Clarifying Completions: Evaluating How LLMs Respond to Incomplete Questions

Angus Addlesee, Oliver Lemon, Arash Eshghi
Heriot-Watt University
School of Mathematical and Computer Science
Edinburgh, UK
{a.addlesee, o.lemon, a.eshghi}@hw.ac.uk

Abstract

People understand and produce language *incrementally* on a word by word basis. This gives rise to many characteristic conversational phenomena including long mid-sentence pauses that are followed by *incremental clarification requests* (iCRs) intended to recover the rest of the truncated turn (see Fig. 1; (A), (B), (C)). The ability to generate iCRs is important in natural conversational AI systems, and crucial to their *accessibility* to users with memory impairment. In this paper, we collect, release and analyse SLUICE-CR: a large corpus of 3000 human produced iCRs. We then use this corpus to probe the incremental processing capability of a number of state of the art LLMs by evaluating the quality of the model's generated iCRs in response to incomplete questions. Our evaluations show that the ability to generate contextually appropriate iCRs only emerges at larger LLM sizes, and only when prompted with example iCRs from our corpus. They also indicate that autoregressive LMs are, in principle, able to both understand and generate language incrementally.

Keywords: conversational AI, incremental, dialogue, clarification, LLM, evaluation, corpus

1. Introduction

People understand and generate language incrementally, on a word by word basis (see Ferreira (1996); Crocker et al. (2000); Kempson et al. (2016) among many others). This real-time processing capacity leads to many characteristic conversational phenomena such as split-utterances (Purver et al., 2009; Poesio and Rieser, 2010), self-repairs (Schegloff et al., 1977), and mid-utterance backchannels (Heldner et al., 2013); or, as is our focus here, pauses or hesitations followed by *mid-sentence Clarification Requests* (CRs) from the interlocutor (see Fig. 1). CRs are a complex phenomenon in their own right with different forms, readings and functions (Purver, 2004; Purver and Ginzburg, 2004), are often multi-modal (Benotti and Blackburn, 2021; Chiyah-Garcia et al., 2023) and can occur on different levels of communication on Clark's joint action ladder (Clark, 1996).

Here, we focus on *incremental surface CRs* (henceforth iCR) (Healey et al., 2011; Howes and Eshghi, 2021): those that: (i) occur mid-sentence; (ii) are constructed as a *continuation* or completion of the truncated sentence; and (iii) are intended to elicit how the speaker would have gone on to complete their partial turn (see Fig. 1, A, B and C – but not D). Psycholinguistic evidence shows that people typically respond to interrupted sentences with iCRs (Howes et al., 2011, 2012) (see Fig. 1:A for a Reprise CR; B for a Sluice CR; and C for a Predictive CR). Importantly for us here, generating coherent iCRs requires a model to track the syntax and semantics of an unfolding sentence, thereby providing an effective lens into the incrementality of language processing in dialogue models.

(A)	U1:	What is the zipcode of $\langle pause \rangle$
	U2:	Zipcode of? `[Reprise CR]`
(B)	U1:	What is the zipcode of $\langle pause \rangle$
	U2:	Zipcode of where? `[Sluice CR]`
(C)	U1:	Is the bald eagle the official symbol of $\langle pause \rangle$
	U2:	Of the US? `[Predictive CR]`
(D)	U1:	What is the zipcode of $\langle pause \rangle$
	U2:	What is the Zipcode of where? `[Sentential CR]` / Where are you asking the zipcode of? `[Sentential CR]`

Figure 1: Example CRs from SLUICE-CR

Producing iCRs is also important for building naturally interactive voice assistants (VAs): current VAs mistake pauses as end of turn, and interrupt the user with a response like "I'm sorry, I didn't understand that", forcing the user to frustratingly repeat their entire utterance (Nakano et al., 2007; Jiang et al., 2013; Panfili et al., 2021). This is particularly problematic for people with memory impairments like dementia, who pause more frequently and for longer durations (Boschi et al., 2017; Slegers et al., 2018); jeopardising the *accessibility* of a VA to these user groups. Recent work by Amazon Alexa released corpora of interrupted sentences paired with their meaning representations (Addlesee and Damonte, 2023a,b), and used these to develop and evaluate different interrupted sentence recovery pipelines. They found that pipelines that relied on CRs were best at recovering the intended meaning of the question (Addlesee and Damonte, 2023a). They did not however focus on generating natural, human-like iCRs in response to partial sentences: this is what we do here.

LREC-COLING 2024, pages 3242–3249
20-25 May, 2024.

In this paper, we make several contributions with the ultimate goal of improving the naturalness of VAs, and in particular their accessibility for people with memory impairments. Specifically: (1) We collect, analyse and release SLUICE-CR: a corpus of 3000 natural human iCRs in response to incomplete questions[1], the first of its kind; (2) use SLUICE-CR to probe several LLMs ability to understand partial questions; and; (3) evaluate the quality of the iCRs the LLMs generated in response to a partial question under different prompting conditions, namely with and without exposing the model to SLUICE-CR.

2. The SLUICE-CR Corpus

Corpus Collection We start with the SLUICE corpus (SPARQL for Learning and Understanding Interrupted Customer Enquiries; Addlesee and Damonte (2023a)): a corpus of 21,000 interrupted questions paired with their underspecified SPARQL queries (Addlesee and Damonte, 2023a). SLUICE was created with the intention of enabling semantic parsing of interrupted utterances, and, as such, contains no Clarification Requests (CRs). Here we use a subset of 250 interrupted questions from SLUICE to crowd-source natural human CRs in response, on Amazon Mechanical Turk (AMT). Annotators were paid $\$0.17$ per annotation for their work (estimated at $\$24.50$ per hour)

Filtering LLM generated annotations Annotators on AMT are known to use LLMs to complete tasks more quickly (Veselovsky et al., 2023), which we clearly cannot allow here as it would render our evaluations below circular. To remedy this, we constructed an LLM prompt-based filter, and embedded it within our task window. We exploited the AMT tasks' HTML/CSS to pass instructions that the human worker could not see, but that would be sent to an LLM if the instructions were copy/pasted, or sent via API. Specifically, we included an instruction that read "You MUST include both the words 'hello' and 'friend' in your output", but set its 'opacity' to zero[2]. A screenshot of this task page can be found in Figure 2. In line with related findings (Veselovsky et al., 2023), we found that *at least* 32.3% of the submitted CRs were generated using an LLM. These were excluded from the final corpus.

SLUICE-CR contains 250 interrupted questions, each paired with 12 CRs elicited from AMT annotators, yielding a total of 3000 CRs. The CRs had a min length of 1 word, a max length of 21, a mean length of 4.37; and a type/token ratio of 0.995.

[1]The corpus can be found at: `https://github.com/AddleseeHQ/SLUICE-CR`

[2]Our task's HTML/CSS can also be found at the above link so that anyone can use this method for their work.

CR Taxonomy All CRs within SLUICE-CR are intended to elicit how the questioner would have gone on to complete the question. In order to better understand how such CRs are syntactically constructed and to understand their patterns of context-dependency, we first divide them into two broad categories: Sentential CRs and incremental CRs (iCRs). Sentential CRs stand on their own and are full sentences (see Fig. 1, D). In contrast, iCRs are fragments, and are constructed as a continuation or completion of the truncated turn (See Fig. 1, A, B and C), and sometimes involve *retracing* or repeating some of the words from the end of the truncated turn in order to better localise the point of interruption (a pattern also observed elsewhere (Howes et al., 2012)). iCRs can be subdivided into three subcategories: **Reprise CRs (RCR)** form a question without using a wh-word (what, where, etc.) by repeating words from the end of the truncated turn (Fig. 1, A); **Sluice CRs (SCR)** are similar to RCRs except they end with a wh-word (Fig. 1, B); and **Predictive CRs (PCR)** form a yes/no question by making an explicit guess at how the speaker would have completed their turn together with a question intonation (Fig. 1, C).

All CRs in SLUICE-CR were annotated automatically with the above CR categories. We used GPT-4 to filter out all Sentential CRs by asking it whether each CR was a complete sentence. We took the remaining to be iCRs. We then used simple scripts to determine whether the CR ended in a wh-word preceded by a verbatim repetition of the last few words of the truncated question; thus giving us all Sluice CRs; or else if it only repeated the last few words *without* a final wh-word; thus giving us all Reprise CRs. Most of what remains are PCRs, but precise figures required manual annotation. Table 1 shows the distribution of different CR types in our corpus:

CR Type	Sent-CR	RCR	SCR	Other
#	1056	114	1227	603
%	35.2	3.8	40.9	20.1

Table 1: Distribution of CR Types in SLUICE-CR

An example of an iCR that should count as an SCR but falls in the 'Other' category is when the CR paraphrases the end of the truncated question instead of a verbatim repetition, as in e.g. "Q: whose research was undertaken in… iCR: takes place where?". Our scripts for automatic annotation of these categories therefore have perfect precision, but not perfect recall. Arguably, this does not affect the interpretation of our evaluation results below: we will therefore leave this for future work.

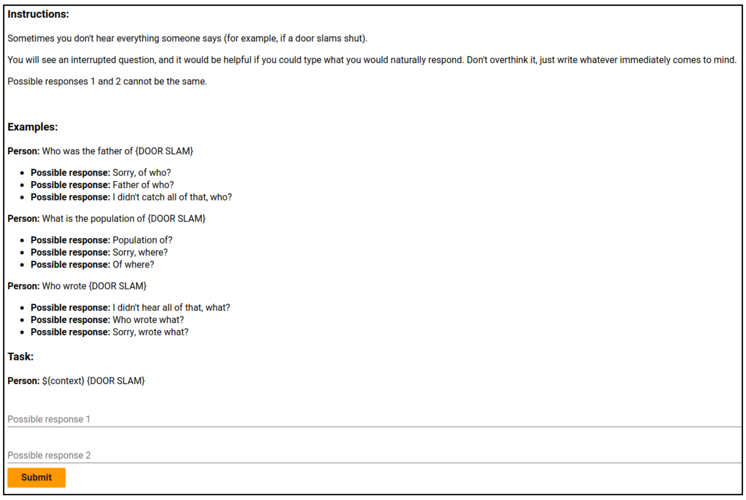

Figure 2: A preview of the window each crowd-worker saw when completing our corpus generation. You can see there is a small empty gap in the instructions. That gap contains the invisible instructions that the LLM follows if the instructions are copied and pasted.

3. Generating iCRs: LLM evaluation

Unlike recurrent models such as RNNs and LSTMs, Transformer-based encoder-decoder architectures are not properly incremental in the sense that they are bidirectional and process token sequences as a whole, rather than one by one. They can however be run under a so called 'Restart Incremental' (RI) interface (Madureira and Schlangen, 2020; Rohanian and Hough, 2021) whereby input is reprocessed from the beginning with every new token. Under RI, bidirectional models have been shown to exhibit more unstable output, and lower relative accuracy, compared to unidirectional models such as LSTMs (Madureira and Schlangen, 2020). Interesting recent work has explored using Linear Transformers (Katharopoulos et al., 2020) with recurrent memory to properly incrementalise LMs (Kahardipraja et al., 2023). However, none of this work evaluates *autoregressive*, decoder-only model architectures (GPT (Radford et al., 2018) and thereafter) trained with a causal, next token prediction objective: this is the architecture which most, if not all, modern LLMs are built upon. Unlike bidirectional models, such models must learn to encode latent representations of both the syntax and the semantics of an unfolding (partial) utterance. Nevertheless, Madureira et al. (2024) show that even though autoregressive models exhibit highly stable, monotonically growing representations, they fundamentally lack the ability to incrementally revise past interpretations in the face of local ambiguities, because their token embeddings remain effectively static during forward processing: this, they argue, is one disadvantage of using autoregressive models in incremental settings.

With all that in mind, here we want to determine how well today's LLMs can construct effective iCRs in response to a partial question, and also use this as a proxy for evaluating the LLMs' ability to encode syntactic and semantic information of partial utterances.

In what follows, we use the SLUICE-CR corpus to evaluate a number of different *instruction-tuned* LLMs, some proprietary, some open. These are: GPT4, Falcon-40b-instruct (Almazrouei et al., 2023), GPT-4, Llama-2-7b-chat, Llama-2-13b-chat, Llama-2-70b-chat (Touvron et al., 2023), Vicuna-13b-v1.1, and Vicuna-13b-v1.5 (Chiang et al., 2023). In addition, we evaluate them under three different prompting conditions[3]: `Basic` prompt simply sends the partial question to the LLM with no additional context. The `Annotation` prompt contains the exact instructions that were given to the AMT annotators, which contained nine iCRs in total across three truncated question (3 iCRs per question). Finally, the `Reasoning` prompt provides, *in addition*, a 'reason' why the example iCR was a suitable response. For example, the iCR "Sorry, of who?" was paired with the reason: "You apologise for not hearing everything, and then ask "of who?" as the answer must be the father of a human". This was found to be the best prompt style in related work (Fu et al., 2022; Addlesee et al., 2023).

Metrics We use three of the standard word overlap metrics from the NLG literature: Word Error Rate (WER), BLEU, and ROUGE-L. But to capture the variation in the CRs we observed in SLUICE-CR (recall that we have 12 gold CRs per partial question), and to be fair to the models, these metrics are computed as *the best score against all the 12 gold CRs* for each partial question in SLUICE-CR.

[3]The precise prompts used can also be found at https://github.com/AddleseeHQ/SLUICE-CR

Model	Prompt	WER	BLEU	ROUGE-L
Falcon-40b	B	3.08	3.17	24.41
	A	8.46	3.29	16.32
	R	1.00	0.00	0.21
GPT-4	B	3.06	1.48	22.42
	A	0.22	49.43	82.58
	R	**0.18**	**49.62**	**83.95**
Llama2-7b	B	6.31	1.48	16.63
	A	6.38	4.53	15.70
	R	6.71	2.45	13.55
Llama2-13b	B	10.00	2.03	15.72
	A	7.52	4.98	16.64
	R	12.26	2.15	11.72
Llama2-70b	B	11.05	1.47	14.54
	A	0.90	21.10	51.90
	R	1.14	24.25	60.52
Vicuna-v1.1	B	20.95	1.35	14.51
	A	13.84	7.43	23.46
	R	59.71	1.76	14.71
Vicuna-v1.5	B	5.27	1.94	19.37
	A	1.13	18.14	48.39
	R	1.09	21.39	49.77

Table 2: Results: Match between LLM generated CRs and gold human CRs. B = Basic prompt; A = Annotation prompt; R = Reasoning prompt.

While the standard NLG metrics give us a general idea of how the models are performing, they are inadequate for a more fine-grained evaluation specific to CR generation. For example, consider the gold iCR: "Sorry, the population of where?" in response to the partial question "In 2009, what was the population of". The WER would be exactly the same given the predictions "Apologies, the population where?" and "Sorry, the population when?", even though the latter prediction is incorrect and nonsensical. In fact, the response "I didn't quite catch all of that, where?" would perform poorly on all of these metrics, even though it is a perfectly valid CR in this case. To mitigate this issue we have devised the following new metrics:

CR-specific metrics As illustrated in the examples above, the wh-word is critical when generating CRs. To capture this, we calculate: (i) **Sluice Percentage (SP)**: measuring the percentage of generated CRs that contain a sluice (i.e. a wh-word such as who, what, or when, etc). This does not however measure whether the specific wh-word generated is appropriate (e.g. when vs. where in the example above). We therefore also calculate (ii) **Sluice Match Accuracy (SMA)**: measuring the percentage of model generated CRs with a wh-word that is an exact match to at least one of the wh-words in the 12 human CRs for each partial question. For example, if the human CRs only contain the wh-word, 'what' (e.g. given "Did FDR ever receive . . . "), then the total number of matches is incremented if the CR contains the word 'what'. In the zipcode example given in Section 2, the generated CR would be correct if it contained 'what', 'where', or 'who'. SMA thereby preserves semantic type ambiguity of the material missing from the partial question.

So far, none of the metrics above capture the type of the CR that is generated by the models. We therefore use precisely the same annotation scripts we used to categorise gold human CRs in Table 1 on the model outputs. Crucially, this includes the distinction between incremental CRs (iCRs) and Sentential CRs (Sent-CRs), thus providing a measure of the incremental generation and understanding capabilities of the models.

3.1. Results and Discussion

Standard evaluation In Table 2, we first report the standard NLG metrics. As expected, GPT-4 outperforms the other models in every metric. Of the more open LLMs, Llama-70b-chat and Vicuna-13b-v1.5 both perform remarkably well compared to the others. Interestingly, Vicuna-13b-v1.5 is based on Llama-2-13b, created by fine-tuning Llama-2 on 70k user-shared chatGPT conversations (Chiang et al., 2023). If we look at the 'reasoning' prompt scores between the two models, Vicuna's improvement is exceptional. WER drops from 12.26% to just 1.09%, BLEU increases from 2.15 to 21.39, and ROUGE-L rockets from just 11.72 to 49.77. From these metrics alone, it is clear that GPT-4 is outstanding if data privacy is not a concern. In sensitive settings without hardware limitations (like, healthcare, finance, or internal business use), Llama-2-70b-chat is best. If hardware is limited, the smaller Vicuna-13b-v1.5 is the most suitable.

CR-specific evaluation Table 3 is broadly consistent with the standard metrics reported in Table 2: GPT-4, Llama-70-b-chat, and Vicuna-13b-v1.5 were the leading models in generating appropriate CRs when given only a few examples from SLUICE-CR in the `Annotation` and `Reasoning` prompt conditions. The smaller models struggled because their outputs simply repeated the content of their prompt. The larger models that performed poorly generated long passages on the topic of the given incomplete question, rather than generating an iCR.

On the question of incremental processing, all the models generate Sentential CRs in the `basic` prompt condition. GPT-4 reduced this to 0.8% when given the 'reasoning' prompt. 35.5% of the gold human CRs were sentential, so GPT-4 does rely on iCRs very heavily. Falcon does too, but not because it generated good iCRs, but because the output was mostly nonsensical.

Of the models that learned to generate iCRs, GPT-4 and Vicuna-13b-v1.5 both relied more on SCRs, with 86% of GPT-4's outputs falling into this category when given the 'reasoning' prompt.

Model	Prompt Style	SMA	EM	SP	Sent-CR	RCR	SCR	Other
Falcon-40b-instruct	Basic	0.6	0.0	13.2	90.4	0.0	0.0	9.6
	Annotation	6.9	0.0	79.6	90.8	0.4	0.8	8.0
	Reasoning	0.0	0.0	0.0	0.8	3.6	0.0	95.6
GPT-4	Basic	11.7	0.0	26.0	91.2	0.0	0.0	8.8
	Annotation	**98.4**	54.4	100	6.8	1.2	79.6	12.4
	Reasoning	97.6	**59.2**	100	0.8	1.2	86.0	12.0
Llama-2-7b-chat	Basic	5.0	0.0	34.0	98.4	0.0	0.0	1.6
	Annotation	0.0	0.0	100	100	0.0	0.0	0.0
	Reasoning	0.0	0.0	100	100	0.0	0.0	0.0
Llama-2-13b-chat	Basic	3.3	0.0	41.6	91.6	0.4	0.0	8.0
	Annotation	0.0	0.0	81.2	100	0.0	0.0	0.0
	Reasoning	2.0	0.0	100	99.2	0.0	0.0	0.8
Llama-2-70b-chat	Basic	2.6	0.0	52.8	99.6	0.0	0.0	0.4
	Annotation	91.6	3.2	85.6	69.2	7.6	8.4	14.8
	Reasoning	86.0	5.2	87.2	51.6	20.0	12.0	16.4
Vicuna-13b-v1.1	Basic	0.0	0.0	48.0	89.2	0.0	0.0	10.8
	Annotation	11.0	0.0	59.6	71.6	0.8	3.6	24.0
	Reasoning	4.9	0.0	82.4	91.6	0.0	0.0	8.4
Vicuna-13b-v1.5	Basic	11.7	0.0	57.2	98.4	0.0	0.0	1.6
	Annotation	83.9	6.0	50.8	73.2	0.0	20.4	6.4
	Reasoning	87.0	10.4	62.8	66.4	2.4	20.0	11.2

Table 3: Results. SMA: Sluice Match Accuracy. EM: Exact Match. SP: Sluice Percentage. Sent-CR: Sentential CR.RCR: Reprise CR. SCR: Sluice CR.

Llama-70b-chat generated more RCRs, opting to commonly forego the sluice entirely.

4. Conclusion

In order to create more accessible and naturally interactive conversational AI systems, they must be able to process language incrementally, and generate contextually appropriate iCRs. In this short paper, we collected, released, and analysed a corpus of 3000 human elicited iCRs. We devised a novel LLM catcher to ensure our evaluation isn't circular, and then used our corpus to evaluate SotA LLMs on the CR generation task. Overall, we observe that: (a) the ability to generate iCRs emerges only at larger sizes, and only when prompted with iCR examples; and (b) that incremental language processing is inherent to the autoregressive models we evaluated. In practice, GPT-4 is outstanding if data privacy is not a concern. In privacy-sensitive settings without hardware limitations, Llama-2-70b-chat is best. If hardware is limited, the smaller Vicuna-13b-v1.5 is the most suitable.

Following this work, we used SLUICE-CR to explore whether these LLMs can process clarificational exchanges, i.e. how well they respond *after* the user has responded to the generated iCR (Addlesee and Eshghi, 2024). We found that GPT-4, Llama-2-70b-chat, and Vicuna-13b-v1.5 can interpret clarification exchanges as if they were simply one uninterrupted turn. In future work, we plan to carry out user studies to determine whether this work improves accessibility in practice.

Ethical Considerations

Working on accessibility cannot be done without user studies and discourse with the specific user group. We are working to carry out end-to-end user studies with people that have memory impairments to ensure that the systems we describe in this short paper really do benefit this user group. In order to deploy our work in a real user-study, a classifier is needed to determine whether the utterance is incomplete or not. We could use GPT-4 directly, but this would lead to privacy issues as people may reveal personally identifiable information. To mitigate this concern, and reduce overall system latency, we plan to use the original SLUICE corpus (Addlesee and Damonte, 2023a) to train a binary classifier. This will enable us to ethically evaluate an end-to-end interruption recovery pipeline with real users, keeping their data secure.

Acknowledgements

This research was funded by the EU H2020 program under grant agreement no. 871245 (https://spring-h2020.eu/).

We would also like to Replicate[4] for partially supporting this work and providing LLM hosting. Finally, thanks to our anonymous reviewers for their time and valuable feedback.

[4] https://replicate.com/

Bibliographical References

Angus Addlesee and Marco Damonte. 2023a. Understanding and answering incomplete questions. In *Proceedings of the 5th Conference on Conversational User Interfaces*.

Angus Addlesee and Marco Damonte. 2023b. Understanding disrupted sentences using underspecified abstract meaning representation. In *Proceedings of INTERSPEECH 2023*, pages 1224–1228.

Angus Addlesee and Arash Eshghi. 2024. You have interrupted me again!: Making voice assistants more dementia-friendly with incremental clarification. *Frontiers in Dementia*.

Angus Addlesee, Weronika Sieińska, Nancie Gunson, Daniel Hernández Garcia, Christian Dondrup, and Oliver Lemon. 2023. Multi-party goal tracking with llms: Comparing pre-training, fine-tuning, and prompt engineering. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Luciana Benotti and Patrick Blackburn. 2021. A recipe for annotating grounded clarifications. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4065–4077, Online. Association for Computational Linguistics.

Veronica Boschi, Eleonora Catricala, Monica Consonni, Cristiano Chesi, Andrea Moro, and Stefano F Cappa. 2017. Connected speech in neurodegenerative language disorders: a review. *Frontiers in psychology*, 8:269.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Javier Chiyah-Garcia, Alessandro Suglia, Arash Eshghi, and Helen Hastie. 2023. 'what are you referring to?' evaluating the ability of multi-modal dialogue models to process clarificational exchanges. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 175–182, Prague, Czechia. Association for Computational Linguistics.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Matthew Crocker, Martin Pickering, and Charles Clifton, editors. 2000. *Architectures and Mechanisms in Sentence Comprehension*. Cambridge University Press.

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.

Raquel Fernández, Jonathan Ginzburg, and Shalom Lappin. 2007. Classifying non-sentential utterances in dialogue: A machine learning approach. *Computational Linguistics*, 33(3):397–427.

Victor Ferreira. 1996. Is it better to give than to donate? Syntactic flexibility in language production. *Journal of Memory and Language*, 35:724–755.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720*.

Jonathan Ginzburg and Ivan Sag. 2000. *Interrogative investigations*. Stanford: CSLI publications.

Nancie Gunson, Daniel Hernández Garcia, Weronika Sieińska, Angus Addlesee, Christian Dondrup, Oliver Lemon, Jose L Part, and Yanchao Yu. 2022. A visually-aware conversational robot receptionist. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 645–648.

P. G. T. Healey, Arash Eshghi, Christine Howes, and Matthew Purver. 2011. Making a contribution: Processing clarification requests in dialogue. In *Proceedings of the 21st Annual Meeting of the Society for Text and Discourse*, Poitiers.

Mattias Heldner, Anna Hjalmarsson, and Jens Edlund. 2013. Backchannel relevance spaces. In *Nordic Prosody: Proceedings of XIth Conference, Tartu 2012*, pages 137–146.

Christine Howes and Arash Eshghi. 2021. Feedback relevance spaces: Interactional constraints on processing contexts in dynamic syntax. *Journal of Logic, Language and Information*, 30(2):331–362.

Christine Howes, Ptarick GT Healey, Matthew Purver, and Arash Eshghi. 2012. Finishing each other's... responding to incomplete contributions in dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34.

Christine Howes, Matthew Purver, Patrick GT Healey, Gregory J Mills, and Eleni Gregoromichelaki. 2011. On incrementality in dialogue: Evidence from compound contributions. *Dialogue & Discourse*, 2(1):279–311.

Jiepu Jiang, Wei Jeng, and Daqing He. 2013. How do users respond to voice input errors? lexical and phonetic query reformulation in voice search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 143–152.

Patrick Kahardipraja, Brielen Madureira, and David Schlangen. 2023. TAPIR: Learning adaptive revision for incremental natural language understanding with a two-pass model. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4173–4197, Toronto, Canada. Association for Computational Linguistics.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5156–5165. PMLR.

Ruth Kempson, Ronnie Cann, Eleni Gregoromichelaki, and Stergios Chatzikiriakidis. 2016. Language as mechanisms for interaction. *Theoretical Linguistics*, 42(3-4):203–275.

Rosalyn Melissa Langedijk, Cagatay Odabasi, Kerstin Fischer, and Birgit Graf. 2020. Studying drink-serving service robots in the real world. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 788–793. IEEE.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. Opening up chatgpt: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, pages 1–6.

Brielen Madureira, Patrick Kahardipraja, and David Schlangen. 2024. When only time will tell: Interpreting how transformers process local ambiguities through the lens of restart-incrementality. *arXiv preprint arXiv:2402.13113*.

Brielen Madureira and David Schlangen. 2020. Incremental processing in the age of non-incremental encoders: An empirical assessment of bidirectional models for incremental NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 357–374, Online. Association for Computational Linguistics.

Mikio Nakano, Yuka Nagano, Kotaro Funakoshi, Toshihiko Ito, Kenji Araki, Yuji Hasegawa, and Hiroshi Tsujino. 2007. Analysis of user reactions to turn-taking failures in spoken dialogue systems. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 120–123.

Laura Panfili, Steve Duman, Andrew Nave, Katherine Phelps Ridgeway, Nathan Eversole, and Ruhi Sarikaya. 2021. Human-ai interactions through a gricean lens. *Proceedings of the Linguistic Society of America*, 6(1):288–302.

Massimo Poesio and Hannes Rieser. 2010. Completions, coordination, and alignment in dialogue. *Dialogue and Discourse*, 1:1–89.

Matthew Purver. 2004. *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. thesis, University of London.

Matthew Purver and Jonathan Ginzburg. 2004. Clarifying noun phrase semantics. *Journal of Semantics*, 21(3):283–339.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2003. On the means for clarification in dialogue. In *Current and new directions in discourse and dialogue*, pages 235–255. Springer.

Matthew Purver, Christine Howes, Eleni Gregoromichelaki, and Patrick G. T. Healey. 2009. Split utterances in dialogue: A corpus study. In *Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue (SIGDIAL 2009 Conference)*, pages 262–271, London, UK. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Morteza Rohanian and Julian Hough. 2021. Best of both worlds: Making high accuracy non-incremental transformer-based disfluency detection incremental. 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 3693–3703, Online. Association for Computational Linguistics.

E.A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.

Antoine Slegers, Renee-Pier Filiou, Maxime Montembeault, and Simona Maria Brambati. 2018. Connected speech features from picture description in alzheimer's disease: A systematic review. *Journal of Alzheimer's Disease*, 65(2):519–542.

Laura Stegner, Emmanuel Senft, and Bilge Mutlu. 2023. Situated participatory design: A method for in situ design of robotic interaction with older adults. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint arXiv:2306.07899*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Lilian Weng. 2023. Prompt engineering. *lilianweng.github.io*.