How Useful is Context, *Actually*? Comparing LLMs and Humans on Discourse Marker Prediction

Emily Sadlier-Brown[a] **Millie Lou**[b] **Miikka Silfverberg**[b] **Carla L. Hudson Kam**[a]
[a]University of British Columbia [b]Independent
emily.sadlier-brown@ubc.ca

Abstract

This paper investigates the adverbial discourse particle *actually*. We compare LLM and human performance on cloze tests involving *actually* on examples sourced from the Providence Corpus of speech around children. We explore the impact of utterance context on cloze test performance. We find that context is always helpful, though the extent to which additional context is helpful, and what relative placement of context (i.e. before or after the masked word) is most helpful differs for individual models and humans. The best-performing LLM, GPT-4, narrowly outperforms humans. In an additional experiment, we explore cloze performance on synthetic LLM-generated examples, and find that several models vastly outperform humans.

1 Introduction

Natural human language utterances can be described as containing different levels of information. The most obvious is the main message or topic of the utterance, but speakers often also aim to convey information that is *about* the main message, e.g., reflecting their beliefs about or stance on the message, and its relation to other utterances in the discourse (Clark, 1996). This *pragmatic* information is an essential component of human linguistic interactions. With the recent advent of highly capable large language models (LLMs), it is also becoming a key focus in research on computational language generation.

In this paper we focus on the English adverbial discourse marker *actually*, a word with pragmatic functions. *Actually* serves to 1) highlight unexpectedness by conveying contrast and contradiction (Oh, 2000; Halliday and Hasan, 1976; Lenk, 1998; Aijmer, 2002), and 2) express reality, truth, certainty, and evidentiality (Quirk et al., 1985; Biber and Finegan, 1988; Glougie, 2016). We expect these functions to result in statistically robust properties in the surrounding linguistic context that could potentially allow an LLM to correctly predict a missing *actually*. Our experiments examine how the relative placement of contextual information affects prediction success of a variety of LLMs. Here we provide models context in the form of preceding and following utterances. We also compare the LLMs' performance to that of humans on the same task.

To evaluate prediction success, we utilize standard cloze tests, which consist of masking a word in an utterance or sequence of utterances and asking a model/human to predict the missing word. Our cloze test items are drawn from the Providence Corpus (Demuth et al., 2006) of transcribed everyday conversational speech around children, which represents an under-explored text-type in computational studies. We recruit human participants through the mTurk platform (Crowston, 2012) and compare their performance to artificial language models: GPT-3.5 and GPT-4 (Achiam et al., 2023; Brown et al., 2020), BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019) and ELECTRA (Clark et al., 2020).

Humans' guessing accuracy on cloze-type tasks has been shown to improve with increased surrounding context and to depend on the placement of that context (Rubin, 1976). We, therefore, vary the amount and type of context available in the cloze examples. Our results show that GPT-4 prediction performance is on par with (in fact higher than) human subjects' performance, echoing very recent findings (Sravanthi et al., 2024) on other pragmatic language processing tasks. Performance depends crucially on surrounding utterance context. Some amount of context is always helpful though different models and humans benefit from different placement of context. For humans and the highest-performing model, GPT-4, preceding context is more helpful than following context.

In an additional experiment, we compare cloze performance on examples sourced from the Provi-

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 231–241
August 15, 2024 ©2024 Association for Computational Linguistics

dence corpus to synthetic examples generated by GPT-3.5. Surprisingly, while human and model performance is quite similar for the corpus-sourced examples, three of the models vastly outperform human subjects on the synthetic examples. This result corroborates earlier findings that LLM-generated data differs in crucial ways from natural data (Das et al., 2024) and that LLMs demonstrate a preference for synthetic text (Panickssery et al., 2024).

Related Work Pragmatic LLM language use is an active research area. Hu et al. (2023) investigate LLM capacity for pragmatically motivated interpretations, finding that humans and models utilize similar cues for pragmatic language use. Sravanthi et al. (2024) present a benchmark of ten pragmatic language use tasks, showing that LLMs achieve near comparable performance with human subjects on many tasks. Our findings lend additional support for this result. Lake and Murphy (2023) raise two important points: 1. current models are strongly linked to text-based patterns and 2. if the aim is human-like language capacity, models should benefit from context in similar ways to humans. We address these observations by targeting (transcribed) spoken language and investigating the impact of context.

Several studies investigate LLM cloze test performance. Lai et al. (2020) compare cloze test performance of BERT and LSTM language models (Hochreiter and Schmidhuber, 1997). Pezzelle et al. (2018) compare LSTMs to human subjects on quantifier prediction in context, observing that humans benefit from broad context, while models do not. Our findings do not agree with this observation because we found that models also benefit from broad context. This possibly reflects differences between LSTMs and LLMs.

Closely related to our approach, Pandia et al. (2021) investigate LLM cloze performance on discourse markers, finding that model performance does not mirror humans on causal connectives. However, in contrast to our approach, they force models to choose a completion from among a set of 66 discourse particles. We instead allow models and humans to freely generate the masked word. We believe our approach to be preferable because artificially restricting the pool of answers makes the task substantially easier. This complicates interpretation of the experimental results and risks inflating performance for models and/or humans.

2 Methods

Data We use both corpus and synthetic data in our experiment. Our corpus data consist of 295 naturally-produced spoken utterances containing the discourse marker *actually* along with a preceding and following context utterance. Utterances are drawn from the Providence corpus (Demuth et al., 2006) of the PhonBank database (Rose and MacWhinney, 2014)[1] which consists of videotaped interactions between six children, family members and other adults in natural situations, usually in the home. All utterances in the corpus are orthographically transcribed and time-aligned with the video.

All *actually*-containing utterances in our dataset are spoken by adults, but many context utterances are child speech, and children are always present or nearby. 70% of target utterances (utterances containing *actually*) are directed at children while the remaining 30% are adult-directed, with the result that the speech style is best described as "speech around children". Preceding and following utterances could be spoken by the same speaker as the target utterance or by a different speaker. In addition to the *actually* examples, 37 distractor examples containing one of three other words, drawn from the same corpus, are included in the dataset. See Appendix A for additional details.

In our experiment with human participants (conducted using Amazon Mechanical Turk (Crowston, 2012), participants control how many cloze items they complete. Because *actually* is always a possible answer, there is a risk that participants will realize that it is the intended completion regardless of the example. To counteract this possibility, we generated 295 *synthetic examples* using GPT-3.5. These are engineered to resemble the *actually* examples, but target a variety of other words (selected based on which word in the synthetic target utterance was predicted with the lowest confidence by BERT). See Appendix A for additional details.

Cloze test To investigate the effect of surrounding context, we created four examples for every *actually* utterance, each accompanied by some combination of the preceding and following context utterance as demonstrated in Figure 1. The four conditions were: T (target utterance only); T+N (target plus next utterance); P+T (preceding utter-

[1] Publicly available at https://sla.talkbank.org/TBB/phon/Eng-NA/Providence

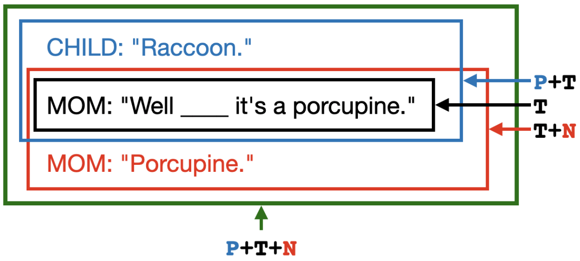

Figure 1: Cloze test (correct answer = *actually*). We investigate different degrees of contextual supervision, asking human participants and models to fill in the missing word given different combinations of: the target utterance (**T**), the following utterance (**N**) and the preceding utterance (**P**).

ance plus target); and P+T+N (preceding, target and next). In total, this results in 2360 examples of which 1180 are *actually* examples and the rest synthetic.

We asked human subjects to fill in the "5 most likely English words" for the target utterance. We chose to ask for five answers because there are normally several reasonable answers for any given example, and overly limiting the number of responses (say, to one or two) would mean our dataset would fail to include words that participants might believe are equally likely. On the other hand, asking for more than five responses could make the task too difficult and time-consuming. All examples and all conditions were randomized such that any given participant received a random mix of *actually* and synthetic examples across a random distribution of conditions. Each example was only completed by one participant (although participants were free to complete as many items as they wanted). In order to ensure the validity of the results, we limited participants to those completing the task in an English-speaking country, and we removed responses which contained high proportions of repeated answers or answers that occurred in the prompt. In total, there were 255 mTurk participants. Given that there were 2360 examples in total, each annotator annotated 8.1 examples on average. See Appendix C for additional details.

We experiment with two types of LLMs: 1) Encoder-only large language models: BERT, RoBERTa and ELECTRA, and 2) Generative large language models: GPT-3.5 and GPT-4. For encoder-only models, which are trained on a masked language modeling objective, we can straightforwardly frame the cloze task as masked prediction. We extracted the five most probable words in the given context. For generative models (GPT-3.5 and GPT-4), we cannot directly frame the cloze task as masked prediction. Instead, we prepared a prompt which asks the model to predict the missing word in the example (see Appendix B for details on the prompts and generation process).

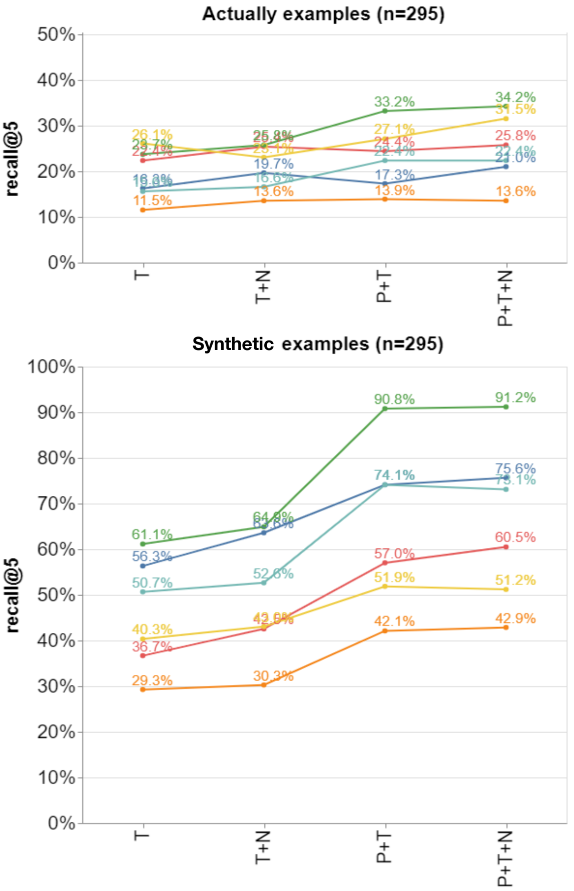

Figure 2: Results (recall@5) on cloze tests for: ● BERT, ● ELECTRA ● RoBERTa, ● GPT-3.5, ● GPT-4 and ● mTurk. We present results for *actually* examples in the top panel and synthetic examples in the bottom panel. Results are presented for all context types: T, T+N, P+T and P+T+N.

For both models and humans, we evaluate recall@5, i.e., we computed how often the correct word is found among the five completions.

3 Results

Experimental results for human subjects and LLMs are presented in Figure 2. The information is shown in tabular format in Table 1.

***Actually* examples** Across all settings, humans do well in comparison to most models on *actually* prediction but GPT-4 outperforms humans in all settings apart from T, where only the target utterance is provided as context. In general, context is

ACTUALLY EXAMPLES

Model	T	T+N	P+T	P+T+N
Human	26.1	23.1	27.1	**31.5**
GPT-3.5	16.0	16.6	**22.4**	**22.4**
GPT-4	23.7	25.8	33.2	**34.2**
BERT	16.3	19.7	17.3	**21.0**
ELECTRA	11.5	13.6	**13.9**	13.6
RoBERTa	22.4	25.4	24.4	**25.8**

SYNTHETIC EXAMPLES

Model	T	T+N	P+T	P+T+N
Human	40.3	42.6	**51.9**	51.2
GPT-3.5	50.7	52.6	**74.1**	73.1
GPT-4	61.1	64.9	90.8	**91.2**
BERT	56.3	63.6	74.1	**75.6**
ELECTRA	29.3	30.3	42.1	**42.9**
RoBERTa	36.7	42.5	57.0	**60.5**

Table 1: Cloze test results (recall@5) for *actually* and synthetic data.

helpful for both models and humans, but the effect of quantity and placement differs across humans and different models. Purely based on the target utterance (T), human recall@5 (0.291) narrowly beats both GPT-4 (0.237) and RoBERTa (0.224), while BERT, GPT-3.5 and ELECTRA deliver substantially lower performance. In the presence of additional supervision in the form of the following utterance (T+N), the general picture remains largely unchanged, although RoBERTa and GPT-4 now narrowly outperform human annotators. Providing the previous utterance as context instead of the following one (P+T), results in a substantial boost in recall for GPT-4 (+0.08) and GPT-3.5 (+0.06), while changes for humans and other models remain small. In this setting, GPT-4 clearly outperforms all other models and humans. Finally, full context (P+T+N) delivers the best performance for GPT-4. In this context, humans' recall@5 gains +0.04 and ends up close to, though still slightly below, GPT-4.

Synthetic examples Performance for all models and human participants is higher on synthetic examples than *actually* examples. In the baseline setting, seeing only the target utterance (T), human participants do 14%-points better on synthetic examples and GPT-4 does 37%-points better. Overall, three of the models– GPT-4, BERT and GPT-3.5– very clearly outperform human participants on the synthetic examples. Given additional context, we see the same overall trend as for *actually* examples: providing the next utterance (T+N) marginally improves performance whereas the preceding utterance (P+T) leads to large improvements for all model types and humans. Providing both context utterances (P+T+N) delivers small additional improvements for RoBERTa and GPT-4 and results in minor degradation for humans and GPT-3.5 compared to P+T. Overall, model performance is extremely high, with GPT-4 achieving 91% recall in the P+T+N setting, which is a whole 40%-points higher than human performance. At the same time, RoBERTa seems to deliver the most humanlike performance. BERT surprisingly delivers far stronger performance than RoBERTa even though the model architectures are very similar.

4 Discussion

Both models and humans achieved moderate success in the cloze task for *actually* and notably higher rates of success on the synthetic examples. Humans' ability to predict *actually* in a variety of contexts generally fell within the range of accuracy of the best-performing models. This suggests that the models were able to generalize to the type of speech from which we drew our examples (largely comprising child-directed speech). GPT-4, in particular, outperformed humans. This is not wholly unexpected given that it is one of the largest LLMs to date.

Context is generally helpful. The preceding context utterance seems to be crucial—the best performance is always achieved either in the condition P+T or P+T+N. For the *actually* examples, the generative models GPT-3.5 and GPT-4, along with humans, derive the largest gains in accuracy from added context, while other models saw smaller improvements. This might indicate that a generative training objective better helps models condition on contextual information in a human-like way compared to a masked language modeling objective. On the other hand, GPT-4 outperforms humans, so it is in fact using context more effectively than our human subjects. Interestingly, although RoBERTa sees little improvement with context, it nevertheless outperforms GPT-3.5 in all conditions.

Models massively outperform humans on synthetic data. Our natural examples proved much more difficult than the synthetic ones. GPT-4, in particular, achieved a stunning 91.2% recall@5 in the synthetic P+T+N condition. As evidenced by far lower human performance at 51.2%, this is unrealistically high. In fact, on synthetic data, humans were outperformed by all models apart from ELECTRA, a pattern of results which stands in stark contrast with the results from the natural data. We hypothesize that synthetically produced examples, unlike real ones, strongly reflect the distribution learned by GPT-3.5 which was used to generate those examples. This makes a cloze task less of a test of generalizability and more of a test of overfitting to training data. Therefore, models in fact benefit from a narrow understanding of language on synthetic data, which makes it less surprising that they outperform humans. The effects of context in the synthetic examples are also much more pronounced and in this setting both humans and all models improve with added context in contrast to the harder *actually* examples.

5 Conclusion and Future Work

In our data, the performance of models and humans fell within the same range. This suggests that models, especially those performing closest to humans, are able to predict the occurrence of the pragmatically-sensitive item *actually*, possibly based on similar aspects of the surrounding context as humans. However, important differences emerged between models and humans in overall accuracy, use of context, and the effect of context location, suggesting differences in how, and how effectively, models and humans utilize context. Our work raises some important questions: where models outperform humans, are they picking up on contextual cues that humans are not sensitive to? If so, what are these cues? Is outperforming humans a desirable goal, or is emulating human behavior—interpreting contextual cues in a human-like way, including failing to make use of certain cues even if they might be useful—more aligned with the goals of language modeling? Finally, our experiments on synthetic examples demonstrate a stark contrast between LLM performance on natural and synthetic data. Consequently, we urge caution when using synthetic data in experiments, especially when comparing human and LLM performance.

6 Limitations

Limitations of working with human subjects There are several limitations related to the human cloze experiment. Although we limited mTurk participants to those performing the task in English-speaking countries, we do not know whether participants are native English speakers, nor do we know their level of English proficiency. In fact, we expect many participants will have been second language speakers, meaning that the results might not carry over to a native English-speaking population. The presumed language background variability of the human participants also reduces comparability between our human results and model results, since the models are trained only on English. Furthermore, as in all experiments involving human subjects, participants' understanding of the task, attention to the task and motivation to follow the instructions cannot be controlled. To mitigate these potential issues as much as possible, we limited participants by experience and approval rating and we automatically filtered out responses that bore the hallmarks of inattention: those that exceeded a pre-determined proportion of repeated answers or answers copied directly from the prompt (see **Appendix C** for details).

Limitations of the experimental design Given that our human participants were allowed to complete as many examples as they wanted, there was a risk that participants who completed more examples would figure out that *actually* is often a possible answer. In an ideal world, each human subject would complete a single example, which would entirely eliminate the effect of seeing multiple *actually* examples, but unfortunately, limiting the task to one example per participant would make it not worthwhile for most participants. We attempted to mitigate the potential effect of seeing multiple *actually* examples by adding the synthetic examples as described in **Methods** and **Appendix A**. In addition, we performed a post-hoc analysis evaluating whether participants who completed more examples in fact guessed *actually* more often. Figure 3 shows the number of *actually* answers as a function of the total number of cloze test items completed by the participant. As the regression line in the plot demonstrates, participants who completed more examples did tend to guess *actually* more frequently than others. However, the effect is moderate. Moreover, as Figure 4 demonstrates, most of our answers

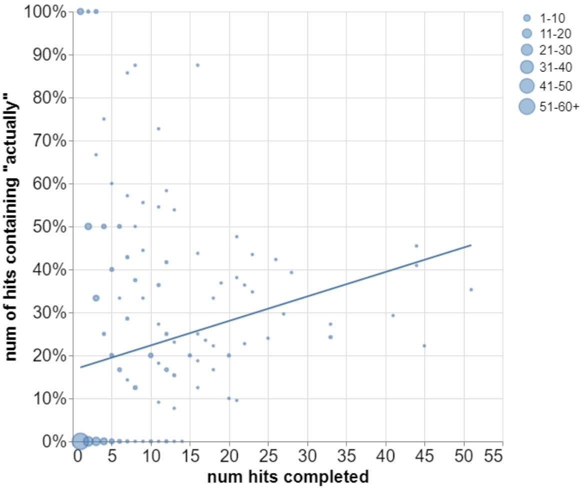

Figure 3: The proportion of correctly identified *actually* answers as a function the number of examples completed. The regression line shows that the proportion of *actually* guesses does tend to increase as the number of completed examples increases.

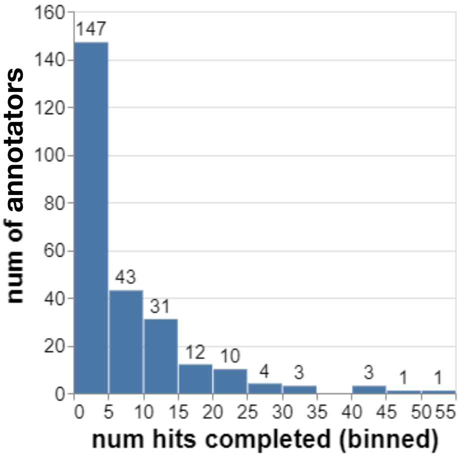

Figure 4: The distribution of the number of completed examples among participants. 58% of participants completed maximally five examples and 87% completed maximally fifteen.

come from participants who completed very few examples. Consequently, most of our correct *actually* responses come from participants who completed very few examples simply because there are far more such participants. This means that the data is unlikely to be very biased on the whole.

Caveat concerning LLMs Finally, there is one major limitation related to the LLMs: while we do not believe that the LLMs would have been exposed to the Providence corpus during their training process, we were only able to check this for BERT, RoBERTa and ELECTRA. For the GPT models it is impossible to know for certain. If these were exposed to the Providence corpus, this might inflate their performance on the *actually* cloze tests.

Acknowledgements

The authors would like to thank Vered Shwartz for suggesting the mturk experiment and for her valuable input in the experiment design phase. The present work was supported by UBC Language Sciences Seed Funding. In addition, we acknowledge the following grant support for PhonBank: NIH-NICHHD RO1-HD051698.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Karin Aijmer. 2002. English discourse particles. *English Discourse Particles*, pages 1–315.

Douglas Biber and Edward Finegan. 1988. Adverbial stance types in english. *Discourse Processes*, 11(1):1–34.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

H. Clark. 1996. Using language. *Cambridge*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Kevin Crowston. 2012. Amazon mechanical turk: A research tool for organizations and information systems scholars. In *Shaping the Future of ICT Research. Methods and Approaches: IFIP WG 8.2, Working Conference, Tampa, FL, USA, December 13-14, 2012. Proceedings*, pages 210–221. Springer.

Debarati Das, Karin De Langis, Anna Martin, Jaehyung Kim, Minhwa Lee, Zae Myung Kim, Shirley Hayati, Risako Owan, Bin Hu, Ritik Parkar, et al. 2024. Under the surface: Tracking the artifactuality of llm-generated data. *arXiv preprint arXiv:2401.14698*.

Katherine Demuth, Jennifer Culbertson, and Jennifer Alter. 2006. Word-minimality, epenthesis and coda licensing in the early acquisition of english. *Language and Speech*, 49(2):137–173.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jennifer Robin Sarah Glougie. 2016. *The semantics and pragmatics of English evidential expressions: the expression of evidentiality in police interviews*. Ph.D. thesis, University of British Columbia.

M. A. K. Halliday and R. Hasan. 1976. *Cohesion in English*. Longman, London.

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

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Yi-An Lai, Garima Lalwani, and Yi Zhang. 2020. Context analysis for pre-trained masked language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3789–3804.

Brenden M Lake and Gregory L Murphy. 2023. Word meaning in minds and machines. *Psychological review*, 130(2):401.

Uta Lenk. 1998. *Marking discourse coherence: Functions of discourse markers in spoken English*, volume 15. Gunter Narr Verlag.

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*.

Sun-Young Oh. 2000. Actually and in fact in american english: a data-based analysis. *English Language & Linguistics*, 4(2):243–268.

L Pandia, Y Cong, and A Ettinger. 2021. Pragmatic competence of pre-trained language models through the lens of discourse connectives. In *Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL)*.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*.

Sandro Pezzelle, Steinert-Threlkeld Shane, Raffaella Bernardi, Szymanik Jakub, et al. 2018. Some of them can be guessed! exploring the effect of linguistic context in predicting quantifiers. In *ACL 2018: The 56th Annual Meeting of theAssociation for Computational Linguistics Proceedings of the Conference Vol. 2 (Short Papers)*, pages 114–119. Association for Computational Linguistics.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Y. Rose and B. MacWhinney. 2014. The phonbank project: Data and software-assisted methods for the study of phonology and phonological development. In J. Durand, U. Gut, and G. Kristoffersen, editors, *Handbook of corpus phonology*, pages 380–401. Oxford University Press, Oxford.

David C Rubin. 1976. The effectiveness of context before, after, and around a missing word. *Perception & Psychophysics*, 19:214–216.

Settaluri Lakshmi Sravanthi, Meet Doshi, Tankala Pavan Kalyan, Rudra Murthy, Pushpak Bhattacharyya, and Raj Dabre. 2024. Pub: A pragmatics understanding benchmark for assessing llms' pragmatics capabilities. *arXiv preprint arXiv:2401.07078*.

A Test Item Creation

Actually items *Actually* examples are drawn from the publicly-available Providence corpus (Demuth et al., 2006), which consists of video-taped spoken English interactions between children (n=6) and their parents and sometimes others. The children ranged in age from 1 year to 4 years. The data was collected in the form of one hour video recordings, collected across an average of 61 sessions per child over the years 2002-2005. The corpus is transcribed in written English, and these transcriptions are time-aligned with the videos. The mean(sd) lengths of example utterances (in number of words) were: for target utterances, 11.1(7.1); for preceding utterance, 5.3(4.8); for following utterance, 5.9(4.9).

We located all utterances containing tokens of *actually* (n=844). As part of a larger project, the *actually*-containing utterance, the utterance before and the utterance after were annotated for a suite of linguistic and other behavioural features (e.g. activity). For the present analysis, the example set was filtered to exclude examples in which the child spoke the target (*actually*-containing) utterance, examples in which there was no behavioural information (e.g. due to speakers being off-camera) and examples in which the utterance before and/or after was missing. Three examples of *actually* test items can be seen in Figure 5.

Synthetic items The synthetic examples were generated by GPT-3.5-turbo-1106 with temperature set to 0.7. The GPT-3.5 prompt is given in Figure 6. To generate a set of synthetic examples that resembled the *actually* examples, we provided GPT with the following variables and populated them with randomly selected values from distributions similar to those of the *actually* dataset:

```
MOM: "Hee hee hee."
MOM: "I think it'd  be a great idea if we
     ____ went to sleep tonight and stayed
     asleep all night."
MOM: "Wouldn't that be great?"
******
CHILD: "More there and there."
MOM: "Nope  we ____ don't need more there."
MOM: "The secret to good wrapping is not to
     use too much tape I think."
******
OTHER ADULT: "You have more of the last
             paper?"
MOM: "It's ____ in my car."
MOM: "Let me get it out."
```

Figure 5: Example test items

- 2 speakers (e.g., MOM, DAD, CHILD)
- emotion of speaker2 (e.g., neutral, happy)
- activity of speaker2 (e.g., playing, conversing)
- location in the house (e.g., living room, kitchen)
- age of the CHILD if CHILD was selected as a speaker (e.g., 15 months, 2 years old)
- "do" or "do not" add a discourse marker to the utterance

Out of 295 examples, 50 had to be manually edited so that the format was correct. A common error was that GPT-3.5-turbo added a fourth utterance in the synthetic example when the instructions only asked for three. After this light editing process, the mean(sd) lengths of the synthetic example utterances (in number of words) were: for target utterances, 8.7(3.1); for preceding utterance, 6.0(2.1); for following utterance, 6.5(2.8).

Once synthetic examples were finalized, we simulated masking each word in the utterance and used BERT to make one word predictions for all masked instances. The word with the the lowest probability among BERT's predictions was masked for the cloze test. In Figure 7, we give three examples of synthetic items (the masked words are *Um*, *I'm* and *bedtime*, respectively).

B LLM Prompt Details

LLMs were asked to predict one example at a time to ensure their responses were not influenced by any text from other examples. A synthetic example was provided so that the LLM responded in the correct json format. The prompt is shown in Figure 8.

C mTurk Experiment Details

We recruited human participants via Amazon Mechanical Turk (Crowston, 2012). Each participant was paid $0.10 per HIT (which consisted of one test question). mTurk participants qualified for the task if they:

- had a HIT approval rate over 95%
- had a number of HITs approved > 500
- were located in one of the 35 most populous countries in which English is an official or predominant language, according to https://en.wikipedia.org/wiki/List_of_countries_and_territories_where_English_is_an_official_language

In addition to the above qualifications, mTurk participants were required to pass a qualification test consisting of three fill-in-the-blank questions of the form given in Figure 9.

The instructions in the HIT were: "Fill in the blank with the 5 most likely English words. No duplicates."

To limit the number of examples a participant could complete, and to prevent participants from completing the same example in more than one condition, we implemented the following:

- examples were released in 100-example batches and participants who completed HITs in one batch were unable to complete the task in another batch
- in one batch, there was only one type of context per example (e.g., if T+N of example1 is in Batch1, then T, P+T, P+T+N of example1 will be in a different batch)
- only one set of responses was collected per example

Once answers were collected, a quality check was conducted to filter out poor responses:

- no responses consisting of repeated answers (e.g., "fun", "fun", "fun", "fun", "fun")
- no one-character answers (e.g., "r", "e", "a", "1", "y")

~~~~
You are a screenwriter who is writing a conversation between two people. Speaker1
and Speaker2 are {location} and Speaker2 {act_cat}.{child_msg} Create a three turn
conversation, make up an action for Speaker2 before the Speaker2 utterance and
{discourse_marker_msg} add a discourse marker frequently found in conversations in
Speaker2's utterance. Follow the specified FORMAT. In the FORMAT, more detailed
instructions will be provided between the delimiter triple backticks, ```.

###FORMAT START###
Speaker2 ```The scene will be provided and should be printed exactly the same in
the output```.
Speaker1: "```Says something brief to Speaker2 that fits the scene.```"
Speaker2 ```Make up an action Speaker2 is doing before the next utterance.```
Speaker2: "```Says something brief to Speaker1.```"
Speaker1: "```Responds to Speaker2.```"
###FORMAT END###

###EXAMPLE START###
INSTRUCTIONS:
- Speaker1 = MOM
- Speaker2 = DAD
- Scene = DAD is changing diaper.
- Emotion = DAD is feeling panicked.
- Location = living room
- Do add a discourse marker frequently found in conversations in Speaker2's
utterance

OUTPUT:
DAD is changing diaper.
MOM: "Did you remember to use baby powder?"
DAD looks up quickly.
DAD: "Huh, what do you mean baby powder? I didn't know she needs it. Is she sick?"
MOM: "No calm down. She's fine."
###EXAMPLE END###

Now it's your turn.

INSTRUCTIONS:
- Speaker1 = {s1}
- Speaker2 = {s2}
- Scene = {s2} {act_cat}.
- Emotion = {s2} is feeling {emotion}.
- Location = {location}
- {discourse_marker_msg} add a discourse marker frequently found in conversations in
Speaker2's utterance

OUTPUT:
~~~~

Figure 6: LLM prompt template for generation of synthetic cloze examples. The LLM generates its answer after `OUTPUT:`.

```
OTHER ADULT: "Hey, do you have the car
              keys?"
MOM: "_____, let me see... Oh, here they
     are."
OTHER ADULT: "Great, let's head out."
******
DAD: "Smells delicious in here, what's for
     dinner?"
MOM: "_____ making your favorite, spaghetti
     and meatballs."
DAD: "That sounds amazing, I can't wait to
     eat!"
******
OTHER ADULT: "I didn't expect to see you
             here."
MOM: "Yeah, I wanted to tidy up a bit
     before _____."
OTHER ADULT: "Well, that's nice of you to
             do."
```

Figure 7: Example synthetic items.

- type-to-token ratio (TTR) was =<0.5
- answer-to-sample-ratio (ATSR) was =>0.1

TTR and ATSR were calculated by participant. We found that low TTR indicates the participant did not meaningfully complete the task due to having repeated a high proportion of words across different HITs. High ATSR indicates that the participant repeatedly used words found in the example utterances as responses, rather than choosing words that fit the context as per task instructions. Examples with rejected answers were put in new batches for another round of mTurk completions. In total, we completed three rounds before all examples received accepted answers.

~~~
You are a fluent English-speaker. In a conversation between two people, there will
be a blank denoted by _____.

TASK:
  1. Read the text between the characters ```
  2. Determine the 5 most likely English words in place of the blank _____; NO
     DUPLICATES IN THE LIST
  3. Create a JSON object like the following: {"word1": "one_word_only",
     "word2":"one_word_only", "word3": "one_word_only", "word4": "one_word_only",
     "word5":"one_word_only"}
  4. Your response should only contain the JSON object.

EXAMPLE:
```MOM likes to _____ cookies.``` A good response is {"word1": "eat", "word2": "make",
"word3": "buy", "word4": "decorate", "word5": "bake"}

TEXT:
~~~
{example}
```

OUTPUT:
```

Figure 8: LLM Prompt template for Recall@5. The variable `{example}` is replaced by a cloze test example. The LLM generates its answer after `OUTPUT:`.

```
DAD is cooking.

MOM: "Did you add salt?"

DAD is standing.

DAD: "Yeah, _____ course."

MOM: "Oh good."
```

Figure 9: One of our qualification questions for mTurk annotators. The correct answer here is *of*.