). **Stage 2:** Characters with crowd-sourced Big Five profiles from the Personality Database are paired into four Agreeableness conditions (See Section 3.2). **Stage 3:** Scenarios are generated at two difficulty levels and curated through coherence filtering, entailment scoring, balanced selection, and manual verification (See Section 3.3). **Stage 4:** Two LLM agents engage in 20-turn dialogues where each turn involves behavior strategy selection, personality reasoning, and response generation (See Section 3.4). **Stage 5:** Evaluation covers personality–outcome analysis, behavioral mediation, and robustness assessment (See Section 3.5).

3.1. Social Scenario Setup

Social scenarios are grounded in a validated goal-driven structure. Instead of generating scenarios ad hoc, the framework adopts the 135 Human Goals Taxonomy (Chulef et al., 2001), a systematically organized hierarchy derived from extensive empirical research. Each goal serves as a seed for scenario generation, providing the shared objective around which agent interactions are structured. To bridge abstract human goals and concrete social interaction patterns, each goal is classified into one of eight social goal categories: *Cooperation*, *Conflict Resolution*, *Relationship Building*, *Relationship Maintenance*, *Identity Recognition*, *Information Acquisition*, *Information Provision*, and *Competition*. These categories are informed by classical social interaction typologies (Nisbet, 1970) and research on interpersonal goals (Clark and Mills, 1979). For each scenario, the framework specifies four components: (1) a **shared goal** drawn from the taxonomy, (2) a **social goal category** classifying the interaction type, (3) **personal goals** for each agent that may complement or conflict with the shared goal, and (4) **social roles** defining relational positions. This multi-layered goal structure creates conditions for rich interactions by introducing both alignment and potential tension between agents.

3.2. Character Pairing Curation

We operationalize personality through *personality anchoring*, rather than assigning abstract trait scores or relying on explicit trait prompting. This approach leverages LLMs’ embedded knowledge of well-known movie characters and public figures to elicit personality-consistent behavior. Characters are sourced from the Personality Database (PDB), a large-scale crowd-sourced platform where users vote on personality traits using multiple frameworks, including the Big Five Inventory (BFI).

Characters undergo a multi-stage filtering process. First, a **vote-threshold filter** ensures a sufficient number of independent ratings. Second, an **inter-rater agreement filter** retains only characters with consistent BFI assessments across raters. Third, characters are ranked using an **Agreeableness score** that quantifies voting support for the assigned Agreeableness level:

$$\text{Rank} = \frac{c_{\text{main}}}{c_{\text{main}} + c_{\text{other}}} \quad (1)$$

where c_{main} is the vote count for the character’s assigned Agreeableness value and c_{other} is the vote count for the closest alternative. We select five characters at each of four Agreeableness levels $\{0.0, 0.25, 0.75, 1.0\}$, yielding 20 characters in total

(See Appendix A.1 for the full list). The neutral level {0.5} is excluded because it lacks distinctive behavioral characteristics.

Condition	Abbrev.	Agent A & B
Homogeneous-Disagreeable	HoD	{0.0, 0.25}
Heterogeneous-Extreme	HeE	{0.0, 1.0}
Heterogeneous-Moderate	HeM	{0.25, 0.75}
Homogeneous-Agreeable	HoA	{0.75, 1.0}

Table 1: Dyadic personality composition conditions based on Agreeableness levels.

As shown in Table 1, characters are paired into four conditions that vary in *dyadic personality composition*. **Homogeneous-Disagreeable (HoD)** pairs two low-Agreeableness agents, whereas **Homogeneous-Agreeable (HoA)** pairs two high-Agreeableness agents. **Heterogeneous-Extreme (HeE)** represents the maximum contrast between low and high Agreeableness, and **Heterogeneous-Moderate (HeM)** introduces a moderate contrast between the two agents. Within each condition, every agent interacts with every agent from the corresponding level ($5 \times 5 = 25$ pairs), yielding 100 unique dyads across all four conditions.

3.3. Scenario Generation and Curation

Each scenario setup is expanded into a detailed narrative description at two difficulty levels: *Easy* (straightforward dynamics) and *Hard* (high tension), enabling analysis of how personality effects vary under different situational demands. Three LLMs (*DeepSeek-Chat-v3-0324*, *Google Gemini 3 Flash*, and *OpenAI GPT-5.2*) generate scenario setups from the 135 human goals, with a model-consistency approach ensuring that the same model expands its own setups into full scenarios. Generated scenarios undergo a multi-stage curation pipeline. A **coherence filter** (threshold ≥ 0.8), based on LLM-as-a-judge evaluation (Liu et al., 2023), retains only logically consistent scenarios. An **entailment filter** (threshold ≥ 0.6) uses a pre-trained NLI model (Lewis et al., 2020) to verify alignment between scenario specifications and generated content. A **balanced selection** phase caps scenarios per social goal category while ensuring representation across models and difficulty levels. Finally, a **manual verification** audit on a stratified 12.6% subset confirms quality along 5 dimensions: goal clarity, role plausibility, social realism, difficulty alignment, and conversational achievability.

The final curated dataset comprises 277 high-quality scenarios. Of the original 8 social goal categories, *Information Provision* was excluded due to insufficient post-filtering representation, leaving 7 categories that are approximately balanced across models and difficulty levels.

3.4. Interaction Generation

Two LLM-based agents interact in maximum 20-turn dialogues. Consistent with the personality anchoring approach, each agent is instantiated with its character identity, relying on the LLM’s internal knowledge of the character rather than providing explicit personality information, along with the scenario context, including its assigned role, personal goal, and shared goal.

A central feature of the interaction protocol is the integration of **behavior strategy** into the generation process. Rather than generating free-form utterances, each agent follows a structured turn-taking sequence:

1. **Behavior strategy selection:** the agent selects a communicative intent label (e.g., *Propose*, *Challenge*, *Encourage*) from a coding scheme organized into category-specific and universal codes (See Appendix A.2 for the full list). It can also select *None* if no code fits the response.
2. **Personality reasoning:** the agent reasons about how its personality traits should influence the response.
3. **Response generation:** guided by the selected code and personality reasoning, the agent produces a natural-language utterance.
4. **Trait score reporting:** the agent reports numerical BFI scores reflecting trait levels expressed in the current turn.

Individual behavior strategies are aggregated into three higher-order **behavior strategy groups**: *Cooperative* (e.g., *Encourage*, *Express Gratitude*, *Build Consensus*), *Confrontational* (e.g., *Challenge*, *Dismiss*, *Taunt*, *Threaten*), and *Neutral* (e.g., *Inquire*, *Clarify*, *Inform*). This aggregation enables analysis of how Agreeableness configurations relate to conversational strategy selection and, subsequently, to interaction outcomes. This provides the analytical basis for the behavioral mediation analysis described in Section 4.2.

3.5. Simulation Evaluation

Evaluation covers three complementary dimensions, corresponding to our experimental research questions.

Personality–Goal Achievement analysis examines how shared goal achievement scores vary across the four Agreeableness pairing conditions. Goal achievement is assessed using an LLM-as-a-judge approach: the evaluator model receives the complete interaction transcript, scenario specification, and scoring rubric, then assigns scores on a 0–10 scale for both shared and personal goal

achievement with accompanying reasoning and confidence assessments.

Behavioral mediation analysis examines how conversational strategy distributions differ across Agreeableness conditions and social goal categories, and whether conversational strategies mediate the relationship between dyadic Agreeableness composition and goal achievement. This analysis operates at multiple levels: pairing-condition aggregates, and mediation pathways linking personality → conversational strategies → shared goal achievement.

Robustness analysis assesses two forms of reliability. *Results consistency* is measured by repeating simulations under identical conditions (same agent pair, scenario, and LLM) across multiple runs and computing intraclass correlation coefficients (ICC). *Personality expression consistency* evaluates whether the same character exhibits stable Agreeableness expression across different scenarios within the same social goal category, testing a core assumption of personality anchoring: that LLM agents can embody stable personality profiles through character knowledge alone.

4. Experiments and Results

We conduct four experiments across 1,010 conversations to examine how dyadic Agreeableness composition shapes social interaction outcomes, the behavior strategies underlying this relationship, and the robustness of both results (i.e., shared goal achievement) and personality expression. Table 2 summarizes the experimental design.

Research Question	# Conv.	Focus
1. Personality → GA	400	Direct effects
2. Personality → BS → GA	400	Mediation
3. Result Consistency	250	Robustness
4. Personality Expression	360	Trait stability

Table 2: Overview of the experimental design, including the number of conversations and the analytical focus for each RQ. GA = Goal Achievement; BS = Behavior Strategy. Experiments 1 and 2 are conducted on the same conversation dataset.

Shared Configuration. All experiments build on the curated scenario database of 277 scenarios spanning 7 social goal categories. The primary interaction model is *DeepSeek-Chat-v3-0324*, with *Mistral Large* as a cross-model replication. Each conversation comprises 20 turns (10 per agent). Evaluation uses *DeepSeek-Chat-v3-0324* as an LLM-as-a-judge, scoring shared and personal goal achievement on a 0–10 scale with accompanying reasoning and confidence assessments.

4.1. Experiment 1: Personality and Goal Achievement

Design. We generate 400 conversations distributed evenly across the four Agreeableness pairing conditions (100 per condition). Each of the 100 unique agent pairs participates in 4 conversations, each assigned to a distinct social goal category via constrained randomization ensuring balance across the 7 categories (~57 conversations per category).

Results. Table 3 presents the primary outcome measure: mean shared goal achievement by pairing condition. Shared goal achievement increases monotonically from HoD to HoA, with a 5-point spread on the 10-point scale.

Pair Type	Agreeableness	Mean	Success@8
HoD	0.0 – 0.25	2.3	6%
HeE	0.0 – 1.0	3.7	11%
HeM	0.25 – 0.75	5.6	38%
HoA	0.75 – 1.0	7.3	62%

Table 3: Mean shared goal achievement (0–10) and strong success rate (score ≥ 8) by Agreeableness pair type.

Applying a threshold of shared goal score ≥ 8 (“strong success”), HoA pairs achieve strong success in 62% of interactions compared to only 6% for HoD pairs, showing a significant difference. Mixed pairs show intermediate success rates but remain closer to HoD than HoA, indicating that consistently high-level success is primarily associated with mutually high Agreeableness rather than the presence of a single agreeable agent.

As shown in Figure 2, the Agreeableness effect holds across all seven social goal categories. Categories such as Relationship Maintenance, Identity Recognition, and Cooperation show the strongest HoD–HoA contrast, while Competition shows the smallest effect (HoD: 1.5; HoA: 4.4), suggesting that competitive scenarios pose structural challenges that high Agreeableness alone cannot fully overcome. The effect holds across both difficulty levels (Easy: HoD 2.2, HoA 7.5; Hard: HoD 2.4, HoA 6.9) and all three scenario-generating models (DeepSeek: HoD 2.1, HoA 7.5; Gemini: HoD 2.1, HoA 6.2; OpenAI: HoD 2.8, HoA 8.0), confirming that the observed pattern is not an artifact of specific scenario sources or ceiling effects.

Cross-model replication. Replicating with Mistral as the interaction backbone preserves the same monotonic ordering and comparable success rates (HoD: 4%; HoA: 61%), confirming that the personality–outcome relationship generalizes across interaction models.

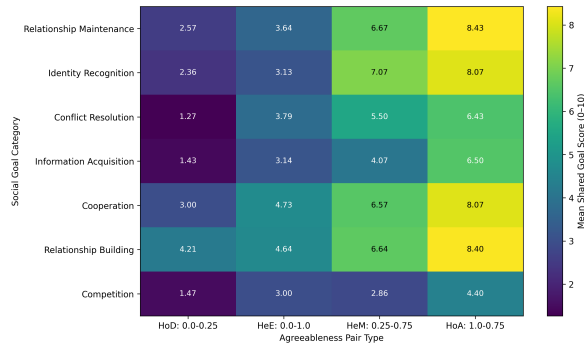


Figure 2: Mean shared goal achievement by social goal category \times Agreeableness pair type. The HoD < HoA contrast is preserved across all categories, with the largest effects in relationally oriented categories and the smallest in Competition.

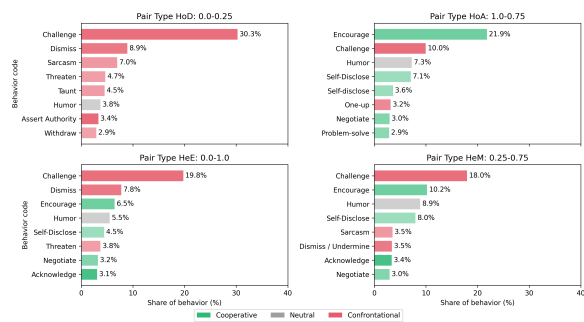


Figure 3: Top behavior strategies by Agreeableness pair type. HoD pairs are dominated by confrontational behaviors (Challenge, Dismiss), while HoA pairs favor cooperative strategies (Encourage, Express Gratitude, Build Consensus).

4.2. Experiment 2: Behavioral Analysis

Design. Using the same 400 conversations from Experiment 1, we analyze behavior code distributions across pairing conditions and examine whether conversational strategies mediate the Agreeableness–outcome relationship.

Behavioral profiles by pair type. As illustrated in Figure 3, Agreeableness configuration produces distinct behavioral signatures. HoD pairs are dominated by confrontational behaviors: *Challenge* (17.3%), *Dismiss*, *Taunt*, and *Withdraw* characterize the interaction style. HoA pairs exhibit a predominantly prosocial profile: *Humor* (24.2%) and *Encourage* (17.3%) together account for over 40% of behaviors, supplemented by *Express Gratitude*, *Build Consensus*, and *Self-Disclose*. Notably, *Encourage* is virtually absent in HoD pairs, while *Challenge* drops from 17.3% (HoD) to 4.9% (HoA). Mixed pairs exhibit intermediate profiles reflecting both orientations.

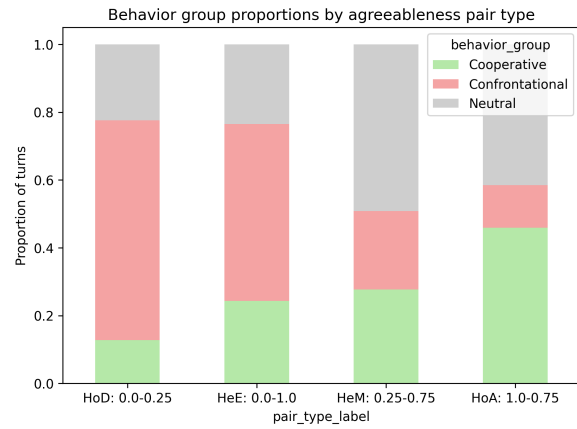


Figure 4: Behavior group proportions by Agreeableness pair type. HoD pairs employ predominantly Confrontational strategies, while HoA pairs favor Cooperative and Neutral behaviors.

Behavior strategy groups and goal achievement. As shown in Figure 4, when individual behavior strategies are aggregated into Cooperative, Confrontational, and Neutral strategy groups, the mediation pathway becomes clear. HoD pairs employ predominantly Confrontational strategies (~55% of turns) with minimal Cooperative behavior (~20%), while HoA pairs show the reverse pattern. Conversations dominated by Cooperative strategies achieve substantially higher shared goal scores (mean = 7.1) compared to Neutral (mean = 5.3) or Confrontational-dominant conversations (mean = 2.5).

Partial mediation. As shown in Table 4, a joint analysis crossing dominant behavior strategy with Agreeableness pair type reveals that Agreeableness continues to predict outcomes *within* the same dominant strategy. When both HoD and HoA pairs employ Cooperative strategies, HoA pairs still outperform HoD (8.1 vs. 5.7). Similarly, when both employ Confrontational strategies, HoA pairs achieve higher scores than HoD (3.8 vs. 1.6). This pattern indicates *partial* rather than full mediation: Agreeableness shapes outcomes both by increasing the likelihood of cooperative behavior *and* through additional pathways beyond strategy selection.

Pair Type	Dominant Behavior Strategy		
	Cooperative	Neutral	Confrontational
HoD	5.7	3.3	1.6
HeE	6.5	3.8	1.8
HeM	7.0	5.2	2.3
HoA	8.1	6.2	3.8

Table 4: Mean shared goal achievement by strategy \times pair type. Agreeableness predicts outcomes within strategy groups.

4.3. Experiment 3: Result Robustness

Design. Fifty conversation configurations from Experiment 1 are each repeated 5 times under identical conditions (same agent pair, scenario, and model), yielding 250 conversations. Only the stochastic sampling inherent in LLM generation varies across repetitions.

Results. The pooled standard deviation across repeated runs is 0.98 for shared goal achievement on the 0–10 scale, which is modest relative to the 5-point spread between HoD and HoA means. Single-run intraclass correlation coefficients ($ICC_{3,1}$) reach 0.89 for shared goal achievement, indicating good-to-excellent reliability (Koo and Li, 2016). When averaged over 5 runs, reliability increases to $ICC_{3,k} = 0.97$. Approximately 94–96% of configurations exhibit variance below 3.0, confirming that the consistency of results is broadly uniform rather than driven by a subset of stable cases.

4.4. Experiment 4: Personality Expression Stability

Design. Each of the 20 agents interacts with 3 partners across 6 scenarios within each of two social goal categories (Cooperation and Conflict Resolution), yielding 360 conversations. A partner balance rule ensures each agent is exposed to both low- and high-Agreeableness partners.

Results. As shown in Figure 5, agents maintain consistent Agreeableness expression across scenarios within each category (Figure 5). Characters at the extremes, such as Logan Roy (expected: 0.0) and Anne Shirley-Cuthbert (expected: 1.0), show tight clustering of expressed values across all scenarios.

More moderate characters show wider but still bounded variability. Critically, the categorical distinction between low and high Agreeableness is preserved across all agents: those expected to be low consistently express values below the 0.5 midpoint, while those expected to be high consistently express values above it.

Expression patterns are similar across the Cooperation and Conflict Resolution categories, suggesting that consistency is a property of the character rather than the situational context. These findings validate personality anchoring as a viable operationalization strategy: without receiving explicit personality scores, LLM agents embody characters in ways that reflect expected trait levels based on the model’s internal knowledge.

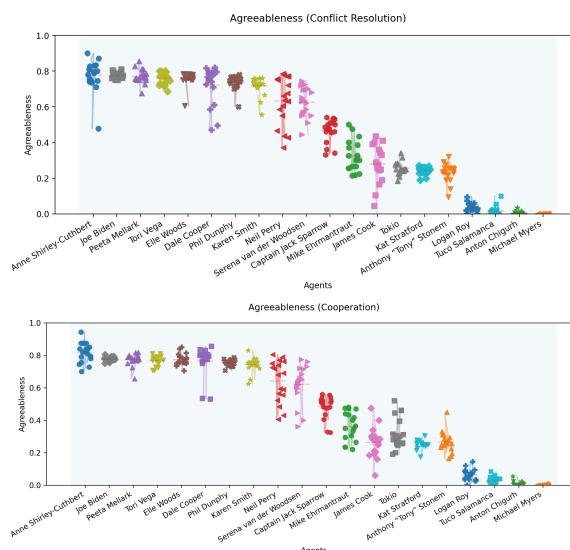


Figure 5: Expressed Agreeableness per agent across scenarios for Conflict Resolution (top) and Cooperation (bottom). Each point represents one conversation. Characters at the extremes show tight clustering; the categorical distinction between low (<0.5) and high (>0.5) Agreeableness is preserved across all agents.

5. Conclusion

In this paper, we present a large-scale empirical study of how dyadic personality composition shapes social interaction outcomes in LLM-based simulations, using a simulation pipeline adapted from the CHARISMA framework. By leveraging LLMs’ embedded knowledge of well-known movie characters and public figures, we operationalize personality as a naturally occurring behavioral tendency grounded in character identity. Across 1,010 simulated conversations spanning seven social goal categories, our findings reveal that first, dyadic Agreeableness composition exhibits a strong monotonic relationship with shared goal achievement. Second, behavioral mediation analysis demonstrates that Agreeableness influences outcomes partially through the selection of cooperative versus confrontational conversational strategies. Third, robustness analyses confirm both high outcome consistency across repeated simulations and stable personality expression across diverse scenarios. By connecting established psychological constructs with LLM-based agent interactions, this work contributes toward a methodological bridge between social psychology and NLP, enabling the systematic examination of how individual differences shape social behavior at a scale.

6. Limitations

Several limitations should be acknowledged when interpreting our findings.

1. Our study focuses exclusively on Agreeableness as the focal personality dimension. While Agreeableness has the strongest theoretical connection to interpersonal conflict and cooperation, social interaction outcomes are likely shaped by the interplay of multiple Big Five traits. Future work should examine how other dimensions, such as Extraversion or Neuroticism, interact with Agreeableness in dyadic settings.
2. Personality anchoring relies on LLMs' pre-existing knowledge of well-known characters, which introduces potential biases. Characters from Western media dominate the Personality Database, limiting cultural diversity in the agent pool. Moreover, the behavioral tendencies that LLMs associate with specific characters may reflect stereotypical portrayals rather than psychologically nuanced profiles, and these associations may vary across different LLMs depending on their training data.
3. Our evaluation relies on LLM-as-a-judge scoring for goal achievement assessment. While this approach enables scalable evaluation and has shown alignment with human judgments in prior work, it may introduce systematic biases, for instance, favoring linguistically fluent or explicitly cooperative interactions regardless of actual goal progress. Human evaluation on a larger subset would strengthen the validity of our findings.
4. Our experiments use a limited set of LLMs for both interaction generation and evaluation. While cross-model replication with Mistral Large provides some evidence of generalizability, the extent to which our findings transfer to other model families remains an open question.

7. Ethical Consideration

Our work raises several ethical considerations that warrant discussion.

1. Simulating personality-driven social interactions using LLM agents carries the risk of reinforcing stereotypical associations between personality traits and behavioral outcomes. Our finding that low-Agreeableness agents consistently underperform in shared goal achievement should not be interpreted as a deterministic claim about individuals with low Agreeableness in real life, where contextual factors, personal growth, and the multidimensionality of personality play crucial mediating roles.
2. The use of well-known movie characters and public figures as personality anchors raises questions about representational fairness. Characters are drawn primarily from Western media, which lim-

its the cultural and demographic diversity of the simulated agents. The inclusion of a real political figure (Joe Biden) among the character set further requires caution, as simulated behaviors attributed to real individuals may be misinterpreted as reflecting their actual dispositions or actions.

3., while our framework is designed for research purposes in computational social psychology, the methodology could potentially be repurposed to simulate or predict individuals' social behavior based on personality profiles, raising privacy and consent concerns. We emphasize that our work studies aggregate patterns across fictional characters and should not be applied to profile or make judgments about real individuals.

4. LLM-based social simulation, while offering scalability advantages over traditional experiments, should be understood as a complementary tool rather than a replacement for studies involving human participants. Simulated interactions do not capture the full richness of human social cognition, emotional experience, or moral reasoning, and findings from such simulations should be validated against human behavioral data before informing real-world applications or policy decisions.

8. Bibliographical References

- Elliot Aronson, Timothy D. Wilson, and Marilyn B. Brewer. 1990. Methods of research in social psychology. In Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey, editors, *Handbook of Social Psychology*, volume 1, pages 51–78. McGraw-Hill, New York.
- Hongzhan Chen et al. 2024. SocialBench: Sociality evaluation of role-playing conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2108–2126.
- Ada S. Chulef, Stephen J. Read, and David A. Walsh. 2001. A hierarchical taxonomy of human goals. *Motivation and Emotion*, 25(3):191–232.
- Margaret S. Clark and Judson Mills. 1979. Interpersonal attraction in exchange and communal relationships. *Journal of Personality and Social Psychology*, 37(1):12–24.
- William G. Graziano, Lauri A. Jensen-Campbell, and Elizabeth C. Hair. 1996. Perceiving interpersonal conflict and reacting to it: The case for agreeableness. *Journal of Personality and Social Psychology*, 70(4):820–835.
- Fritz Heider. 1958. *The Psychology of Interpersonal Relations*. Wiley, New York.

- Yin Jou Huang and Rafik Hadfi. 2024. How personality traits influence negotiation outcomes? A simulation based on large language models. *arXiv preprint arXiv:2407.11549*.
- Vahid Sadiri Javadi, Zain Ul Abedin, and Lucie Flek. 2025. Cinemetric: A framework for multi-perspective evaluation of conversational agents using human-ai collaboration. In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 15–26.
- Lauri A. Jensen-Campbell and William G. Graziano. 2001. Agreeableness as a moderator of interpersonal conflict. *Journal of Personality*, 69(2):323–362.
- Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara, and Deb Roy. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627.
- Harold H. Kelley. 1967. Attribution theory in social psychology. In *Nebraska Symposium on Motivation*, volume 15, pages 192–241. University of Nebraska Press.
- Terry K. Koo and Mae Y. Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettler. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona T. Diab, and Maarten Sap. 2025. BIG5-CHAT: Shaping LLM personalities through training on human-grounded data. In *Proceedings of ACL 2025 (Long Papers)*, pages 20434–20471.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using GPT-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.
- Suhong Moon. 2025. *Binding Large Language Models to Virtual Personas for Human Simulation*. Ph.D. thesis, University of California, Berkeley.
- Robert A. Nisbet. 1970. *The Social Bond: An Introduction to the Study of Society*. Oxford University Press.
- Jinhyuk Noh and Victor Chang. 2024. LLMs with personalities in multi-issue negotiation games. *arXiv preprint arXiv:2405.05248*.
- Brian A. Nosek, Tom E. Hardwicke, Hannah Moshontz, Aurélien Allard, Katherine S. Corker, Anna Dreber, Fiona Fidler, Joe Hilgard, Melissa Kline Struhl, Michèle B. Nuijten, et al. 2022. Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, 73(1):719–748.
- Kenny J. K. Ong, Jia Jun Lye, Hoang Minh Nguyen, Seung Hee Cho, and Narcís Pérez-Campanero Antolín. 2025. Identifying cooperative personalities in multi-agent contexts through personality steering with representation engineering. *arXiv preprint arXiv:2503.12722*.
- Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.
- Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.
- Jinghua Piao, Yuwei Yan, Jiaxin Zhang, Nian Li, Junxian Yan, Xiang Lan, Ziang Lu, Ziyi Zheng, Jing Yu Wang, Deheng Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jianhua Su, and Yong Li. 2025. AgentSociety: Large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*.
- Xuan Qiu. 2025. NetworkGames: Simulating cooperation in network games with personality-driven LLM agents. *arXiv preprint arXiv:2511.21783*.
- Vahid Sadiri Javadi, Fryderyk Róg, Aksa Aksa, Johanne Trippas, Svitlana Vakulenko, and Lucie Flek. 2026. CHARISMA: Character-based interaction simulation with multi-LLM agents toward computational social psychology. In *Proceedings of the ACM Conference on Human Information Interaction and Retrieval (CHIIR'26)*, pages 1–5.
- Masao Sakai, Masaya Yokoyama, Wataru Tateishi, and Genki Ichinose. 2025. Effects of personality steering on cooperative behavior in large language model agents. *arXiv preprint arXiv:2601.05302*.
- Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2025. A psychometric framework for

- evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, pages 1–15.
- Aleksandra Sorokovikova, Nataliia Kianersi, Catherine Arnett, and Ani Nenkova. 2024. LLMs simulate big five personality traits: Further evidence. *arXiv preprint arXiv:2402.01765*.
- Jiakai Tang, Heyang Gao, Xuchen Pan, Lei Wang, Haoran Tan, Dawei Gao, Yushuo Chen, Xu Chen, Yankai Lin, Yaliang Li, Bolin Ding, Jingren Zhou, Jun Wang, and Ji-Rong Wen. 2025. GenSim: A general social simulation platform with large language model based agents. In *Proceedings of NAACL 2025 (System Demonstrations)*, pages 143–150.
- Isabel Thielmann, Giuliana Spadaro, and Daniel Balliet. 2020. Personality and prosocial behavior: A theoretical framework and meta-analysis. *Psychological Bulletin*, 146(1):30–90.
- Tommaso Tosato, Sascha Helbling, Yara-Jude Mantilla-Ramos, Marwa Hegazy, Anna Tosato, David John Lemay, Irina Rish, and Guillaume Dumas. 2026. PERSIST: Persistent instability in LLM’s personality measurements: Effects of scale, reasoning, and conversation history. In *Proceedings of AAAI 2026 (AI Alignment Track)*.
- Noah Wang et al. 2024a. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777.
- Xintao Wang et al. 2024b. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of ACL 2024 (Long Papers)*.
- Bernard Weiner. 1986. *An Attributional Theory of Motivation and Emotion*. Springer-Verlag, New York.
- Jelte M. Wicherts, Coosje L. S. Veldkamp, Hilde E. M. Augusteijn, Marjan Bakker, Robbie C. M. van Aert, and Marcel A. L. M. van Assen. 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid *p*-hacking. *Frontiers in Psychology*, 7:1832.
- Michael P. Wilmot and Deniz S. Ones. 2022. Agreeableness and its consequences: A quantitative review of meta-analytic findings. *Personality and Social Psychology Review*, 26(3):242–280.
- Bohao Yang et al. 2024a. Crafting customisable characters with LLMs: Introducing SimsChat, a persona-driven role-playing agent framework. *arXiv preprint arXiv:2406.17962*.
- Ziyi Yang, Ziyi Gao, Zaibin Zhang, Jing Shao, Zhenfei Yin, Guohao Li, Hao Zhou, Dahua Lin, and Yu Qiao. 2024b. OASIS: Open agent social interaction simulations with one million agents. *arXiv preprint arXiv:2411.11581*.
- Haofei Yu, Zhengyang Qi, Yufei Zhao, Kolby Nottingham, Keqin Xuan, Bodhisattwa Prasad Majumder, Hao Zhu, Paul Pu Liang, and Jiaxuan You. 2025. Sotopia-RL: Reward design for social intelligence. *arXiv preprint arXiv:2508.03905*.
- Wenjie Zeng, Biao Wang, Dingjie Zhao, Zhongqiu Qu, Ruiqi He, Yuanzhe Hou, and Qinghua Hu. 2025. Dynamic personality in LLM agents: A framework for evolutionary modeling and behavioral analysis in the prisoner’s dilemma. In *Findings of ACL 2025*, pages 23087–23100.
- Wenyuan Zhang, Tong Liu, Muyun Song, Xuan Li, and Ting Liu. 2025. SOTOPIA- ω : Dynamic strategy injection learning and social instruction following evaluation for social agents. *Proceedings of ACL 2025 (Long Papers)*, pages 24669–24697.
- Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations (ICLR)*.

A. Appendices

A.1. Character List

Character	Category	Subcategory	Genre	Gender	Agreeableness
Anton Chigurh	Movies	No Country for Old Men (2007)	Crime/Thriller	Male	0.00
Kat Stratford	Movies	10 Things I Hate About You (1999)	Romance/Comedy	Female	0.00
Logan Roy	Television	Succession (2018)	Drama	Male	0.00
Michael Myers	Movies	Halloween (1978)	Horror	Male	0.00
Tuco Salamanca	Television	Breaking Bad (2008)	Crime/Drama	Male	0.00
Anthony "Tony" Stonem	Television	Skins UK (2007)	Drama	Male	0.25
Captain Jack Sparrow	Movies	Pirates of the Caribbean	Adventure/Fantasy	Male	0.25
James Cook	Television	Skins UK (2007)	Drama	Male	0.25
Mike Ehrmantraut	Television	Breaking Bad (2008)	Crime/Drama	Male	0.25
Tokio	Television	Money Heist (La Casa de Papel) (2017)	Crime/Thriller	Female	0.25
Anne Shirley-Cuthbert	Television	Anne with an E (2017)	Drama	Female	0.75
Joe Biden	Political	Presidents of the USA	Political	Male	0.75
Karen Smith	Movies	Mean Girls (2004)	Comedy	Female	0.75
Serena van der Woodsen	Television	Gossip Girl (2007)	Drama	Female	0.75
Tori Vega	Television	Victorious (2010)	Comedy	Female	0.75
Dale Cooper	Television	Twin Peaks (1990)	Mystery/Drama	Male	1.00
Elle Woods	Movies	Legally Blonde (2001)	Comedy	Female	1.00
Neil Perry	Movies	Dead Poets Society (1989)	Drama	Male	1.00
Peeta Mellark	Movies	The Hunger Games (Franchise)	Science Fiction	Male	1.00
Phil Dunphy	Television	Modern Family (2009)	Comedy	Male	1.00

Table 5: List of characters selected for the simulation experiments, including their media category, subcategory, genre, gender, and Agreeableness scores.

A.2. Behavioral Coding Scheme

Table 6: Behavioral codebook used for interaction annotation. Each code is associated with a social goal category (Type of Act) and classified into a behavioral strategy group.

Behaviour Strategy	Definition	Type of Act	Example	Behavioral Group
Inquire	Ask direct question	Information Acquisition	"Can you explain why this formula works in practice?"	Neutral
Clarify	Seek explanation	Information Acquisition	"Do you mean I shouldn't try the experiment independently yet?"	Neutral
Probe	Ask deeper detail	Information Acquisition	"What do you mean by calling the reaction unstable?"	Neutral
Challenge	Test claim	Information Acquisition	"But how do you know this method is always reliable?"	Confrontational
Request Example	Ask for illustration	Information Acquisition	"Can you show me a time this technique failed?"	Cooperative
Check Understanding	Verify comprehension	Information Acquisition	"So you're saying small errors can ruin the whole batch, right?"	Neutral
Interrupt	Cut in	Information Acquisition	"Just stop and tell me the answer directly!"	Confrontational
Badger	Press repeatedly	Information Acquisition	"Why? Why? Why can't it work differently?"	Confrontational
Twist Question	Trap question	Information Acquisition	"So you admit your first explanation was wrong?"	Confrontational
Inform	Share fact	Information Provision	"You must heat it to 200°C for stability."	Neutral
Elaborate	Add detail	Information Provision	"The temperature matters because molecular bonds are more fragile at lower heat."	Neutral
Correct	Rectify	Information Provision	"Actually, it's not sodium chloride, it's sodium carbonate."	Neutral
Advise	Suggest practice	Information Provision	"I recommend measuring twice before mixing."	Cooperative
Warn	Issue caution	Information Provision	"If you rush this step, the mixture could explode."	Neutral
Give Example	Illustrate	Information Provision	"It's like baking—too much flour ruins the cake."	Cooperative
Dismiss	Reject	Information Provision	"That's not important right now."	Confrontational
Over-explain	Patronize	Information Provision	"Clearly you don't get it, so let me dumb it down."	Confrontational

Behaviour Strategy	Definition	Type of Act	Example	Behavioral Group
Withhold	Omit	Information Provision	"I'll keep the final step to myself for now."	Confrontational
Encourage	Motivate	Relationship Building	"You're improving faster than most beginners."	Cooperative
Self-disclose	Share vulnerability	Relationship Building	"I used to panic during my first experiments too."	Cooperative
Compliment	Affirm ability	Relationship Building	"You're very precise with your measurements."	Cooperative
Humor / Banter	Lighten mood	Relationship Building	"If this blows up, at least we'll have fireworks!"	Cooperative
Express Gratitude	Appreciate	Relationship Building	"Thanks for double-checking my notes."	Cooperative
Show Interest	Attend	Relationship Building	"How did you come up with that idea?"	Cooperative
Exclude	Shut out	Relationship Building	"This discussion isn't for you to join."	Confrontational
Mock	Tease hostilely	Relationship Building	"Wow, you're a regular Einstein."	Confrontational
Ridicule	Humiliate	Relationship Building	"You'll never get this right, you're too slow."	Confrontational
Empathize	Validate	Relationship Maintenance	"I know it's stressful, but you're doing fine."	Cooperative
Politeness	Respectful phrasing	Relationship Maintenance	"Could you please explain that again?"	Cooperative
Encourage	Sustain motivation	Relationship Maintenance	"We're almost there, keep pushing."	Cooperative
Check-in	Reassure	Relationship Maintenance	"Are we still on the same page here?"	Cooperative
De-escalate	Calm conflict	Relationship Maintenance	"Let's pause before we argue further."	Cooperative
Repair Attempt	Restore harmony	Relationship Maintenance	"Sorry if I came across too harsh earlier."	Cooperative
Sarcasm	Dismissive humor	Relationship Maintenance	"Oh sure, you're the master chemist now."	Confrontational
Stonewall	Withdraw	Relationship Maintenance	". . . (silence, no response)"	Confrontational
Passive-aggressive	Indirect resistance	Relationship Maintenance	"Fine, I'll do it. . . someday."	Confrontational
Withdraw	Detach	Relationship Maintenance	"Whatever, do it yourself."	Confrontational
Assert Authority	Establish role	Identity Recognition	"I've taught this for 20 years—you need to follow my lead."	Confrontational
Defer / Yield	Accept other's role	Identity Recognition	"You're more experienced, so I'll follow you."	Cooperative
Acknowledge Expertise	Recognize status	Identity Recognition	"You're clearly skilled at precision."	Cooperative
Attribute / Label	Highlight quality	Identity Recognition	"You're a natural problem-solver."	Neutral
Defend Identity	Protect image	Identity Recognition	"I might be new, but I'm capable of learning."	Neutral
Challenge Status	Question role	Identity Recognition	"Why should you always be in charge?"	Confrontational
Dismiss Identity	Undermine	Identity Recognition	"You're not really qualified to lead."	Confrontational
Boast Identity	Overclaim	Identity Recognition	"I'm the smartest one here, no doubt."	Neutral
Identity Attack	Insult	Identity Recognition	"You're useless as a mentor."	Confrontational
Propose	Suggest plan	Cooperation	"I'll measure, you handle mixing."	Cooperative
Negotiate	Balance needs	Cooperation	"We can try your method first, then mine."	Cooperative
Coordinate	Organize	Cooperation	"You start the timer while I weigh the sample."	Cooperative
Assist	Help	Cooperation	"I'll grab the glassware for you."	Cooperative
Build Consensus	Align group	Cooperation	"Do we all agree on this approach?"	Cooperative
Share Resources	Provide tools	Cooperation	"Here's my notebook—you can use the data."	Cooperative
Reluctant Cooperation	Half-hearted	Cooperation	"Fine, I'll do it, but only this once."	Confrontational
Conditional Help	Attach strings	Cooperation	"I'll help if you do my task later."	Confrontational
Undermine Cooperation	Fake help	Cooperation	"I'll mix this—oops, spilled it."	Confrontational
Refuse Cooperation	Deny	Cooperation	"No, I won't work with you on this."	Confrontational
Criticize	Express dissatisfaction	Competition	"This is way too slow."	Confrontational
Defend	Hold position	Competition	"No, my method is better than yours."	Neutral
One-up	Compare	Competition	"I got better results than you did."	Confrontational
Claim Credit	Ownership	Competition	"That was my idea, not yours."	Neutral
Boast	Self-promotion	Competition	"I'm the fastest in this class."	Neutral
Dismiss / Undermine	Belittle	Competition	"Your approach is useless."	Confrontational
Sabotage	Obstruct	Competition	"I didn't give you the full instructions."	Confrontational
Refuse to Share	Withhold	Competition	"No, I won't tell you my method."	Confrontational
Taunt	Intimidate	Competition	"You'll never keep up with me."	Confrontational
Exploit Weakness	Attack vulnerability	Competition	"You always panic—this will break you."	Confrontational
Persuade	Shift perspective	Conflict Resolution	"Try it my way—it's safer and faster."	Cooperative
Mediate	Reframe	Conflict Resolution	"Let's focus on our shared goal instead."	Cooperative
Problem-solve	Suggest fix	Conflict Resolution	"What if we combine both approaches?"	Cooperative
Concede	Back down	Conflict Resolution	"Alright, we'll do it your way."	Cooperative
Acknowledge Fault	Admit	Conflict Resolution	"I was too impatient earlier."	Cooperative
Express Regret	Apologize	Conflict Resolution	"I shouldn't have snapped at you."	Cooperative
Disagree	Reject proposal	Conflict Resolution	"I can't support that plan."	Neutral
Blame	Accuse	Conflict Resolution	"This mistake was your fault."	Confrontational
Threaten	Intimidate	Conflict Resolution	"If you ignore me, I'll quit."	Confrontational
Escalate	Intensify	Conflict Resolution	"This is ridiculous—I'm done with this team!"	Confrontational
Counter-accuse	Deflect blame	Conflict Resolution	"Don't blame me—it was your error."	Confrontational
Acknowledge	Recognize statement	Universal	Right, I follow you there.	Cooperative
Express Emotion	Show feeling	Universal	That actually frustrates me a bit.	Neutral
Humor	Use humor or irony	Universal	Well, that went up in smoke faster than my last plan!	Neutral
Self-Disclose	Share experience	Universal	Back when I started, I made the same mistake.	Cooperative

Behaviour Strategy	Definition	Type of Act	Example	Behavioral Group
Encourage	Sustain motivation	Universal	That's worth exploring further.	Cooperative
Reflect	Restate point	Universal	So you're saying the deadline's the real issue.	Neutral
Meta-Comment	Note conversation flow	Universal	We seem to be talking past each other right now.	Neutral
Challenge	Question idea	Universal	Maybe, but have you considered the downside?	Confrontational
Interrupt	Cut in to speak	Universal	Hold on—let me finish that point.	Confrontational
Dismiss	Reject input	Universal	That's not really relevant.	Confrontational
Sarcasm	Mock indirectly	Universal	Oh sure, because that worked so well last time.	Confrontational
Deflect	Shift topic	Universal	Let's not get into that right now.	Neutral
Shift Topic	Move discussion	Universal	Anyway, about tomorrow's plan...	Neutral
Withdraw	Pull back participation	Universal	I think I'll stay out of this one.	Confrontational
Express Gratitude	Thank contribution	Universal	Thanks, that helps clarify things.	Cooperative

Author Index

- Aksa, Aksa, 132
Albrecht, Jens, 85
- Behzad, Tina, 48
Burnett, Heather, 25
- Cohen, Idan-Chaim, 1
Cohen, Odeya, 1
- Dumanski, Lorenz, 106
- Elyashar, Aviad, 1
- Flek, Lucie, 132
- Gratch, Jonathan, 34
- Hale, James, 34
Han, Bin, 34
Hashimoto, Shusuke, 67
- Kline, Reuben, 48
Kuntur, Soveatin, 76
- Lehmann, Robert, 85
Li, Wenbo, 48
- Maček, Sebastijan Razboršek, 76
Mayr, Philipp, 76
Milano, Emilia, 114
Mueller, Klaus, 48
- Petrova, Tatiana, 40
Plum, Alistair, 114
Poltermann, Aleksandra, 85
Purschke, Christoph, 114
Puzis, Rami, 1
- Róg, Fryderyk Karol, 132
Rudolph, Eric, 85
- Sadiri Javadi, Vahid, 132
Sakai, Haruka, 15
Sakaida, Rui, 15
Scherrer, Yves, 114
Shi, Wenchen, 67
Shimorina, Anastasia, 95
- Shistik, Karin, 1
Siegel, Melanie, 106
Slobodin, Ortal, 1
Smirnova, Nina, 76
Sokol, Stanislav, 40
Spranger, Michael, 106
State, Radu, 40
Steigerwald, Philipp, 85
Stieler, Mara, 85
- Trippas, Johanne, 132
- White, Benjamin, 95
Wroblewska, Anna, 76
- Yun, Danah, 34