Investigating the use of pragmatic inferences and the predictive power of language models in sentence processing

Dingyi Pan & Andrew Kehler
Department of Linguistics
UC San Diego
La Jolla, CA 92093, USA
{dipan, akehler}@ucsd.edu

Abstract

Large language models (LLMs) have demonstrated competence in various syntactic and pragmatic tasks. However, it is unclear whether LLMs can leverage their pragmatic inference abilities in syntactic processing. In addition, it remains unclear whether improvements in pragmatic abilities with larger model sizes also translate to better prediction of human reading times. In this study, we first test whether LLMs can use pragmatic inferences to make attachment decisions of ambiguous relative clauses in English. We then examine their abilities to predict human reading time. Our results suggest that larger and more recent models are in fact able to use their pragmatic inference ability in downstream syntactic processing, but among those models that can, larger and instruction-tuned models do not always have the best psychometric power for predicting reading time.

1 Introduction

Since the introduction of large language models (LLMs), researchers have sought to evaluate their linguistic abilities in a variety of respects. Whereas early results revealed success in certain areas pertaining to syntactic competence (Linzen et al., 2016; Futrell et al., 2019; Warstadt et al., 2020, *inter alia*), pragmatics has historically proven to be more challenging for LLMs, plausibly due to the fact that it requires the integration of inference, real-world knowledge, and contextual cues (Chang & Bergen, 2024). That notwithstanding, whereas early transformer models (e.g., GPT-2) struggle with certain pragmatic tasks such as inferring discourse coherence relations between clauses (Beyer et al., 2021), more recent LLMs have demonstrated the ability to draw pragmatic inferences across a range of phenomena (Hu et al., 2023, *inter alia*). Central to the present study are findings that show that humans not only draw coherence-based pragmatic inferences but also leverage them in sentence processing tasks such as RC attachment (Rohde et al., 2011; Hoek et al., 2021). Although various studies have demonstrated the syntactic and pragmatic abilities of LLMs independently, to our knowledge none have shown that recent models are able to leverage their pragmatic inference abilities to guide syntactic processing (cf. Davis & van Schijndel (2020a)).

Larger LLMs tend to perform better than smaller ones in a variety of pragmatic tasks. Specifically, a sharp increase in performance is observed at the 1-billion-parameter threshold, whereas little difference is observed among models above and below this size (Hu et al., 2023). In contrast, larger models do not necessarily have better abilities to predict human reading time when difficult syntactic constructions are processed (Oh & Schuler, 2023; Shain et al., 2024; Kuribayashi et al., 2024). Here, we ask whether this generalization holds for sentence processing tasks that rely on pragmatic inferences, i.e., whether or not larger models exhibit better psychometric predictive power for such tasks due to their enhanced pragmatic abilities. To address this question, we examine the attachment bias in English relative clauses (RCs) with potentially ambiguous attachment sites, which has been shown to be sensitive to pragmatic inferences drawn earlier in the sentence (Rohde et al., 2011; Hoek et al., 2021).

English has a default low attachment bias for ambiguous RCs (Frazier & Clifton, 1996; Carreiras & Clifton, 1999, *inter alia*). For instance, although the RC in (1a) can be used to modify either noun phrase (NP) in the complex NP structure "the children of the musicians," English speakers are more likely to judge that it is the musicians, rather than the children, who are arrogant and rude. This preference for the low attachment site is commonly accounted for by the structurally shorter distance between the RC and the low NP as compared to the high NP (Frazier & Fodor, 1978).

1. (a) Melissa babysits the children of the musicians who are arrogant and rude.
   (b) Melissa detests the children of the musicians who are arrogant and rude.

However, there is reason to expect that this bias might shift in examples like (1b), despite the only difference with (1a) being the matrix verb. The rationale unfolds in the following three steps. First, implicit causality (IC) verbs such as "detest" in (1b) create a strong expectation that an explanation of the event it describes will ensue (Kehler et al., 2008). Second, the immediately following RC can be employed by the speaker to convey an explanation. Lastly, "detest", an object-biased IC verb, carries a strong bias toward an ensuing explanation to be about its direct object–"the children" in (1b)–which is the high attachment site for the RC. Therefore, if comprehenders are able to integrate these three pieces of pragmatic information and use them to inform a syntactic processing decision, we would expect a greater bias toward high attachment in object-biased IC contexts than in nonIC contexts, the latter of which create neither a strong expectation for an upcoming explanation nor a strong expectation that any such explanation would be about the direct object. These predictions were confirmed in an offline sentence completion task and a self-paced reading task (Rohde et al., 2011): On the production side, human participants were more likely to produce an explanation-providing RC when an IC verb was used than when a nonIC verb was used, which in turn led to more modifications of the high NP. Further, the results of the self-paced reading task provided even stronger evidence: participants incorporated this pragmatic inference to predict the attachment site, even before they encountered the full RC. Note that this pragmatic inference about the explanation is not mandated by any syntactic relationship or other linguistic felicity requirement that applies to the sentence. This can be seen in (2), which is likewise perfectly felicitous even though it will typically not convey an inference analogous to that in (1b), i.e., that causally relates Melissa's detesting to the place where the children live.

2. Melissa detests the children of the musicians who live in La Jolla.

While humans use pragmatic inference during sentence processing, it is not clear whether LLMs show similar evidence of being able to integrate and use pragmatic information in making syntactic attachment decisions. In addition, although studies have examined the correlation between the next-word log probability and human reading in syntactic tasks, it is also unclear whether results concerning the predictive power of language models can be generalized to sentence processing tasks that are driven by pragmatic inferences.

In Experiment 1, we test whether the more recent LLMs demonstrate the ability to use this pragmatic inference in making RC attachment decisions. In Experiment 2, we examine whether these models can quantitatively predict human reading times for these sentences.

2 Experiment 1: Deciding relative clause attachment site

Previous studies with earlier non-transformer-based models have shown mixed results in terms of the models' default attachment preference for ambiguous RCs. For instance, Davis & van Schijndel (2020b) showed that recurrent neural network (RNN) LMs exhibit a low attachment preference for ambiguous relative clauses in English. This cannot be taken as evidence that models have learned the human-like syntactic preference, however, since the LMs also exhibited a low attachment bias for languages that are known to have a default high attachment bias, such as Spanish. Likewise, Kamerath & De Santo (2025) showed that models also failed to show the human-like attachment bias of ambiguous RCs in Italian, while the English results were mixed in terms of attachment bias and the effect of lexical

items on these preferences. More relevant to the current inquiry, there has been no evidence supporting neural LMs' sensitivity to pragmatic inferences that should, in theory, alter the syntactic attachment bias. Davis & van Schijndel (2020a) found that long short-term memory networks (LSTMs) failed to use the causal relation triggered by IC verbs in RC attachment. Interestingly, the larger transformer-based GPT-2 XL also did not show the effect of IC verbs in syntactic processing, even though it was sensitive to IC bias in other tasks, such as pronoun resolution.

Therefore, in this experiment, we use sentence pairs like those in (1) to test more recent LLMs in their ability to use pragmatic inferences, as triggered by the IC verbs, to guide the attachment decision of ambiguous relative clauses.

2.1 Methods

2.1.1 Models

We evaluated the performance of five models that vary in size and training methods: two pre-trained base models from the Llama model families, namely Llama-3.2-1B (1.23B parameters) and Llama-3.2-3B (3.21B parameters), the instruction-tuned version of the two models, Llama-3.2-1B-Instruct and Llama-3.2-3B-Instruct, and the GPT-2 model (Radford et al., 2019, 124M parameters). The two Llama base models differ minimally from the two Instruct models, primarily in the lack of supervised fine-tuning and reinforcement learning with human feedback (RLHF). All models were accessed through Hugging Face.

2.1.2 Stimuli

Sixty pairs of sentences following the format of (3) were created, where one sentence in each pair has an IC verb as the matrix predicate and one has a nonIC verb. All verbs were taken from Rohde et al. (2011), and each verb occurs three times, each time paired with a different verb. In addition, 20 of the 60 pairs of sentences were identical to those used in the self-paced reading task in Rohde et al. (2011), and the remaining 40 pairs were newly created. This was designed to prevent the possibility that the models had seen the stimuli used in the original study during training.

As in (3), the direct object of the main verb is always a complex NP containing a singular NP and a plural NP, both of which are the possible attachment sites for the RC, followed by the relative pronoun *who*. This resembles the setup of the self-paced reading task in Rohde et al. (2011), where human participants did not see the full RC but anticipated which NP the RC would modify based on the verb type.

3. (a) Melissa detests the children of the musician who _____
   (b) Melissa babysits the children of the musician who ____

2.1.3 Task

We obtained the raw log probability of each of the possible auxiliary verbs that agree with either the number of the high NP or of the low NP (i.e., "is" and "are", respectively, in the example below).

Sentence: Melissa detests/babysits the children of the musician who is/are

As in the self-paced reading task in Rohde et al. (2011), this design directly probes whether the models incorporate pragmatic inference to predict the next word and make the attachment decision before an RC is even seen.

2.1.4 Evaluation

We calculated the difference in the log probabilities of each of the two auxiliary verbs in each sentence by subtracting the log probability of the critical word that agrees with the second NP in number (i.e., the low attachment site), from the log probability of the critical word that

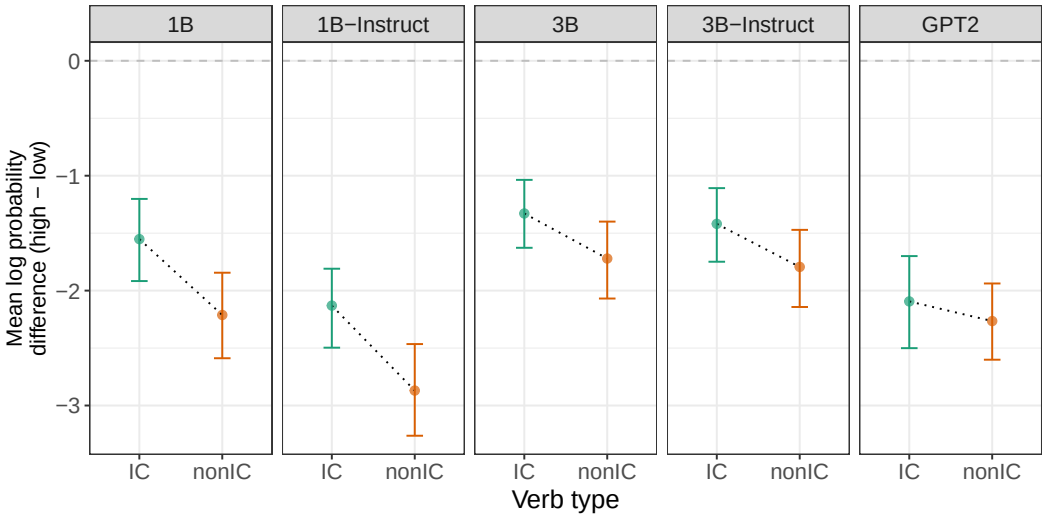

Figure 1: Mean log probability difference between the critical word that agrees with the high NP and that agrees with the low NP given the two verb types.

agrees with the first NP in number (i.e., the high attachment site), i.e., $\log(p_{high}) - \log(p_{low})$. Positive values indicate that the model prefers the high attachment over the low attachment.

2.2 Results

Fig. 1 shows the mean difference between the log probability of the critical word revealing a high attachment preference and that of the word revealing a low attachment preference. All models, except GPT-2, show a strong low attachment bias when nonIC verbs were used, where the mean log probability difference is negative. Yet, for at least some models, this bias is mitigated with IC verbs, where the mean log probability difference of the critical verb that shows the high/low attachment bias is smaller, suggesting there is a greater high attachment bias triggered by the IC verb. This is in line with human results, where the high attachment bias was 36.5% when nonIC verbs were used, revealing a low-attachment preference, but was increased to 50.6% when IC verbs were used (Rohde et al., 2011). Although most models show a similar effect of verb type as human participants, they nonetheless still strongly prefer the low attachment site even in the presence of an IC verb, as reflected in negative values in both verb type conditions in Fig. 1.

We fit a Bayesian mixed-effects linear regression predicting the log probability difference from the main effect of verb type and the maximal random-effects structure that allowed the model to converge, which includes the by-item random intercept. There was a significant main effect of verb type for all four Llama models, such that the log probability difference was lower when a nonIC verb was used than when an IC verb was used (Llama-3.2-1B: $\beta = -0.66$, $CrI = [-1.06, -0.26]$; Llama-3.2-1B-Instruct: $\beta = -0.74$, $CrI = [-1.07, -0.42]$; Llama-3.2-3B: $\beta = -0.39$, $CrI = [-0.70, -0.10]$; Llama-3.2-3B-Instruct: $\beta = -0.38$, $CrI = [-0.66, -0.08]$). This result suggests that for the Llama models, the high attachment preference is stronger when an IC verb is used than when a nonIC verb is used. However, the effect of verb type was not significant for GPT-2 ($\beta = -0.17$, $CrI = [-0.46, 0.11]$).

2.3 Discussion

The results show that all Llama models show a greater bias toward the high attachment site when an IC verb is used. Moreover, the content of the RC does not affect the attachment decision since it is not presented to the models. Therefore, these results suggest that the models anticipate the possibility of an explanation and incorporate that expectation in

predicting the attachment site for the RC and ultimately the auxiliary that follows the relative pronoun. However, GPT-2 does not show the expected behavior, suggesting that it cannot use pragmatic inference to make RC attachment decisions. In addition, unlike human participants, all models show a low-attachment preference in the IC verb condition, suggesting that there might still be other RC processing differences between humans and LLMs. Taken together, these results suggest that larger models can use pragmatic inference to guide downstream syntactic processing in ways that pattern with the behavior of human participants, whereas smaller models are less capable in this regard.

3 Experiment 2: Predicting reading time

It has been shown in the sentence processing literature that the more probable a word is in the context, the less time it takes for people to read it. This is quantified by the information-theoretic measure of surprisal, which is the negative log probability of the word, i.e., $-\log P(x_i|x_{<i})$. The surprisal theory predicts that the processing difficulty of a word is proportional to its surprisal (Hale, 2001; Levy, 2008).

Studies have used language models to estimate surprisal and have shown that the relationship between reading time and surprisal is linear (Smith & Levy, 2013). In addition, the model's predictive power for human reading has been used to measure the model's resemblance to the underlying psychological mechanisms of human sentence processing (Frank & Bod, 2011; Fossum & Levy, 2012; Goodkind & Bicknell, 2018). It has been shown that models with better ability to predict next words, as indicated by a lower perplexity value, have better predictive power (Fossum & Levy, 2012; Goodkind & Bicknell, 2018; Wilcox et al., 2020). However, Huang et al. (2024) challenged the explanatory power of surprisal estimates from LLMs in accounting for processing difficulties across a range of syntactically complex sentences, including RCs with ambiguous attachment sites. Specifically, even though the models they examined capture processing difficulties, they systematically underestimate the magnitude of the effect, suggesting that the role of predictability might be smaller than assumed by the surprisal theory in online sentence processing. In addition,models trained on extremely large datasets are not always better in predicting reading time, and the relationship between perplexity and their predictive power is sometimes reversed, due to their "superhuman" ability in next-word prediction (Oh & Schuler, 2023; Shain et al., 2024). Similarly, Kuribayashi et al. (2024) showed that the predictive power of instruction-tuned LLMs on reading time is worse than that of base LLMs.

Hence, if the negative relationship between perplexity and psychometric predictive power holds in larger models, then one might expect Llama-3.2-3B and Llama-3.2-3B-Instruct to be worse at predicting human reading time than their corresponding 1B Llama models. On the other hand, if the larger number of parameters in the 3B models results in enhanced pragmatic inference abilities, and these abilities are leveraged to make better syntactic predictions for the types of stimuli examined here, then one might expect to see the larger models perform better at predicting reading time.

3.1 Methods

3.1.1 Models

We evaluated the same five models as in Experiment 1, including Llama-3.2-1B, Llama-3.2-1B-Instruct, Llama-3.2-3B, Llama-3.2-3B-Instruct, and GPT-2.

3.1.2 Stimuli

All 40 sentences from the self-paced reading task in Rohde et al. (2011) were used, comprising 20 pairs of IC and nonIC verbs with an RC that is intended to provide an explanation to the event described in the matrix clause. There were 920 reading-time regions, and each region consisted of a minimum of a single word to a maximum of three words.

3.1.3 *Evaluation*

We obtained the surprisal of a word,[1] which is the negative log probability of that word in the sentence, i.e., $-\log P(w_i|w_{<i})$, from each model.

We used generalized additive models (GAM, Wood, 2017) to predict the average reading time across participants in the self-paced reading study in Rohde et al. (2011) from the surprisal value obtained from each model.[2] Since some words are tokenized into multiple subtokens and some regions in the self-paced reading study contain multiple words, we summed the surprisal of each token within the word and region, given the chain rule. In addition to surprisal, the regression models also include the effects of word length and frequency. Since word length and frequency also affect reading time, these factors were commonly included as control variables to delineate the effect of surprisal on reading time (e.g., Frank & Bod, 2011; Smith & Levy, 2013; Goodkind & Bicknell, 2018). In addition, we also included these measures of the two preceding words as predictors of the reading time of the current word to capture the possible spillover effects (Smith & Levy, 2013).

The unigram surprisal was used as the frequency measure, similar to Shain (2024). This measure was chosen since for regions that only contain one word, the unigram surprisal is proportional to the negative log frequency that is commonly used in other reading time studies (e.g., Wilcox et al., 2020; Oh & Schuler, 2023). Yet, in the current study, there are regions with multiple words that typically consist of a shorter but frequent function word and a longer and less frequent content word, as in noun phrases like "the teacher." Hence, we added the surprisal of each unigram w_i in the multi-word region x_i, i.e., $S(x_i) = \sum_i S(w_i)$, where $S(x_i)$ represents the surprisal of that region and $S(w_i)$ represents the surprisal of each word in that region. Aggregating the unigram surprisal takes into account all words in the region but weights the effect of less frequent content words more heavily than that of more frequent function words.[3] To obtain the unigram surprisal of each word, we measured the frequency count using the Corpus of Contemporary American English (COCA, Davies, 2008) and then calculated its probability by dividing the raw frequency count by the total number of words in the COCA dataset. The surprisal was obtained by taking the negative log probability: $S(w_i) = -\log P(w_i) = -\log \frac{\#w_i}{|V|}$, where $\#w_i$ is the raw frequency count of the word w_i, and $|V|$ is the total number of words in COCA.

We evaluated each LLM by comparing the full model to a baseline model that predicts reading time from the control variables, including the word length and the frequency (as measured by the unigram surprisal). Following the previous literature (e.g., Goodkind & Bicknell, 2018; Wilcox et al., 2020), the difference between the log-likelihood of the full model and the baseline model, i.e., ΔLogLik, is used to capture the model's psychometric predictive power. In addition, we calculated each model's perplexity to assess its next-word prediction accuracy and examine how this correlates with its ability to predict human reading time. A lower perplexity suggests that the model is more accurate in predicting the next word. The model's perplexity was calculated as the exponential of the average surprisal across all N regions, i.e., $e^{\frac{1}{N}\sum_i S(x_i)}$, where $S(x_i)$ represents the surprisal of each region x_i.

Moreover, we also fit a GAM model to the measures in the non-critical regions and then use that model to predict the reading time of the critical regions, which include the auxiliary verb of the RC and the two spillover regions immediately after. We used root mean squared error (RMSE) to measure the difference between the predicted reading time and the actual

[1] All log probabilities in the current study are base e, and thus surprisal is in nats instead of bits.

[2] The GAM model was run using the `mgcv` package (Wood, 2011) in R (R Core Team, 2022).

[3] An alternative way to measure the frequency of multi-words in a single region, x_i, will be to use the negative conditional probability of the current word given the previous word(s) as the surprisal value for the second word w_2 and third word w_3 in the region, i.e., $S(x_i) = S(w_1) + S(w_2) + S(w_3) = -\log P(w_1) - \log P(w_2|w_1) - \log P(w_3|w_2, w_1)$. Instead of using the unigram surprisal for all three words, the negative log probability of $-\log P(w_2|w_1)$ and $-\log P(w_3|w_2, w_1)$ can be obtained from a bigram and a trigram model, respectively. This would better capture the frequency of the entire region, taking into account the relationship between words in that region. We will adapt this method for future work.

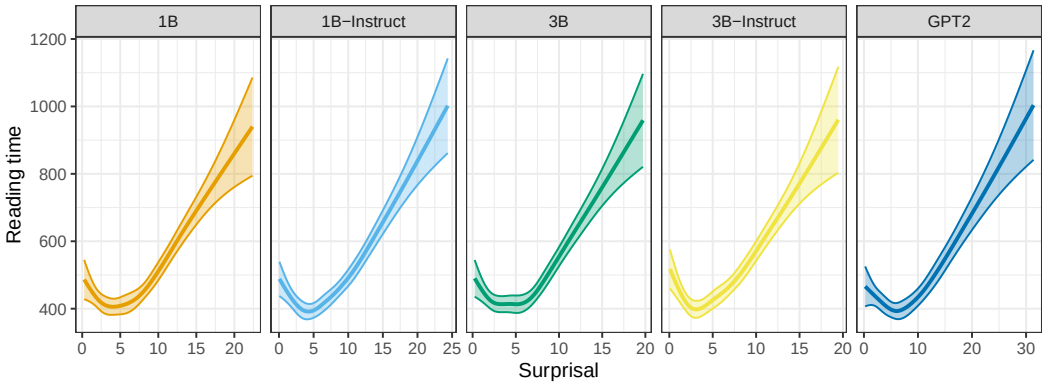

Figure 2: The relationship between reading time and surprisal. Each line represents the regression from the GAM model.

reading time. The smaller the RMSE value, the better the model is at predicting reading time in the critical region.

3.2 Results

Fig. 2 shows the relationship between reading time and surprisal, as fit by the GAM model. This captures the effect of surprisal on reading time above and beyond the effects of frequency and word length. The relationship is close to linear, especially when the surprisal is higher.

Fig. 3 shows the relationship between the model's predictive power, as measured by ΔLogLik, and its performance in next word prediction, as measured by perplexity. Based on the results of the five models in the current study, the relationship between perplexity and ΔLogLik is negative, suggesting that the better the model predicts the next word (i.e., the lower the perplexity is), the better it simulates reading time (i.e., the larger the ΔLogLik is). This negative relationship is similar to the findings with smaller non-transformer-based models (Frank & Bod, 2011; Goodkind & Bicknell, 2018; Wilcox et al., 2020) but contrary to the findings with more recent LLMs where the relationship is reversed (Kuribayashi et al., 2024; Oh & Schuler, 2023). This negative tendency seems to be driven by the high ΔLogLik and relatively low perplexity of Llama-3.2-3B-Instruct, while the ΔLogLik of the other Llama models is similar. GPT-2, on the other hand, has the largest perplexity and smallest ΔLogLik, suggesting that it is worse in predicting the next word and predicting reading times, as compared to the Llama models.

Lastly, we compared the root mean squared error (RMSE) of each model when fitting the GAM model to the measures in the non-critical regions to predict the reading time in the critical regions. GPT-2 had the largest RMSE value (107.81), followed by the two instruction-tuned Llama models (Llama-3.2-1B-Instruct: 99.36; Llama-3.2-3B-Instruct: 102.47). The two base Llama models exhibited relatively low RMSE values (Llama-3.2-1B: 89.81; Llama-3.2-3B: 93.06), suggesting that the relationship between surprisal and reading time in non-critical regions generalized more effectively to the critical regions, especially for the smaller model.

3.3 Discussion

The inverse relationship between perplexity and ΔLogLik does not align with the findings in Kuribayashi et al. (2024) or Oh & Schuler (2023), both of which claim that recent LLMs with lower perplexity tend to have worse psychometric predictive power. In fact, the larger instruction-tuned model, Llama-3.2-3B-Instruct, seems to have lower perplexity and better psychometric predictive power than the other smaller or base models. Since we only tested on five models on 40 sentences, where there were only two pairs that minimally differ in terms of their sizes and two pairs that minimally differ in terms of whether instruction-

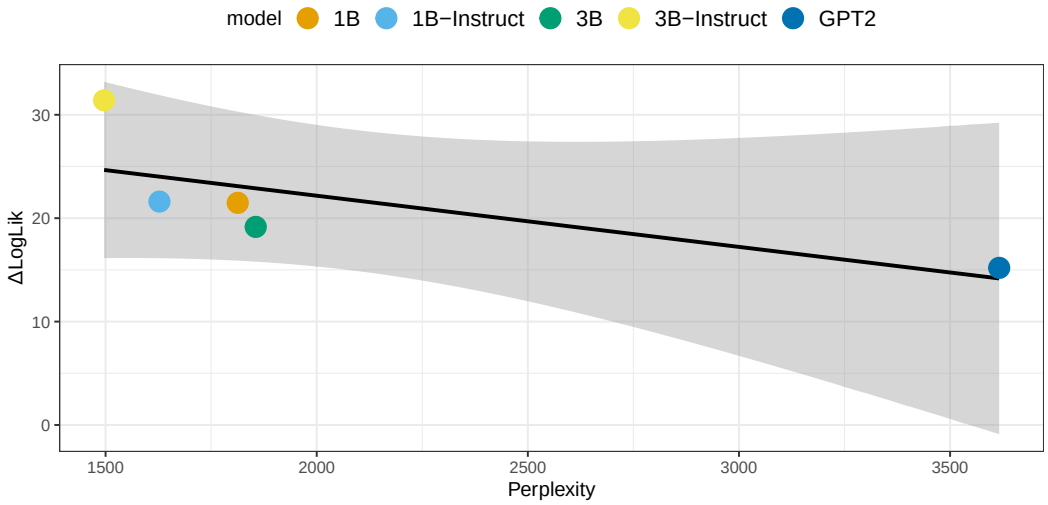

Figure 3: The Δlog likelihood against model perplexity.

tuning is used, the conclusion about the potential effects of model size and instruction tuning is preliminary, and we caution against over-interpreting these results. Below, we sketch out some hypotheses that could potentially explain the mismatch between the results in the current experiment and findings in the previous studies.

This difference with previous studies might be because the reading-time difference we observed with human participants, especially in the critical disambiguating regions of the auxiliary verb in the RC, was due to pragmatic inference. Thus, for models that fail to capture the qualitative high attachment bias in the IC condition, such as GPT-2, which are smaller models with higher perplexity, their psychometric predictive power will be worse than those that capture the attachment bias. This possibility is supported by the relatively high root mean squared error (RMSE) for GPT-2, suggesting that the relationship between surprisal and human reading time in non-critical regions did not generalize well to the critical region. Interestingly, larger instruction-tuned models, such as Llama-3.2-3B-Instruct, performed worse than smaller base models like Llama-3.2-1B in predicting reading times in the critical regions. This pattern is the opposite of what we observed when fitting the model to the entire dataset, yet it aligns with the findings in Kuribayashi et al. (2024), which suggest that instruction tuning may lead to a divergence from human reading behavior. In this case, larger models with better pragmatic inference abilities might not always simulate human reading time better. It is possible that the contextual factors required to decide the RC attachment site might not be fully captured by LLMs, leading to the modest correlation between surprisal estimated by models and processing difficulty in humans, as suggested in Huang et al. (2024).As mentioned above, these are speculative explanations of the observed patterns, supported by evidence from only a small subset of models, and future work needs to include more models with different sizes and use more sentences to reach a definitive conclusion.

4 General Discussion

The results of the experiments presented here suggest that LLMs have the ability to make pragmatic enrichments, with larger and more recent models demonstrating sensitivity to the influence of pragmatic inferences on syntactic processing. Overall, our findings contribute to the positive evidence of the pragmatic abilities of LLMs and their ability to leverage pragmatic inference in guiding downstream processing tasks.

The results from Experiment 1 indicate that GPT-2 fails to bring the pragmatic factors described herein to bear in predicting the likely attachment site for an ensuing RC, a result

that is predicted from the fact that it appears to lack the ability to draw the pragmatic inference in the first place. This result in fact aligns with prior studies showing its at-chance performance on other pragmatic tasks (Beyer et al., 2021; Hu et al., 2023). In contrast, the Llama models, regardless of their sizes and whether additional instruction-tuning is used, all demonstrate behavior consistent with their ability to use pragmatic inferences in the downstream tasks.

The reading time predictions in Experiment 2 indicate that models lacking pragmatic inference abilities, which tend to be smaller models, also exhibit reduced psychometric validity in simulating human reading behavior in a sentence processing task that is modulated by pragmatic inferences. In contrast, models that appear to possess such pragmatic abilities show mixed results in their ability to predict reading times: While there appears to be a negative relationship between the model's perplexity and the ΔLogLik, with Llama-3.2-3B-Instruct having the best psychometric predictive power among all attested Llama models, smaller base models are better at generalizing the reading time in the non-critical region to predict reading time in the critical region. Future studies should evaluate a wider range of models, varying training objectives and model sizes, to confirm this finding. In addition, since Experiment 2 was based on only 20 pairs of stimuli from the original Rohde et al. (2011) study, incorporating a greater number of items with comparable human results would further strengthen the findings.

References

Anne Beyer, Sharid Loáiciga, and David Schlangen. Is incoherence surprising? targeted evaluation of coherence prediction from language models. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4164–4173, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.328. URL https://aclanthology.org/2021.naacl-main.328/.

M. Carreiras and C. Clifton, Jr. Another word on parsing relative clauses: Eyetracking evidence from Spanish and English. *Memory and Cognition*, 27:826–833, 1999.

Tyler A Chang and Benjamin K Bergen. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1):293–350, 2024.

Mark Davies. Word frequency data from *the Corpus of Contemporary American English (COCA)*, 2008. Data available at https://www.wordfrequency.info.

Forrest Davis and Marten van Schijndel. Discourse structure interacts with reference but not syntax in neural language models. In Raquel Fernández and Tal Linzen (eds.), *Proceedings of the 24th Conference on Computational Natural Language Learning*, pp. 396–407, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.conll-1.32. URL https://aclanthology.org/2020.conll-1.32/.

Forrest Davis and Marten van Schijndel. Recurrent neural network language models always learn English-like relative clause attachment. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1979–1990, Online, July 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.179. URL https://aclanthology.org/2020.acl-main.179/.

Victoria Fossum and Roger Levy. Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd workshop on cognitive modeling and computational linguistics (CMCL 2012)*, pp. 61–69, 2012.

Stefan L Frank and Rens Bod. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological science*, 22(6):829–834, 2011.

L. Frazier and C. Clifton, Jr. *Construal*. MIT Press, Cambridge, Mass, 1996.

Lyn Frazier and Janet Dean Fodor. The sausage machine: A new two-stage parsing model. *Cognition*, 6(4):291–325, 1978.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. Neural language models as psycholinguistic subjects: Representations of syntactic state. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 32–42, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1004. URL https://aclanthology.org/N19-1004/.

Adam Goodkind and Klinton Bicknell. Predictive power of word surprisal for reading times is a linear function of language model quality. In Asad Sayeed, Cassandra Jacobs, Tal Linzen, and Marten van Schijndel (eds.), *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pp. 10–18, Salt Lake City, Utah, January 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-0102. URL https://aclanthology.org/W18-0102/.

John Hale. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*, 2001.

Jet Hoek, Hannah Rohde, Jacqueline Evers-Vermeul, and Ted JM Sanders. Expectations from relative clauses: Real-time coherence updates in discourse processing. *Cognition*, 210:104581, 2021.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 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)*, pp. 4194–4213, 2023.

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510, 2024.

Michael Kamerath and Aniello De Santo. Investigating syntactic biases in multilingual transformers with rc attachment ambiguities in italian and english. *arXiv preprint arXiv:2504.09886*, 2025.

Andrew Kehler, Laura Kertz, Hannah Rohde, and Jeffrey L Elman. Coherence and coreference revisited. *Journal of semantics*, 25(1):1–44, 2008.

Tatsuki Kuribayashi, Yohei Oseki, and Timothy Baldwin. Psychometric predictive power of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1983–2005, 2024.

Roger Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177, 2008.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016.

Byung-Doh Oh and William Schuler. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350, 2023.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL https://www.R-project.org/.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Hannah Rohde, Roger Levy, and Andrew Kehler. Anticipating explanations in relative clause processing. *Cognition*, 118(3):339–358, 2011.

Cory Shain. Word frequency and predictability dissociate in naturalistic reading. *Open Mind*, 8:177–201, 2024.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121, 2024.

Nathaniel J Smith and Roger Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319, 2013.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. BLiMP: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020.

Ethan G Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P Levy. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 42, 2020.

S. N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73 (1):3–36, 2011.

Simon N Wood. *Generalized additive models: an introduction with R*. chapman and hall/CRC, 2017.