Probing Large Language Models for Scalar Adjective Lexical Semantics and Scalar Diversity Pragmatics

Fangru Lin, Daniel Altshuler, Janet B. Pierrehumbert
University of Oxford
Oxford, United Kingdom
{fangru.lin, daniel.altshuler}@ling-phil.ox.ac.uk, janet.pierrehumbert@oerc.ox.ac.uk

Abstract

Scalar adjectives pertain to various domain scales and vary in intensity within each scale (e.g. *certain* is more intense than *likely* on the likelihood scale). Scalar implicatures arise from the consideration of alternative statements which could have been made. They can be triggered by scalar adjectives and require listeners to reason pragmatically about them. Some scalar adjectives are more likely to trigger scalar implicatures than others. This phenomenon is referred to as scalar diversity. In this study, we probe different families of Large Language Models such as GPT-4 for their knowledge of the lexical semantics of scalar adjectives and one specific aspect of their pragmatics, namely scalar diversity. We find that they encode rich lexical-semantic information about scalar adjectives. However, the rich lexical-semantic knowledge does not entail a good understanding of scalar diversity. We also compare current models of different sizes and complexities and find that larger models are not always better. Finally, we explain our probing results by leveraging linguistic intuitions and model training objectives.

Keywords: Scalar Adjective, Scalar Implicature, Lexical Semantics, Pragmatics

1. Introduction

Scalar adjectives (SAs) are words such as *likely*, *certain*, *warm*, and *scalding*. They describe different scales of properties. For instance, *warm* and *scalding* describe temperature, while *likely* and *certain* describe probabilities. Scalar adjectives can describe the same scale while differing in intensity. For example, *certain* is more intense than *likely* on the likelihood scale because it is used to make a logically stronger statement about a given situation.

Scalar implicatures (SIs) arise from the consideration of alternative statements that could have been made (Figure 1). They can be triggered by SAs. For instance, when a speaker utters 'It is *likely* to rain', a hearer may conclude that the speaker is *not certain* that it will rain. In particular, the hearer may reason that the speaker could have provided the logically stronger statement, 'it is *certain* to rain', but did not do so. Psycholinguistic studies indicate that some SAs are more likely to generate implicatures than others (Van Tiel et al., 2014; Gotzner et al., 2018; Ronai and Xiang, 2022). This phenomenon is referred to as *scalar diversity*. For instance, *likely* tends to indicate *not certain*, while *good* does not tend to indicate *not excellent*.

SI is a long-standing topic of research in pragmatics because it reveals fundamental aspects of human linguistic and cognitive capabilities in areas such as the Theory of Mind (Feng et al., 2021). Accordingly, SIs pose important challenges for the development of NLP models with human-like capabilities (Sap et al., 2022). SIs are also important in practice for downstream tasks that include Natural Language Inference (Williams et al., 2018), Sentiment Analysis (Socher et al., 2013), and indirect Question Answering (de Marneffe et al., 2010). State-of-the-art large Language Models (LLM) such as GPT-4[1] have remarkable performance on many classic benchmarks. However, they have proved to be fragile in some semantic and pragmatic tasks that are easy for humans (Liu et al., 2023a; Lin et al., 2024). The goal of this paper is to probe this discrepancy with an in-depth investigation of SIs with SAs.

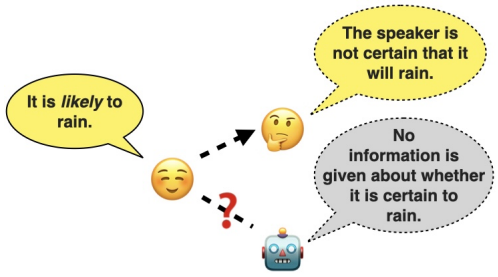

Figure 1: People often mean more than what they literally say. Humans can easily infer implied messages, while LLMs often fail to do so.

Recent research on LLMs' understanding of SA has focused on: (i) probing the lexical semantics of SAs (Liu et al., 2023b) or (ii) their scalar diversity (Hu et al., 2022, 2023). In this paper, we investigate both (i) and (ii). We first utilize two probing methods to evaluate LLMs' knowledge about SAs' scale membership and intensity information. Using previously published datasets, we then assess whether LLMs show human-like scalar diversity in judgments about SA items. Finally, we compare LLMs' pragmatic and lexical-semantic knowledge

[1] https://openai.com/research/gpt-4

LREC-COLING 2024, pages 13033–13049
20-25 May, 2024.

and explain our observations based on linguistic theory.

Our paper includes three main findings. First, LLMs generally encode rich lexical-semantic information about SAs (§3). Second, LLMs have unsatisfying performance in capturing scalar diversity despite encoding rich lexical-semantic information about SAS (§4). Third, the size of the LLMs does not correlate with how well they perform on our tasks (§3 and §4): While the increase in model size is sometimes claimed to invariably improve performance (the so-called "scaling law"), in our study, some large models do worse than smaller models that have different architectures or training objectives.[2]

2. Related works

2.1. Scalar Adjective Lexical Semantics

Since the seminal work by Kamp (1975), SAs have received a lot of attention in formal semantics and the philosophy of language. The same cannot be said about NLP research, where SAs have only recently garnered attention. Some of this research has been concerned with the lexical semantics of SAs, particularly focusing on their scale membership and their intensities (de Melo and Bansal, 2013; Kim and de Marneffe, 2013; Shivade et al., 2015; Wilkinson and Tim, 2016; Cocos et al., 2018). One feature of this research is the use of static rather than contextualized word embeddings from LLMs. This is important because Garí Soler and Apidianaki (2020) have found that BERT-base (Devlin et al., 2019) contextualized word embeddings encode richer information about scalar intensities in vector space. Nevertheless, Liu et al. (2023b) report that many state-of-the-art models, even after fine-tuning on MNLI (Williams et al., 2018), have unsatisfying adjective degree estimations in textual inference.

2.2. Scalar Implicature and Scalar Diversity Pragmatics

In a study of SIs, Schuster et al. (2020) have found that LSTM (Hochreiter and Schmidhuber, 1997) can be trained to infer *some-not all* implicatures (e.g. 'Some student cheated' implicates that not all students cheated). Jeretic et al. (2020) report that BERT fine-tuned on MNLI nearly always predicts that the determiner *some* entails *not all*, but that SAs are treated as synonyms regardless of their intensity. This means that the model may lack relevant pragmatic knowledge about adjectives. In contrast, two recent studies (Hu et al., 2022, 2023) achieved more favorable results with surprisal measures derived from GPT2. String-based surprisal, which considers the surprise level of a given strong word to appear as an alternative to a weak word out of all possible strings in a context, is estimated by the likelihood of a strong word to appear in its position by a GPT-2 model (Radford et al., 2019). Concept-based surprisal treats an alternative as a member of a string set with similar concepts. The surprisal rate for the alternative is considered as an average over that of all strings in such a set. Using these two measures, they found that scalar diversity correlates with string-based and concept-based surprisal. Ruis et al. (2024) tested different LLMs on general conversational implicature understanding. They found that GPT-4 is the best-performing model with 30-shot chain-of-thought prompting (Wei et al., 2022), achieving 88.66% in generalized implicature calculation. This is the category that our adjective-triggered SIs fall under.

2.3. Direct and Indirect Probing

Direct and indirect probing methodologies are widely used to understand LLMs' knowledge. Direct probing is used to analyze the hidden representations encoded in LLMs (Garí Soler and Apidianaki, 2020). This is only possible when the LLM is open-source. Indirect probing is used to analyze the representations of closed-source models. This involves testing performance on various semantic tasks. Typical tasks use textual prompts to assess the prompted answers (Ettinger, 2020). Petroni et al. (2019) argue that indirect probing only reveals the lower bound of model capabilities. Follow-up works have provided various methods to increase the bound (Zhong et al., 2021; White et al., 2023).

We build on these results as follows. Since direct probing provides a good estimate of a model's competence, but closed-source LLMs only allow for indirect probing, we use both methods for some open-source models. We then use the results on these models to ground the results of indirect probing for other models. We make the following assumption: if the two probing methods show similar trends regarding model performance for the open-source models, we deem our indirect probing methods as valid in comparing the relative capabilities of different closed-source models.

3. Probing Lexical Semantics

Recent research on the lexical semantics of SAs have mostly focused on deriving their intensity information from open-source models (Garí Soler and Apidianaki, 2020; Liu et al., 2023b). In this section, we describe how we probe three model families, for

[2]Code is in https://github.com/fangru-lin/llm_scalar_adj.

two aspects of SA lexical semantics: (i) scale membership (e.g. *likely* and *certain* are on the likelihood scale), and (ii) adjective intensity (e.g. *certain* is more intense than *likely* within the same scale).

We assess three model families, containing eight models in total, to understand how different model architectures, training objectives, and sizes affect lexical-semantic knowledge. All models except GPT-4 are accessed via huggingface library (Wolf et al., 2020), and GPT-4 is accessed via OpenAI API.[3]

Encoder Models BERT-base/large(b/l) (110M/340M) (Devlin et al., 2019), RoBERTa-base/large(b/l) (123M/354M) (Liu et al., 2019)

Decoder Models Falcon-7B-instruct (Falcon) (Almazrouei et al., 2023), GPT-4[4]

Encoder-decoder Models Flan-T5-xl/xxl (3.5B/11B) (Chung et al., 2022)

3.1. Datasets for Lexical-Semantic Probing

3.1.1. Scalar Adjective Datasets

Following Garí Soler and Apidianaki (2020), we use three SA datasets: DEMELO (**DM**) (de Melo and Bansal, 2013), CROWD (**CD**) (Cocos et al., 2018), and WILKINSON (**WK**) (Wilkinson and Tim, 2016) which contain 185 half-scales in total (Table 1). For example, the positive side of the quality scale ranges from *good* to *awesome*. Each contains different categories of SAs. Some contain adjectives describing physical appearance while others do not.

	DM	CD	WK
Half-scales	87	77	21
Distinct adjective pairs	548	330	61

Table 1: Overview of half-scale counts and distinct adjective pairs in SA datasets.

DM Half-scales are first collected from WordNet 3.0 (Gross, Derek and Miller, Katherine J, 1990) and then annotated for intensity by two native English speakers.

CD Adjectives are first collected from the Paraphrase Data Base (Ganitkevitch et al., 2013; Pavlick et al., 2015) and then annotated by crowd workers. Annotators are asked to identify whether given adjectives are on the same scale for multiple rounds, and then annotate the intensity of adjectives.

WK Adjectives are collected via crowdsourcing. Crowd workers are first presented with prompt words that belong to a full scale and asked to list other semantically related adjectives. Scales are, then, cleaned automatically based on workers' competence and annotated manually for adjective intensities. This study uses partitioned **WK** with half scales in Cocos et al. (2018), which are derived from the Paraphrase Data Base (Ganitkevitch et al., 2013; Pavlick et al., 2015).

3.1.2. Context Sentence Datasets

To compute contextualized word embeddings for adjectives, we use the context sentence dataset **ukWac** from Garí Soler and Apidianaki (2020). **ukWac** provides 10 context sentences for each half-scale *s* in **DM**, **CD**, and **WK**. All adjectives on *s* share the same context set of 10 different sentences. Each candidate sentence contains an adjective on *s* and allows lexical substitution of its scale-mates. The acceptability after substitution is ensured by computing their acceptability scores using context2vec (Melamud et al., 2016). An example is given below, where the italic word is an adjective on *s*, and appears in the original context sentence. Other scale alternatives to it are inside the brackets.

> They are extremely catchy and are not only great to listen to, but they are also *thrilling* (interesting/moving/exciting) to sing.

3.2. Probing Scale Membership

In this subsection, we assess whether LLMs encode the notion of scale membership. Ideally, we would expect models to be able to assign adjectives to their corresponding scales in context (e.g. *warm* belongs to the temperature scale).

3.2.1. Scale Membership Direct Probing Method

We directly probe scale membership by getting contextualized representations of SAs and half-scales and calculating their similarities. We retrieve word embeddings using **ukWac** by sampling different sentences for each adjective a on a half-scale s to obtain its contextualized representation $\vec{a}$. We derive a scale vector $\vec{s}$ for each s by adding the representations of the weakest and strongest adjectives on s. For instance, for the scale *adequate-fine-fitting-good*, where *adequate* is the least intense and *good* is the most intense adjective, $\vec{s}$ is $\vec{adequate} + \vec{good}$. Then, the cosine similarity $cos(\vec{a}, \vec{s})$ between each adjective a in a SA dataset D_a and all other $\vec{s}$ in the dataset is computed. The scales are ranked by $cos(\vec{a}, \vec{s})$ for each adjective; the scale with the highest $cos(\vec{a}, \vec{s})$ is considered to be the most likely scale for a to belongs to, and so

[3] See appendix 11.6 for more implementation details.

[4] We do not use LLaMA (Touvron et al., 2023) due to our institutional requirement. We use gpt-4-0613 for GPT-4.

forth. For evaluation, the ranking of the scale s that a belongs to ($rank_s$) is considered. We consider Mean Reciprocal Rank (MRR) as the evaluation metric; see Equation (1).

$$MRR = \frac{1}{|D_a|} \sum_{s \in D_a} \frac{1}{rank_s} \quad (1)$$

Intuitively, the higher the scale that an adjective belongs to is ranked, the closer MRR will be to 1. For instance, if all adjectives are aligned on their corresponding scales, the MRR will be 1. If models fail to rank the correct scales as closest, it means that they do not encode word polysemes and fine-grained distinction among different scales.

3.2.2. Scale Membership Direct Probing Experiment and Results

When target adjectives are tokenized as multiple segments, token representations are averaged as the adjectives' final representations. We use layer-wise adjective word embeddings to evaluate information in each layer (e.g. $\vec{a}$ and $\vec{s}$ have 12 representations using a 12-layer model). The experiment is repeated ten times with different random seeds for context sentences. We use fast-text static word embeddings[5] (Mikolov et al., 2018) trained on 600B Common Crawl data as a baseline. For the baseline, we add the representations of strongest and weakest adjectives on *s* as $\vec{s}$, then also rank $cos(\vec{a}, \vec{s})$ for each *a* to calculate MRR. Here, we only report the results for the best-performing layer for non-baseline models in Table 2.[6]

Models	DM	CD	WK
fast-text	0.842	0.716	0.983
BERT-b	0.829$_{\pm 0.010}$	0.797$_{\pm 0.010}$	**0.997**$_{\pm 0.004}$
BERT-l	**0.853**$_{\pm 0.007}$	**0.805**$_{\pm 0.011}$	**0.997**$_{\pm 0.006}$
RoBERTa-b	0.668$_{\pm 0.014}$	0.705$_{\pm 0.007}$	0.906$_{\pm 0.018}$
RoBERTa-l	0.777$_{\pm 0.011}$	0.757$_{\pm 0.008}$	0.977$_{\pm 0.010}$

Table 2: Direct scale membership probing results, subscripted numbers are standard deviation in ten runs. The best results per dataset are in bold.

The fast-text baseline is quite strong considering its training data size. All LLMs encode rich scale information as they mostly outperform (with some behaving on par with) the baseline. Model-wise, BERT models encode more information than RoBERTa. Within the same architecture, the bigger a model is, the better they are. Across datasets, **WK** is relatively easy because it (i) has fewer polysemous adjectives with the same forms on different scales and (ii) has fewer scales for classification than the other two datasets.

3.2.3. Scale Membership Indirect Probing Method

Inspired by Lorge and Pierrehumbert (2023) and de Melo and Bansal (2013), we use four templates including *ADJ$_{weak}$, if not ADJ$_{strong}$* to indirectly assess LLMs' knowledge about scale membership. Adjectives on the same scale are more likely to co-occur in these constructions than those that are not (e.g. *warm, if not hot* should be more likely to appear than *warm, if not thin*).[7]

For each adjective on a half scale that is not the strongest item, we use it as *ADJ$_{weak}$* in our templates and obtain 5 most likely words as *ADJ$_{strong}$*. For instance, for scale *adequate-fine-fitting-good*, we prompt models to answer what is likely to be *ADJ$_{strong}$* in *adequate/fine/fitting, if not* ADJ$_{strong}$. We consider a case to be correct if any top-5 word in *ADJ$_{strong}$* is on the same scale with *ADJ$_{weak}$*, and the overlap is not trivial (e.g. *adequate or even adequate* is a trivial completion, *good or even adequate* is considered to be correct despite the wrong intensity ranking).

3.2.4. Scale Membership Indirect Probing Experiment and Results

For encoder models, we put [MASK] and a comma in the position of *ADJ$_{strong}$* (e.g. *ADJ$_{weak}$, if not [MASK],*) as preliminary experiment results show that BERT mostly predicts punctuation marks as top 5 predictions for [MASK] when it appears as the last token. The comma is used because it does not constrain whether its preceding adjective is attributive (i.e. *can* be immediately followed by a noun) or predicative (i.e. *cannot* be immediately followed by a noun). For non-baseline models except GPT-4, we directly generate *ADJ$_{strong}$* given the preceding prompt. We use a similar generation objective for GPT-4 with the best template in other models (see appendix 11.7). We use the Google Ngram corpus slice for 2019 (Lin et al., 2012) as the indirect probing baseline with a similar setting, to obtain the 5 most likely completions for *ADJ$_{strong}$* in each construction. In Table 3, we report the best results among all templates in the assessed dataset (we report results based on selecting the best templates with held-out datasets in Table 13).

Indirect probing always underperforms direct probing for all models. This result aligns with Petroni et al. (2019). The relative rankings of different models are largely consistent. BERT is better than RoBERTa given the same size class, and larger encoder models are better than smaller ones

[5] https://github.com/facebookresearch/fastText/

[6] We also tried other alternative methods to compute membership, which does not generally work as well as the method we report here. See Appendix 11.1.

[7] See Appendix 11.3 for all templates.

Models	DM	CD	WK
Google Ngram	0.287_1	0.075_1	0.368_1
BERT-b	0.425_1	0.066_1	0.167_1
BERT-l	0.504_1	0.109_3	0.250_1
RoBERTa-b	0.271_4	0.064_3	0.114_1
RoBERTa-l	0.463_1	0.100_1	0.286_1
Falcon	0.265_1	0.045_1	0.167_1
GPT-4	0.540	**0.273**	**0.500**
Flan-T5-xl	**0.544**$_3$	0.177_3	**0.500**$_3$
Flan-T5-xxl	0.515_3	0.156_3	0.364_3

Table 3: Indirect probing results for scale membership. The best results per dataset are in bold. Subscripts refer to the indices of the best templates.

with the same architecture. These observations support the validity of our methods.

We find that all models except Falcon encode rich information about scale membership, in that they either behave on par with the baseline or outperform it. GPT-4 and Flan-T5-xl each achieve the best results on two benchmarks. While GPT-4 has competitive results with Flan-T5-xl in **DM** and **WK**, it is much better for **CD**. Therefore, we conclude that GPT-4 is the best-performing model in this task. Within the same model family, the scaling law tends to hold (i.e., larger is better). However, scaling does not fully explain the pattern of results. For instance, Falcon underperforms much smaller encoder models and even the baseline, and Flan-T5-xxl underperforms Flan-T5-xl.

3.3. Probing Scalar Intensity

In this section, we describe how we probe whether LLMs understand that SAs on the same scale have varied intensities (e.g. *hot* denotes a higher temperature than *warm*).

3.3.1. Scalar Intensity Direct Probing Method

Unlike previous methods which ground SAs in identical contexts and calculate them in separate runs (e.g. It is a *good/great/wonderful/awesome* movie) (Garí Soler and Apidianaki, 2020) (G&A), we use a novel method, which involves binding all SAs on the same scale as one single input to obtain their contextualized representations (e.g. *good great wonderful awesome*), as illustrated in Figure 2. Our assumption is that people are more likely to notice scalar words have different intensities when they are salient in the same context (Ronai and Xiang, 2023a).

For each half-scale s, we shuffle and bind all SAs a on s ten times to get 10 different inputs (e.g. *good great wonderful awesome*, *great wonderful good awesome*, etc.). For each a, we average the representations of it in 10 inputs to get its final representation. This is to avoid potential heuristic associations between word order and scalar intensity.

After getting the representation of all a, we utilize the best method (DiffVec) in Garí Soler and Apidianaki (2020) (G&A) to get the intensity ranking for them. Specifically, the intensity of all adjectives a in dataset D_a is measured by their cosine similarities to an external global intensity vector $\vec{dVec}$ derived from another SA dataset D_{vec} ($D_{vec} \neq D_a$).[8]

For every half-scale s in D_{vec}, the mildest and the extreme adjectives on s are notated as a_{ms} and a_{es}. We calculate $\vec{dVec}$ with Equation (2).

$$\vec{dVec} = \frac{1}{|D_{vec} \setminus D_a|} \sum_{s \in D_{vec} \setminus D_a} \vec{a_{es}} - \vec{a_{ms}} \quad (2)$$

For all $s \in D_a$, adjectives on s are ranked by their cosine similarities to $\vec{dVec}$: higher similarity means higher intensity. If two adjectives have the same similarity, they are treated as equally intense.

3.3.2. Scalar Intensity Direct Probing Experiment and Results

We probe layer-wise word embeddings in BERT-b/l, RoBERTa-b/l and report the best results for pairwise accuracy across layers using $\vec{dVec}$ in different datasets. We again provide fast-text embeddings as a baseline here. For the baseline, we use static embeddings of $\vec{a_{es}}$ and $\vec{a_{ms}}$ for all $s \in D_{vec} \setminus D_a$ to derive $\vec{dVec}$ as in Equation 2.

Results are averaged over adjective pairs and shown in Table 4.

Model	Method	DM	CD	WK
fast-text	-	0.637_{WK}	0.685_{DM}	0.836_{CD}
BERT-b	G&A	0.646_{CD}	0.735_{DM}	0.902_{DM}
	Ours	0.639_{CD}	0.706_{DM}	**0.967**$_{DM}$
BERT-l	G&A	**0.695**$_{CD}$	0.731_{DM}	0.918_{DM}
	Ours	0.673_{CD}	0.727_{DM}	0.902_{DM}
RoBERTa-b	G&A	0.557_{WK}	0.645_{DM}	0.820_{DM}
	Ours	0.648_{CD}	0.748_{DM}	0.934_{DM}
RoBERTa-l	G&A	0.595_{CD}	0.682_{DM}	0.836_{DM}
	Ours	0.664_{CD}	**0.752**$_{DM}$	0.934_{DM}

Table 4: Direct scalar intensity probing with our method and G&A on different models. Subscripted letters refer to the source dataset of $\vec{dVec}$. The best results per dataset are in bold. The best results using the same model are underlined.

We find that our simple method is generally better than G&A.[9] Globally, our method achieves the best-ranking results in two out of three datasets. Although G&A marginally outperforms our method

[8] We don't use the global intensity vector in D_a due to the concern of generalisability.

[9] Our method is also superior to Garí Soler and Apidianaki (2020) in other metrics. See appendix 11.2.

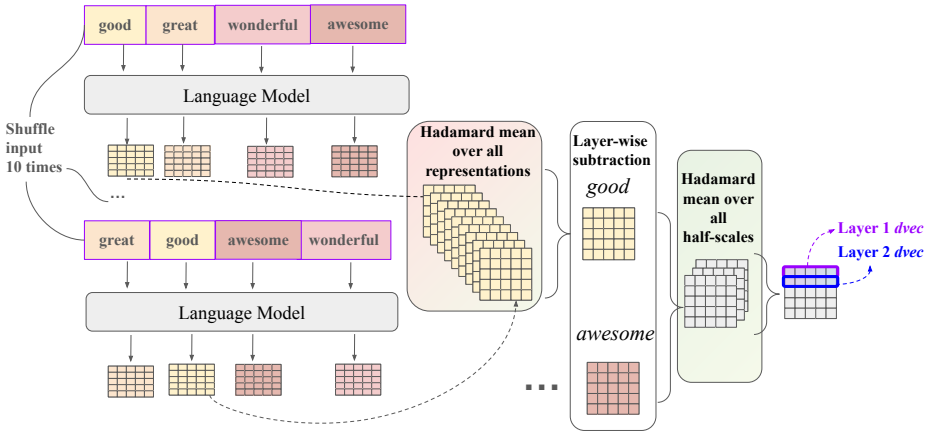

Figure 2: The process to derive intensity vector $\vec{d_{vec}}$. First, an adjective half-scale is randomly shuffled ten times for the order of adjectives as inputs to a language model. Then the encoded word vectors for the same word in different inputs are conducted with the Hadamard mean to derive the final representation of the word. After that, intensity vector $\vec{d_{vec}}$ is calculated by subtracting layer-wise representation of the weakest adjective from the strongest adjective ($\vec{awesome} - \vec{good}$ in this case) then averaging over all relevant half-scale subtractions in a dataset. Then layer-wise $\vec{d_{vec}}$ is used to probe language models' knowledge for adjective intensities.

in some of the datasets in BERT, our method uniformly largely outperforms G&A in RoBERTa, showing that it is more robust across different models and datasets.

All models encode rich information about SA intensity by outperforming the baseline. Across models, BERT models again dominate in two out of three datasets, showing that they encode more intensity information than RoBERTa, which indicates that larger models are not always better. Across datasets, **WK** is again relatively easy because it does not contain ties.

3.3.3. Scalar Intensity Indirect Probing Method

To complement closed-source models, we indirectly probe intensity using a similar objective as in Lorge and Pierrehumbert (2023). Specifically, for each pair of SAs to compare intensities, we compare the perplexity of minimal-pair prompts containing these adjectives such as '*good* but not *awesome*' and '*awesome* but not *good*', with the former featuring correct order of scalar adjectives and the latter incorrect. Ideally, models should always compute lower perplexity for the first prompt, as the latter is infelicitous or even ungrammatical.

We adopt all 34 templates used in the data collection of de Melo and Bansal (2013). [10] For each template, we compare the perplexity of correct and incorrect constructions. We consider that LLMs encode the scalar intensity for the corresponding adjective pair if the correct one has lower perplexity. We consider two adjectives to be equally intense if switching their positions in a template produces equally unlikely phrases.

3.3.4. Scalar Intensity Indirect Probing Experiment and Results

For encoder models, we estimate pseudo-perplexity by masking out tokens one by one and multiplying the likelihood of the original token appearing in the mask position to compute the sequence's pseudo-likelihood (Salazar et al., 2020). For other models except GPT-4, we directly compute perplexity for each sequence. For GPT-4, we use the best-performing template for other models and mimic the setting as in Appendix 11.7. As a baseline, we use Google Ngram corpus slice in 2019 to retrieve the likelihood of constructions as an analogy to perplexity. We then consider a case to be classified correctly if the correct construction is more likely to appear than the incorrect one. We use pair-wise accuracy as an evaluation metric and report the best results among all templates in the assessed dataset. Final results are averaged over pairs and shown in Table 5 (we report results based on selecting the best templates with held-out datasets in Table 14).

[10] Except *ADJ$_{strong}$(,) or very ADJ$_{weak}$* because some adjectives examined in this study are extreme ones (Paradis, 1998) which cannot be modified by *very* (e.g. *very terrific*). See Appendix 11.4 for the full template list.

Models	DM	CD	WK
Google ngram	0.312_{0}	0.222_{0}	0.442_{0}
BERT-b	0.569_{21}	0.539_{20}	0.770_{1}
BERT-l	0.569_{32}	0.558_{21}	0.738_{32}
RoBERTa-b	0.544_{14}	0.527_{19}	0.689_{14}
RoBERTa-l	0.524_{29}	0.585_{17}	0.754_{4}
Falcon	0.452_{11}	0.567_{18}	0.623_{7}
GPT-4	0.666	**0.739**	0.852
Flan-T5-xl	$\mathbf{0.684}_{17}$	0.621_{17}	$\mathbf{0.869}_{16}$
Flan-T5-xxl	0.633_{16}	0.655_{17}	0.787_{16}

Table 5: Indirect scalar intensity ranking table. Subscripted numbers are the best-performing template number. The best results across models are in bold.

We found that models all encode rich information about adjective intensity by significantly outperforming the baseline. For BERT and RoBERTa, we again find that direct and indirect probing show similar relative performance, with absolute results for indirect probing lower than those for direct probing.

GPT-4 dominates again in this task as it largely outperforms all models in one dataset and has competitive performance with Flan-T5-xl in the other two. The scaling law does not explain the pattern of results; across different families, Falcon is much larger than BERT and RoBERTa models but performs much worse. Even within encoder models, RoBERTa also fails to outperform BERT.

4. Probing Scalar Diversity Pragmatics

Given that LLMs have different levels of lexical-semantic knowledge about SAs, we can now ask whether these different levels correlate with the LLMs' ability to draw pragmatic inferences. This section assesses one particular aspect of this question, namely whether models can reason about the scalar diversity of SAs.

4.1. Scalar Diversity Probing Models

In this section, we describe how we use naturalistic settings to probe LLMs to answer questions about scalar diversity. We only use instruction-tuned models in the previous sections, namely Falcon, GPT-4, Flan-T5-xl/xxl so that they can properly follow instructions.

4.2. Scalar Diversity Probing Dataset

We use all SI instances triggered by SAs from **PVT** (Pankratz and van Tiel, 2021), **GZ** (Gotzner et al., 2018), and **RX** (Ronai and Xiang, 2022), providing a total of 152 instances. About 40 human participants answer yes or no to each prompt such as 'Mary: The problem is *hard*. Would you conclude from this that Mary thinks the problem is not *unsolvable*?' Since 40 annotators may not represent very fine-grained human judgments, we convert these datasets to binary classifications: we force models to answer `yes` or `no` given a question similar to the approach used in Ruis et al. (2024) (see the next subsection for an illustration), and consider the answers to be *yes* for instances to which at least half of the participants infer SIs, otherwise *no*.

An overview of these datasets can be found in Table 6.

Dataset	Total instance	Yes	No
PVT	50	13	37
GZ	70	19	51
RX	32	5	27

Table 6: Overview of total instances and answer counts in scalar diversity datasets.

The above datasets are relatively small. However, there are no existing large-scale datasets for scalar diversity triggered by SAs. By using three such datasets, we hope to provide a reasonable evaluation of the models' capabilities.

4.3. Scalar Diversity Probing Method

We note that some models may have inherent preferences for *yes* or *no* (e.g. more likely to answer *yes* to neutral prompts or vice versa). To debias them for such preferences, we adopt the following strategies. First, we feed the full prompt to models and record the probability for *yes* and *no* as the immediate token following the prompt and denote them as *sy* and *sn*, respectively. For models other than GPT-4, we apply three strategies to retrieve their best performance: (i) *sy*, (ii) weighted probability for *sy* (denoted as *wy*) and (iii) calibrated probability of *wy* (denoted as *cy*). For each strategy, we consider the models' answer to be *yes* if the probability is at least 0.5, *no* otherwise.

The value of *wy* is calculated as Equation 3:

$$wy = \frac{sy}{sy + sn} \tag{3}$$

cy is the calibrated probability of *wy*, which we calibrate with the probability distribution after softmax as Zhao et al. (2021) in practice. We use three neutral fillers [N/A], empty string, and [MASK] in positions of scalar claims, which give us six different neutral contexts for each template. An example is provided below, where the black part is prompt and the blue part is the answer option.

`Question:` `Imagine that your friend Mary says, "[MASK]" Would you conclude from this`

```
that Mary thinks [N/A]?
Only answer yes or no.
Answer: yes/no
```

Given these neutral prompts, we average the probabilities of *wy* in each prompt and calculate a calibration weight matrix W which calibrates the averaged *wy* to be 0.5 (i.e., neutralize it). We then use this matrix to calibrate *wy* in non-neutral prompts to derive *cy*.

4.4. Scalar Diversity Experiment and Results

Because the datasets are unbalanced, we use macro F1 as the evaluation metric. We use a logistic regression (LR) model with string-based and concept-based surprisal features (Hu et al., 2023) as a baseline.[11] When making SI predictions in one dataset, we use either of the remaining datasets or both of them to train an LR model. We report the best performance and strategies for all models in Table 7.[12]

Model	RX	GZ	PVT	Avg
LR	0.726_{RX}	0.713_{RX+PVT}	0.688_{GZ}	0.709
Falcon	0.458_{sy}	0.534_{cy}	0.578_{cy}	0.464_{cy}
Flan-T5-xl	0.835_{cy}	0.757_{cy}	$\mathbf{0.746}_{cy}$	0.779_{cy}
Flan-T5-xxl	$\mathbf{0.897}_{sy}$	$\mathbf{0.867}_{cy}$	0.726_{cy}	$\mathbf{0.816}_{cy}$
GPT-4	0.759	0.598	0.729	0.695

Table 7: Macro-F1 scores for all models across datasets. Subscripted letters for the baseline model (LR) refer to its training dataset. The average result for LR is taken over the best per-dataset results as a universal training strategy for all datasets does not exist. Subscripted letters for non-baseline models refer to strategies used. The best results across models are in bold.

Generally, we find that decoder models have unsatisfying performance in this task. For instance, Falcon performs below baseline, and GPT-4 only performs on par with the baseline and underperforms Flan-T5 models. Flan-T5-xxl shows the best results, despite having much space to reach ceiling performance. This result is slightly different from generalized implicature probing results from Ruis et al. (2024), which reports that GPT-4 outperforms Flan-T5 models.[13]

[11] We drop 4 datapoints in **GZ** in the baseline as they do not have the surprisal information.

[12] See full results for non-baseline models (except GPT-4, which we do not have access to its probability distribution by the time we conduct this experiment) in Appendix 11.8.

[13] We hypothesize this result could be for two reasons. First, their dataset contains very few direct scalar adjective comparisons. Second, we report zero-shot results instead of few-shot and chain-of-thought results, although Ruis et al. (2024) notes that these techniques do not help GPT-4 performance.

5. A Critical Appraisal of Results

5.1. 'Bad' Prompts, Good Results

Recall the following surprising result from §3.3: using simply concatenated adjective strings as inputs – which are unnatural – yields better direct scalar intensity ranking results than natural contexts with single adjectives (G&A).

From the perspective of human language processing, recent experimental linguistic works show that humans derive more robust scalar contrasts when both strong and weak terms appear salient in the same context (Ronai and Xiang, 2023a). It seems reasonable to assume that on different days, a speaker may describe a movie as *good* or *awesome*, but actually have the same degree estimate mentally (e.g. *the movie is a 7/10*) despite these terms having different semantic intensities in general. However, when using these contrastive adjectives in the same context, it is less likely that they have the same mental degree estimates. LLMs may work similarly when computing adjective intensities.

However, we also note that previous works suggest that LLMs tend to memorize information from training (Zhong et al., 2021) while our prompts are not natural at all (e.g. *good awesome wonderful great* does not exist in Google Ngram corpus). It is unclear why the 'bad' prompting method performs better than natural G&A templates. We hypothesize that attention only picks up salient discourse elements (i.e. the comparative adjectives), which is visualized in Figure 3. Natural contexts containing these adjectives, which might appear in training, may not differ as much from our prompt when computing SA representations.

5.2. Indirect Probing: Lower-bounded Absolutely, Faithful Relatively

Indirect probing shows unsatisfying absolute results compared to direct probing in §3. This finding aligns with Petroni et al. (2019), albeit the relative results remain unchanged. There are several possible explanations. First, although LLMs encode SA lexical semantics in word embeddings, they cannot reason about them in complex linguistic constructions. Second, direct probing in our study utilizes the best layer-wise information. The indirect method may not be as good for comprehensively selecting useful information across model layers. Third, although we have many prompts to get the best performance out of models, they could nevertheless be biased by the training corpus distribution if these prompts are not frequent enough (Zhong et al., 2021).

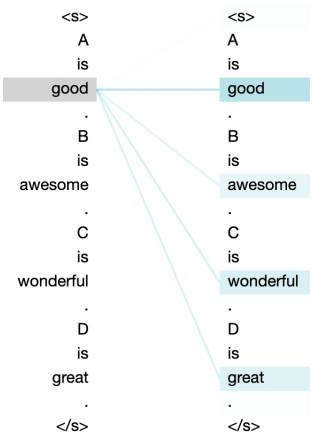

Figure 3: Attention visualization by Bertviz (Vig, 2019). Attention head 10 in the last layer of RoBERTa-b picks up *good*, *great*, *wonderful*, *awesome* when computing *good* in the context of 'A is good. B is awesome. C is wonderful. D is great.'

5.3. Scaling Does Not Explain Performance

In the lexical-semantic task (§3), the scaling law tends to hold. However, it does not explain all the patterns in the results: RoBERTa underperforms smaller BERT, Falcon underperforms smaller encoder models, and Flan-T5-xl. Lorge and Pierrehumbert (2023) also reports similar observations that BERT has better representations of scalar adverbs than RoBERTa. We hypothesize that this can be explained by training objectives. First, models that model sentence relations in pre-training tend to learn fine-grained lexical semantics easier than those without: BERT features next-sentence prediction, and Flan-T5 models sentence relations in the encoders and decoders, which forces them to learn semantics contrastively, which aligns with the findings in Merrill et al. (2024). Moreover, the fact that Falcon underperforms RoBERTa, which also lacks sentence relation training, can be explained by that the decoders only model left-hand context, which constrains them from reasoning about useful right-hand contexts. However, it is nevertheless surprising that Flan-T5-xxl underperforms the smaller Flan-T5-xl. We leave this question for future investigations.

5.4. Good Lexical Semantics Does Not Entail Good Pragmatics

In scalar diversity reasoning (section 4), GPT-4 performs worse than the much smaller Flan-T5 models, even though it is the best-performing model in the lexical-semantic task. We find that GPT-4 predicts *no* for at least 90% of the instances. Instead of forcing a *yes* or *no* answer in section 4, we try free generation here to gain further insights. Figure 4 displays one instance where humans are confident in saying *yes*, but GPT4 says *no*. We see that the model is quite conservative in making statements about implicit status.[14]

🤔 **Prompt:** Mary says: It is likely. Would you conclude from this that, according to Mary, it is not certain?

🤖 **GPT-4:** Not necessarily. The term "likely" suggests a high probability, but it does not explicitly rule out certainty. Mary could believe it is certain but is using the term "likely" to express this belief.

Figure 4: Free generation results for GPT-4 using a prompt from **GZ** without forcing `yes` or `no` answers.

Considering linguistic capabilities, we may not always expect models that are good at one task to be good at another. Weissweiler et al. (2022) also note that better syntax does not always mean better semantics. Specifically, the gap between semantic and pragmatic capabilities in our work can be explained by the fact that a weaker scalar alternative does not always give rise to the negation of the stronger one. Humans consider various factors in addition to lexical semantics such as the discourse context, the frequency of the implicature (Pankratz and van Tiel, 2021), politeness (Gotzner and Mazzarella, 2021), and other social factors. Even more confoundingly, humans also need to consider the possibility that conversational implicatures might be canceled in the subsequent discourse (Potts, 2015). Due to these complicated concerns, models need to both understand SA lexical semantics and reason pragmatically about various contextual factors to derive implicatures. Thus, good semantic models may not be good pragmatic models in reasoning about communicative intentions.

6. Conclusion

This paper has shown that LLMs of different sizes and architectures encode rich information about SA lexical semantics. Moreover, we have shown that good lexical-semantic knowledge of SAs does not always give rise to good performance in the pragmatic reasoning task about their scalar diversity. Finally, we leveraged both linguistic intuitions and model training objectives to provide an analysis for our probing results.

[14] Although conversational implicatures can be canceled (Grice, 1975) for discussion), we wish to emphasize that the paper investigates the degree to which LLM behavior matches that of an independently collected, normed, dataset. The possibility of cancellation is less evident to the experimental participants in the dataset collection.

7. Limitations

One limitation of our research is that we only use a binary classification for labels. We recognize that scalar diversity can be more fine-grained than simple *yes* or *no*. Moreover, we also note that the datasets that we use for scalar diversity task are relatively small and only constrained to SAs, which is due to the sparsity of large-scale scalar reasoning datasets[15] We encourage future works to undertake larger-scale collections of fine-grained datasets to evaluate models and train them to suit relevant needs.

Finally, some of our test datasets (e.g. ukWac) are published before the release of our models. LLMs may have already seen the test materials. As discussed in (La Malfa et al., 2024; Drinkall et al., 2024), evaluations of LLMs may be artifactually high because of the difficulties in ensuring that the test data are truly invisible during training. Future works can investigate this problem further.

8. Acknowledgements

FL is supported by Clarendon and Jason Hu studentships. JBP is supported by the Engineering and Physical Sciences Research Council (EP/T023333/1). This work is supported by research funding provided by St Catherine's College and the Faculty of Linguistics, Philology, and Phonetics at the University of Oxford. We are grateful to all the people who have offered generous help and feedback on all versions of this paper.

9. Bibliographical References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling Instruction-Finetuned Language Models. *arXiv preprint arXiv:2210.11416*.

Anne Cocos, Skyler Wharton, Ellie Pavlick, Marianna Apidianaki, and Chris Callison-Burch. 2018. Learning Scalar Adjective Intensity from Paraphrases. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1752–1762, Brussels, Belgium. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2010. "Was It Good? It Was Provocative." Learning the Meaning of Scalar Adjectives. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 167–176, Uppsala, Sweden. Association for Computational Linguistics.

Gerard de Melo and Mohit Bansal. 2013. Good, Great, Excellent: Global Inference of Semantic Intensities. *Transactions of the Association for Computational Linguistics*, 1:279–290.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Felix Drinkall, Eghbal Rahimikia, Janet B Pierrehumbert, and Stefan Zohren. 2024. Time Machine GPT. In *NAACL 2024 Findings*.

Allyson Ettinger. 2020. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Wangshu Feng, Hongbo Yu, and Xiaolin Zhou. 2021. Understanding particularized and generalized conversational implicatures: Is theory-of-mind necessary? *Brain and Language*, 212:104878.

Aina Garí Soler and Marianna Apidianaki. 2020. BERT Knows Punta Cana is not just beautiful, it's gorgeous: Ranking Scalar Adjectives with Contextualised Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7371–7385, Online. Association for Computational Linguistics.

Nicole Gotzner and Diana Mazzarella. 2021. Face Management and Negative Strengthening: The Role of Power Relations, Social Distance, and Gender. *Frontiers in psychology*, 12:602977.

Nicole Gotzner, Stephanie Solt, and Anton Benz. 2018. Scalar Diversity, Negative Strengthening, and Adjectival Semantics. *Frontiers in Psychology*, 9.

[15] Some datasets containing SI triggers such as determiners and verbs, but they are also small (Ronai and Xiang, 2023b; Van Tiel et al., 2014).

Herbert P Grice. 1975. Logic and Conversation. In *Speech acts*, pages 41–58. Brill.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jennifer Hu, Roger Levy, Judith Degen, and Sebastian Schuster. 2023. Expectations over Unspoken Alternatives Predict Pragmatic Inferences. *Transactions of the Association for Computational Linguistics*, 11:885–901.

Jennifer Hu, Roger Levy, and Sebastian Schuster. 2022. Predicting scalar diversity with context-driven uncertainty over alternatives. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 68–74, Dublin, Ireland. Association for Computational Linguistics.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are Natural Language Inference Models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Hans Kamp. 1975. Two theories about adjectives. In *Formal Semantics of Natural Language*. Cambridge University Press.

Joo-Kyung Kim and Marie-Catherine de Marneffe. 2013. Deriving adjectival scales from continuous space word representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1625–1630.

Emanuele La Malfa, Christoph Weinhuber, Orazio Torre, Fangru Lin, Anthony Cohn, Nigel Shadbolt, and Michael Wooldridge. 2024. Code Simulation Challenges for Large Language Models. *arXiv preprint arXiv:2401.09074*.

Fangru Lin, Emanuele La Malfa, Valentin Hofmann, Elle Michelle Yang, Anthony Cohn, and Janet B Pierrehumbert. 2024. Graph-enhanced Large Language Models in Asynchronous Plan Reasoning. *arXiv preprint arXiv:2402.02805*.

Yuri Lin, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the Google Books NGram corpus. In *Proceedings of the ACL 2012 System Demonstrations*, pages 169–174, Jeju Island, Korea. Association for Computational Linguistics.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2023a. We're Afraid Language Models Aren't Modeling Ambiguity. *arXiv preprint arXiv:2304.14399*.

Wei Liu, Ming Xiang, and Nai Ding. 2023b. Adjective Scale Probe: Can Language Models Encode Formal Semantics Information? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13282–13290.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Isabelle Lorge and Janet Pierrehumbert. 2023. Not Wacky vs. Definitely Wacky: A Study of Scalar Adverbs in Pretrained Language Models. pages 296–316.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany. Association for Computational Linguistics.

William Merrill, Zhaofeng Wu, Norihito Naka, Yoon Kim, and Tal Linzen. 2024. Can You Learn Semantics Through Next-Word Prediction? The Case of Entailment. *arXiv preprint arXiv:2402.13956*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Elizabeth Pankratz and Bob van Tiel. 2021. The role of relevance for scalar diversity: a usage-based approach. *Language and Cognition*, 13(4):562–594.

Carita Paradis. 1998. Degree Modifiers of Adjectives in Spoken British English.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong

Kong, China. Association for Computational Linguistics.

Christopher Potts. 2015. Presupposition and Implicature. *The handbook of contemporary semantic theory*, pages 168–202.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language Models Are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9.

Eszter Ronai and Ming Xiang. 2022. Three Factors in Explaining Scalar Diversity. In *Proceedings of Sinn und Bedeutung*, volume 26, pages 716–733.

Eszter Ronai and Ming Xiang. 2023a. Degree estimates as a measure of inference calculation. *Proceedings of the Linguistic Society of America*, 8:5537.

Eszter Ronai and Ming Xiang. 2023b. Memory Versus Expectation: Processing Relative Clauses in a Flexible Word Order Language. *Cognitive Science*, 47(1):e13227.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2024. The Goldilocks of Pragmatic Understanding: Fine-Tuning Strategy Matters for Implicature Resolution by LLMs. *Advances in Neural Information Processing Systems*, 36.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked Language Model Scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sebastian Schuster, Yuxing Chen, and Judith Degen. 2020. Harnessing the Linguistic Signal to Predict Scalar Inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5387–5403, Online. Association for Computational Linguistics.

Chaitanya Shivade, Marie-Catherine de Marneffe, Eric Fosler-Lussier, and Albert M Lai. 2015. Corpus-Based Discovery of Semantic Intensity Scales. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–493.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

Bob Van Tiel, Emiel Van Miltenburg, Natalia Zevakhina, and Bart Geurts. 2014. Scalar Diversity. *Journal of Semantics*, 33(1):137–175.

Jesse Vig. 2019. A Multiscale Visualization of Attention in the Transformer Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. The better your Syntax, the better your Semantics? Probing Pretrained Language Models for the English Comparative Correlative. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10882, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. *arXiv preprint arXiv:2302.11382*.

Bryan Wilkinson and Oates Tim. 2016. A Gold Standard for Scalar Adjectives. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2669–2675, Portorož, Slovenia. European Language Resources Association (ELRA).

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding Through Inference. In *Proceedings of the 2018 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-shot Performance of Language Models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual Probing Is [MASK]: Learning vs. Learning to Recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online. Association for Computational Linguistics.

10. Language Resource References

Cocos, Anne and Wharton, Skyler and Pavlick, Ellie and Apidianaki, Marianna and Callison-Burch, Chris. 2018. *Learning Scalar Adjective Intensity from Paraphrases*. Association for Computational Linguistics. PID https://github.com/acocos/scalar-adj/.

de Melo, Gerard and Bansal, Mohit. 2013. *Good, Great, Excellent: Global Inference of Semantic Intensities*. MIT Press. PID http://demelo.org/gdm/intensity/.

Ganitkevitch, Juri and Van Durme, Benjamin and Callison-Burch, Chris. 2013. *PPDB: The Paraphrase Database*. Association for Computational Linguistics. PID http://paraphrase.org//.

Garí Soler, Aina and Apidianaki, Marianna. 2020. *BERT Knows Punta Cana is not just beautiful, it's gorgeous: Ranking Scalar Adjectives with Contextualised Representations*. Association for Computational Linguistics. PID https://github.com/ainagari/scalar_adjs/tree/master/data.

Gotzner, Nicole and Solt, Stephanie and Benz, Anton. 2018. *Scalar Diversity, Negative Strengthening, and Adjectival Semantics*. PID https://github.com/jennhu/expectations-over-alternatives/blob/master/cross-scale/human_data/g18.csv.

Gross, Derek and Miller, Katherine J. 1990. *Adjectives in WordNet*. Oxford University Press, ISLRN 379-473-059-273-1.

Pankratz, Elizabeth and van Tiel, Bob. 2021. *The role of relevance for scalar diversity: a usage-based approach*. Cambridge University Press. PID https://github.com/jennhu/expectations-over-alternatives/blob/master/cross-scale/human_data/pvt21.csv.

Pavlick, Ellie and Rastogi, Pushpendre and Ganitkevitch, Juri and Van Durme, Benjamin and Callison-Burch, Chris. 2015. *PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification*. Association for Computational Linguistics. PID http://paraphrase.org//.

Ronai, Eszter and Xiang, Ming. 2022. *Three factors in explaining scalar diversity*. PID https://github.com/jennhu/expectations-over-alternatives/blob/master/cross-scale/human_data/rx22.csv.

Wilkinson, Bryan and Tim, Oates. 2016. *A Gold Standard for Scalar Adjectives*. European Language Resources Association (ELRA), ISLRN 691-938-401-573-9.

11. Appendix

11.1. Alternative Methods for Scale Membership Probing

Instead of summing up the weakest and strongest adjectives on each half-scale as the scale vector, and ranking the cosine similarity between each adjective and scale vectors for their scale memberships, we also experimented with computing the cosine similarity between each adjective and all adjectives on different scales, and averaging the similarity (we skip the assessed adjective itself when computing similarity with the scale it belongs to). We report results for the best-performing layers below. We observe that using the end adjectives provides better results than all adjectives (i.e. the method in the main content).

Models	DM	CD	WK
fast-text	**0.748**	0.445	**0.958**
BERT-b	$0.660_{\pm 0.009}$	$0.461_{\pm 0.020}$	$0.941_{\pm 0.012}$
BERT-l	$0.707_{\pm 0.013}$	$\mathbf{0.481}_{\pm 0.014}$	$0.938_{\pm 0.013}$
RoBERTa-b	$0.342_{\pm 0.021}$	$0.203_{\pm 0.015}$	$0.612_{\pm 0.043}$
RoBERTa-l	$0.535_{\pm 0.016}$	$0.326_{\pm 0.016}$	$0.856_{\pm 0.019}$

Table 8: Direct scale membership probing results by comparing averaged cosine similarities between the target adjective to all adjective members on different scales (in comparison to calculating the cosine similarity between the target adjective and the scale vector in the main content). Subscripted numbers are standard deviation in ten runs. The best results per dataset are in bold.

Next, we show some alternatives to mean pooling, the pooling method used in the main content. In contextualized representation computation, we used mean pooling when words are segmented. Instead of mean pooling, we also tried max/min pooling. We report the results below (we use end adjectives in this experiment). Generally, mean pooling performs better.

Models	DM	CD	WK
BERT-b	$0.815_{\pm 0.009}$/$0.807_{\pm 0.006}$	$0.783_{\pm 0.011}$/$0.775_{\pm 0.010}$	$0.997_{\pm 0.004}$/$0.996_{\pm 0.004}$
BERT-l	$0.835_{\pm 0.011}$/$0.832_{\pm 0.011}$	$0.787_{\pm 0.010}$/$0.789_{\pm 0.009}$	$0.998_{\pm 0.003}$/$0.998_{\pm 0.003}$
RoBERTa-b	$0.642_{\pm 0.009}$/$0.641_{\pm 0.008}$	$0.682_{\pm 0.007}$/$0.684_{\pm 0.006}$	$0.881_{\pm 0.023}$/$0.881_{\pm 0.016}$
RoBERTa-l	$0.727_{\pm 0.010}$/$0.723_{\pm 0.006}$	$0.721_{\pm 0.007}$/$0.720_{\pm 0.008}$	$0.954_{\pm 0.011}$/$0.950_{\pm 0.014}$

Table 9: Direct scale membership probing results using min/max pooling (left/right). Subscripted numbers are standard deviation in ten runs.

11.2. Full Table of Scalar Intensity Ranking Results

Model			DM			CD			WK		
	Dvec	Method	P-ACC	τ	ρ	P-ACC	τ	ρ	P-ACC	τ	ρ
BERT-base	DM	G&A	-	-	-	0.735	0.668	0.745	0.902	0.803	0.820
		Ours	-	-	-	0.706	0.604	0.681	**0.967**	**0.934**	**0.951**
	CD	G&A	0.646	0.431	0.509	-	-	-	0.852	0.705	0.756
		Ours	0.639	0.417	0.522	-	-	-	0.902	0.803	0.883
	WK	G&A	0.584	0.303	0.317	0.704	0.602	0.685	-	-	-
		Ours	0.584	0.295	0.360	0.694	0.581	0.659	-	-	-
BERT-large	DM	G&A	-	-	-	0.731	0.663	0.717	0.918	0.836	0.815
		Ours	-	-	-	0.727	0.650	0.703	0.902	0.803	0.873
	CD	G&A	**0.695**	**0.531**	**0.623**	-	-	-	0.918	0.836	0.868
		Ours	0.673	0.488	0.606	-	-	-	0.918	0.836	0.900
	WK	G&A	0.613	0.362	0.372	0.707	0.605	0.649	-	-	-
		Ours	0.628	0.391	0.445	0.706	0.605	0.685	-	-	-
RoBERTa-base	DM	G&A	-	-	-	0.645	0.485	0.536	0.820	0.539	0.727
		Ours	-	-	-	0.748	0.698	0.759	0.934	0.869	0.929
	CD	G&A	0.540	0.207	0.222	-	-	-	0.820	0.640	0.710
		Ours	0.648	0.431	0.558	-	-	-	0.918	0.836	0.893
	WK	G&A	0.557	0.233	0.253	0.599	0.377	0.450	-	-	-
		Ours	0.597	0.326	0.469	0.661	0.500	0.601	-	-	-
RoBERTa-large	DM	G&A	-	-	-	0.682	0.561	0.620	0.836	0.672	0.807
		Ours	-	-	-	**0.752**	**0.702**	**0.783**	0.934	0.869	0.912
	CD	G&A	0.595	0.323	0.378	-	-	-	0.820	0.639	0.690
		Ours	0.664	0.465	0.612	-	-	-	0.902	0.803	0.880
	WK	G&A	0.558	0.241	0.261	0.645	0.478	0.571	-	-	-
		Ours	0.642	0.417	0.544	0.685	0.558	0.673	-	-	-

Table 10: Full evaluation results of applying our methods to different models with additional metrics Kendall's τ and Spearman's ρ. The best results across models and methods are marked in bold. Best results using the same model and *Dvec* resource are underlined.

11.3. Construction templates and examples for scale membership indirect probing

Index	Template	Example for adjs on the same scale	Example for adjs on different scales
1	ADJ_{weak} **or even** ADJ_{strong}	*warm* or even *hot*	#*warm* or even *tasty*
2	ADJ_{weak} **if not** ADJ_{strong}	*warm*, if not *hot*	#*warm* if not *tasty*
3	ADJ_{weak}, **or even** ADJ_{strong}	*warm*, or even *hot*	#*warm* or even *tasty*
4	ADJ_{weak}, **if not** ADJ_{strong}	*warm*, if not *hot*	#*warm*, if not *tasty*

Table 11: Construction templates with which adjectives on the same scale are more likely to appear than those on different scales. # means pragmatically bad expressions.

11.4. All Templates Used for Indirect Scalar Adjective Intensity Ranking Probing

Index	weak-strong	Index	strong-weak
0	ADJ$_{weak}$ but not ADJ$_{strong}$	22	not ADJ$_{strong}$ just ADJ$_{weak}$
1	ADJ$_{weak}$ and almost ADJ$_{strong}$	23	not ADJ$_{strong}$, just ADJ$_{weak}$
2	ADJ$_{weak}$ and even ADJ$_{strong}$	24	not ADJ$_{strong}$ but just ADJ$_{weak}$
3	ADJ$_{weak}$ or even ADJ$_{strong}$	25	not ADJ$_{strong}$, but just ADJ$_{weak}$
4	ADJ$_{weak}$ although not ADJ$_{strong}$	26	not ADJ$_{strong}$ but still ADJ$_{weak}$
5	ADJ$_{weak}$ if not ADJ$_{strong}$	27	not ADJ$_{strong}$, but just ADJ$_{weak}$
6	ADJ$_{weak}$ though not ADJ$_{strong}$	28	not ADJ$_{strong}$ still ADJ$_{weak}$
7	ADJ$_{weak}$ or almost ADJ$_{strong}$	29	not ADJ$_{strong}$, still ADJ$_{weak}$
8	ADJ$_{weak}$, and even ADJ$_{strong}$	30	not ADJ$_{strong}$ although still ADJ$_{weak}$
9	ADJ$_{weak}$, or even ADJ$_{strong}$	31	not ADJ$_{strong}$, although still ADJ$_{weak}$
10	ADJ$_{weak}$, or almost ADJ$_{strong}$	32	not ADJ$_{strong}$, though still ADJ$_{weak}$
11	ADJ$_{weak}$, and almost ADJ$_{strong}$	33	not ADJ$_{strong}$ though still ADJ$_{weak}$
12	ADJ$_{weak}$ though not ADJ$_{strong}$		
13	ADJ$_{weak}$, although not ADJ$_{strong}$		
14	ADJ$_{weak}$, but not ADJ$_{strong}$		
15	ADJ$_{weak}$, if not ADJ$_{strong}$		
16	not only ADJ$_{weak}$ but ADJ$_{strong}$		
17	not just ADJ$_{weak}$ but ADJ$_{strong}$		
18	ADJ$_{weak}$ even ADJ$_{strong}$		
19	ADJ$_{weak}$ almost ADJ$_{strong}$		
20	ADJ$_{weak}$, even ADJ$_{strong}$		
21	ADJ$_{weak}$, almost ADJ$_{strong}$		

Table 12: All templates used in computing scalar adjective intensity ranking.

11.5. Held-out Prompt Engineering Results

In the main content of the paper, we report indirect probing results using the best-performing template for each model. This may incur some concerns about generalizability. Here, we choose the best-performing prompt in the other two datasets when evaluating one dataset. For GPT-4, we simply use the best template found in other models. Our general conclusion remains unchanged.

Models	DM	CD	WK
Google Ngram	0.287_1	0.075_1	0.368_1
BERT-b	0.425_1	0.066_1	0.167_1
BERT-l	0.482_3	0.102_1	0.250_1
RoBERTa-b	0.266_1	0.050_1	0.114_1
RoBERTa-l	0.463_1	0.100_1	0.286_1
Falcon	0.265_1	0.045_1	0.167_1
GPT-4	0.540	**0.273**	**0.500**
Flan-T5-xl	$\mathbf{0.544}_3$	0.177_3	$\mathbf{0.500}_3$
Flan-T5-xxl	0.515_3	0.156_3	0.364_3

Table 13: Indirect probing results for scale membership with the best prompt found in held-out datasets. The best results per dataset are in bold. Subscripted digits are the best template indexes.

Models	DM	CD	WK
Google ngram	0.312_{0}	0.222_{0}	0.442_{0}
BERT-b	0.546_{1}	0.509_{1}	0.623_{21}
BERT-l	0.529_{21}	0.518_{32}	0.738_{32}
RoBERTa-b	0.544_{14}	0.527_{19}	0.689_{14}
RoBERTa-l	0.513_{4}	0.555_{4}	0.721_{6}
Falcon	0.451_{18}	0.567_{18}	0.607_{18}
GPT-4	**0.666**	**0.739**	**0.852**
Flan-T5-xl	0.648_{16}	0.612_{16}	0.754_{17}
Flan-T5-xxl	0.564_{17}	0.582_{16}	0.770_{17}

Table 14: Indirect scalar intensity ranking table. Subscripted numbers are the best-performing template number. The best results across models are in bold.

11.6. Model Implementation Details

Flan-T5 and Falcon models are run with float16 precision. All open-source models are run on V100, A100, and M1 chip. GPT-4 is queried via OpenAI API in 2023. Temperature is set to 0 and top_p is 1 where applicable.

11.7. Prompts Used for GPT-4

In this section, we provide two prompt examples for GPT-4.

Adjective scale alignment *Do not provide explanations. Give five most likely words following the phrase: ADJ$_{weak}$ or even*

Adjective intensity ranking *Prompt: Do not provide explanations. Which of the following phrases is more natural? Answer none if they are equally unnatural. A. not just ADJ$_{weak}$ but ADJ$_{strong}$ B. not just ADJ$_{strong}$ but ADJ$_{weak}$'.*

In adjective intensity ranking, we randomly shuffle the correct answer index to avoid heuristics. The model is expected to return the first phrase as the answer for adjective pairs with unequal intensities, and none for those with equal intensities.

11.8. Scalar Diversity Results for All Non-baseline Models (except GPT-4) and Strategies

Strategy	Datasets	Falcon	Flan-T5-xl	Flan-T5-xxl
sy	RX	**0.458**	0.714	**0.897**
	GZ	0.421	0.683	0.850
	PVT	0.425	0.688	0.689
	Average	0.435	0.695	0.812
wy	RX	0.135	0.714	0.855
	GZ	0.323	0.683	0.850
	PVT	0.206	0.688	**0.726**
	Average	0.221	0.695	0.811
cy	RX	0.281	**0.835**	0.855
	GZ	**0.534**	**0.757**	**0.867**
	PVT	**0.578**	**0.746**	**0.726**
	Average	**0.464**	**0.779**	**0.816**

Table 15: Full table of scalar diversity results. The best results per model per dataset are marked in bold.