Incremental Comprehension of Garden-Path Sentences by Large Language Models: Semantic Interpretation, Syntactic Re-Analysis, and Attention

Andrew Li, Xianle Feng, Siddhant Narang, Austin Peng, Tianle Cai, Raj Sanjay Shah, Sashank Varma
{ali403, xianle.feng, snarang37, apeng39, tcai38, rajsanjayshah, varma}@gatech.edu
Georgia Institute of Technology

Abstract

When reading temporarily ambiguous garden-path sentences, misinterpretations sometimes linger past the point of disambiguation. This phenomenon has traditionally been studied in psycholinguistic experiments using online measures such as reading times and offline measures such as comprehension questions. Here, we investigate the processing of garden-path sentences and the fate of lingering misinterpretations using four large language models (LLMs): GPT-2, LLaMA-2, Flan-T5, and RoBERTa. The overall goal is to evaluate whether humans and LLMs are aligned in their processing of garden-path sentences and in the lingering misinterpretations past the point of disambiguation, especially when extra-syntactic information (e.g., a comma delimiting a clause boundary) is present to guide processing. We address this goal using 24 garden-path sentences that have optional transitive and reflexive verbs leading to temporary ambiguities. For each sentence, there are a pair of comprehension questions corresponding to the misinterpretation and the correct interpretation. In three experiments, we (1) measure the dynamic semantic interpretations of LLMs using the question-answering task; (2) track whether these models shift their implicit parse tree at the point of disambiguation (or by the end of the sentence); and (3) visualize the model components that attend to disambiguating information when processing the question probes. These experiments show promising alignment between humans and LLMs in the processing of garden-path sentences, especially when extra-syntactic information is available to guide processing.

Keywords: Ambiguity; Garden-Path Sentences; Semantic Interpretation; Syntactic Parse Trees; Large Language Models

Introduction

Language is rife with ambiguity. Investigating how the human sentence parser handles this ambiguity has been important for revealing its processes, representations, and memory capacities. Of particular interest are *garden-path sentences*, which are sentences that are temporarily ambiguous between two structural interpretations. Readers often choose the incorrect interpretation, for example, the one that is statistically more frequent in the linguistic environment MacDonald et al. (1994). When they later reach the point of disambiguation, they are surprised and must then reanalyze the sentence to construct the correct parse tree and semantic interpretation. Surprisingly, the misinterpretation can linger and still be active at the end of the sentence (Christianson et al., 2001; Patson et al., 2009).

Large Language Models (LLMs) are deep neural networks trained on large corpora and range from multi-million parameter models like BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019) to state-of-the-art multi-billion parameter models like LLaMA-2 (Touvron et al., 2023) and GPT-4 (OpenAI, 2023). They have become more capable and have been claimed to reach human-level performance in general cognitive domains such as decision-making and problem-solving (Brown et al., 2020; Stiennon et al., 2022) and also on language tasks suhc as reading comprehension, grammar processing, and inference (Ye et al., 2023; Koubaa, 2023).

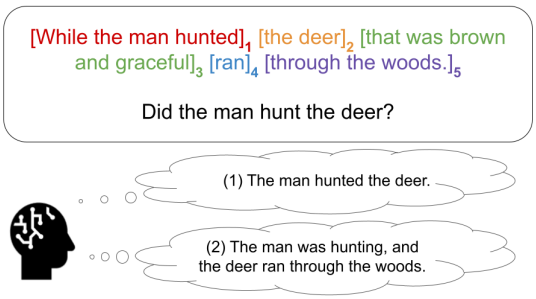

Figure 1: An example demonstrating the garden-path effect. During the incremental processing of this sentence, readers initially expect *deer* to be the object of *hunted*. Upon reaching the second verb *run*, they realize that *deer* is actually the subject of the second clause. This prompts a reanalysis of the sentence to the correct interpretation. However, the misinterpretation sometimes remains active even after reanalysis, and people still verify that "the man hunted the deer".

These successes have raised the question of whether LLMs are more than just engineering successes – whether they are also viable *scientific* models of human cognition. To support this claim, their performance must be measured and their representations examined for alignment to human behavioral signatures (Ivanova, 2023; Shah et al., 2023; Kallini et al., 2024; Bhardwaj et al., 2024; Vemuri et al., 2024). Only then can their correspondence to human cognition be properly evaluated. Here, we do so for the processing of temporarily ambiguous sentences. We move beyond standard surprisal-based measures (Hale, 2001; Levy, 2008; Wilcox et al., 2020a) and directly probe the semantic interpretations, implicit syntactic parse trees, and attention mechanisms as transformer-based LLMs incrementally process garden-path sentences.

Garden-Path Effects

Garden-path phenomena have long been studied in psycholinguistics. Here, we focus on two studies of the phenomenon exemplified in Figure 1. Christianson et al. (2001) had participants read 24 such garden-path sentences and tested their

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

final understanding using two yes/no comprehension questions. For the sentence shown in Figure 1, the questions were:

(1) Did the man hunt the deer?

(2) Did the deer run through the woods?

Interpretation (1) is consistent with the transitive verb interpretation of the sentence, which is assumed during the ambiguous region (i.e., chunk 3) but ultimately proves to be incorrect. In particular, it is inconsistent with the occurrence of the second verb *ran* in chunk 4, which is the point of disambiguation. And yet this incorrect interpretation lingers: Participants incorrectly verified (1) (i.e., responded 'yes') about 60% of the time. That said, they also computed the correct interpretation, verifying (2) nearly 90% of the time.

Additionally, Christianson et al. (2001) studied the effect of adding extra-syntactic information – a comma between chunks 1 and 2 (i.e., "While the man hunted, the deer ran through the woods.") – to signal the correct interpretation. This successfully minimized the ambiguity: a greater percentage of participants correctly rejected (1), the probe consistent with the misinterpretation, when the comma was present versus absent. Patson et al. (2009) found a similar result in their paraphrasing experiment.

LLMs and Psycholinguistics

A number of studies have investigated the alignment of LLMs with human sentence processing (Marvin & Linzen, 2018; Wilcox et al., 2019). Some of these studies, like the current study, have considered the processing of garden-path sentences (Jurayj et al., 2022; Wilcox et al., 2021, 2023). However, most have focused on predicting word-by-word reading times, the coin of the realm in psycholinguistics, using *surprisal* values derived from LLMs. Thus, they have shed little light on the research questions which animate the current study, which concern direct measurement of the semantic and syntactic interpretations that models form when processing garden-path sentences, and whether these interpretations shift following the point of disambiguation.

NLP researchers are becoming increasingly aware of the importance of representing ambiguity in LLMs, which is currently a challenge (A. Liu et al., 2023). However, the human alignment of the technical solutions these researchers are developing is outside the scope of the current study.

The Current Study

The current study compared humans and LLMs on the incremental processing of garden-path sentences. We probed the online processing and final interpretations of a range of models to address the following research questions:

1. Do LLMs represent the *semantic* misinterpretation of a garden-path sentence during the ambiguous region (i.e., chunk 3), and after they reach the point of disambiguation (i.e., chunk 4), do they switch to the correct interpretation?
2. Is the switch from the misinterpretation to the correct interpretation at the point of disambiguation also reflected in the implicit *syntactic* parse trees that LLMs construct?
3. Is the attention mechanism of transformer-based LLMs sensitive to the point of disambiguation?

We address these research questions using three novel methods that capitalize on the fact that LLMs can be directly interrogated in ways that human minds cannot. For the first research question, we present garden-path sentences to LLMs chunk by chunk. After each chunk, we used the comprehension questions – examples (1) and (2) above – to probe the strength of the misinterpretation and correct interpretation, respectively. This is a direct comparison between LLMs and humans that does not require indirect measures like *surprisal* (Hale, 2001; Levy, 2008; Wilcox et al., 2020a).

A parse tree is a hierarchical representation of the syntactic structure of a sentence. To address the second research question, we use the technique developed by Manning et al. (2020) to extract the parse tree at each chunk as an LLM incrementally processes a garden-path sentence. We evaluate whether this structure shifts at the point of disambiguation (i.e., chunk 4) from the misinterpretation to the correct interpretation.

The third research question is more exploratory in nature. We examine the attention weights of LLMs for evidence of sensitivity to the point of disambiguation (e.g., *ran* in the example in Figure 1). These reflect how LLMs weigh the different elements of the input. Vig (2019) and Clark et al. (2019) introduced tools to visualize these weights in transformer models. A substantial amount of linguistic information can be found in the attention weights of models (Clark et al., 2019; Y. Liu et al., 2019). We ask whether this includes information about the point of disambiguation.

Method

Large Language Models

Our study evaluated four LLMs ranging in performance, size, and architecture. We tested GPT-2 (Radford et al., 2019) and LLaMA-2 (Touvron et al., 2023), both decoder-only models with the latter being instruction-tuned. The third model was Flan-T5 (Chung et al., 2022), an encoder-decoder LLM. Finally, we evaluated RoBERTa (Y. Liu et al., 2019) as the most performant encoder-only-architecture model. RoBERTa is trained using a different paradigm (masked language modeling) than GPT-2 (next-token prediction) and LLaMA-2 and Flan-T5 (next-token prediction and instruction tuning).

Tasks and Datasets

We used two tasks originating in Christianson et al. (2001). The first used garden-path sentences and yes/no comprehension questions like examples (1) and (2) above. Preliminary testing revealed that all models correctly answered probe question (2) corresponding to the correct interpretation. We therefore focused on the models' endorsement of probe question (1) corresponding to the misinterpretation.

To simulate the incremental processing, we presented sentences to models in chunks and interrogated their unfolding state. Specifically, we split sentences at the following points:

(1) through the initial verb; (2) the misinterpreted "direct object"; (3) the descriptive clause; (4) the second verb, or point of disambiguation where we expect models to reanalyze the semantics and syntax of the sentence; and (5) the rest of the sentence, where reanalysis might spill over to; see Figure 1. We prompted the models as follows: prompt completion for GPT-2 and Flan-T5; in a chat format for LLaMA; and masked token prediction for RoBERTa. In all cases, we follow a question-answering template, prompting the language model with the garden-path sentence as context, the corresponding question, and "Answer: ".

The second task spans the same 24 garden-path sentences and yes/no questions as the first task, but with a comma inserted after chunk 1 (i.e., the first verb; *hunted* in Figure 1). This extra-syntactic information rules out the incorrect transitive interpretation of the first verb, and thus the misinterpretation that the noun phrase in chunk 2 (e.g., *the deer*) is its direct object, potentially disambiguating the sentence.

With respect to the datasets, one is from (the final) Experiment 3B of (Christianson et al., 2001), where the performance measure was accuracy on the comprehension question (1) corresponding to the misinterpretation. The second is from (Patson et al., 2009) who replicated this study but used a different accuracy measure: alignment of sentence paraphrases to the misinterpretation. Both datasets showed the same pattern of results: Inaccurate comprehension in the first task, but a positive effect of the extra-syntactic information and more veridical comprehension in the second task.

Experiments and Measures

Surprisal Surprisal has been the dominant metric for linking the processing of NLP models to psycholinguistic measures such as reading time (Hale, 2001; Levy, 2008; Wilcox et al., 2020b). To establish a baseline and continuity with prior works, we also compute the surprisal values of the 4 models, averaged within each of the 5 chunks, for both tasks.

Tracking Semantic Interpretations For probe (1) corresponding to the misinterpretation, we used token probabilities as an index of the likelihood of an LLM incorrectly answering "yes". That is, we collected the probability scores (logits) for the tokens "yes" and "no". We collected these measures after the processing of each of the five chunks.

We also computed the final answer accuracy for uniform comparison across models. Specifically, after all five chunks had been processed, we tabulated which token – "yes" corresponding to the misinterpretation and "no" to the correct interpretation – had the higher probability.

Incrementally Extracting Parse Trees A parse tree represents the syntactic structure of a sentence. We extracted the incremental parse tree after processing each chunk. We first measured whether a model is misled during the ambiguous region (i.e., chunk 3; *the deer that was brown and graceful* in Figure 1) and constructed the parse tree corresponding to the misinterpretation. We also measured whether, following the point of disambiguation (i.e., chunk 4; *ran* in Figure 1), the model reanalyzes and constructs the parse tree corresponding to the correct interpretation.

We did so using the technique developed by Manning et al. (2020) to train and extract the parse tree embedded implicitly in the word vectors predicted by an LLM. This is done by training and applying a linear transformation on the embedding word vectors in the hidden layers, and thereafter constructing the minimum spanning tree (MST). In this MST, the distances of the words are the norms of the word vectors, and the dependencies of the words are the edges.

We trained the parse tree probe on GPT-2 and RoBERTa-large. (We did not have access to the computational resources to do so for Flan-T5 and LLaMA-2.) We extracted the parse tree after each chunk and looked for re-analysis following the point of disambiguation (i.e., chunk 4; *ran* in Figure 1).

Visualizing Attention Weights Attention is a key concept in transformer-based LLMs Vaswani et al. (2017). The self-attention heads of a model capture the inter-token relations given as input to the model as $(key, query, value)$ tuples. In BERT-like transformer architectures, there is a multi-head self-attention attached to each layer. This allows the model to capture a variety of structural relations within the token through a weighted dot product.

The attention heads can be visualized as heatmaps to promote explainability. Clark et al. (2019) and Manning et al. (2020) used this technique to find a correlation between different attention heads and various linguistic relations such as direct object and subject-verb agreement. We visualize the attention heads, specifically the attention between pairs of tokens, as heat maps. We focus on the attention weight (1) between the initial verb in chunk 1 and the misinterpreted "direct object" in chunk 2 and (2) between the misinterpreted "direct object" in chunk 2 and the second verb in chunk 4, for which it is the correct subject. Thus, (1) represents the strength of the misinterpretation and (2) the strength of the correct interpretation. We then subtract (1) from (2) and interpret a positive value as evidence that the attention head is sensitive to the disambiguating information (and a negative value as evidence that it is not). We compute this value for all attention heads in all layers of the models.

Results and Discussion

Surprisal Baseline

In our baseline experiment, we compute surprisal for the four models on the first and second tasks, plotted in Figure 2 and Figure 3. For the first task, all models see a modest increase in surprisal at the point of disambiguation (chunk 4), as expected. For the second task, where the comma should disambiguate the sentence, RoBERTa and Flan-T5 show modest evidence of taking advantage of this extra-syntactic information and avoiding the garden path. By contrast, LLaMA and GPT-2 continue to show an increase in surprisal at chunk 4.

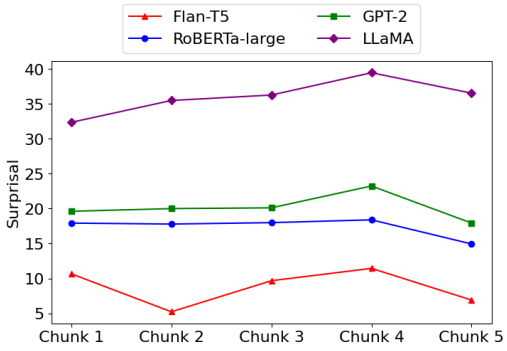

Figure 2: Surprisal in the first task (sentences with comma absent) across the four models.

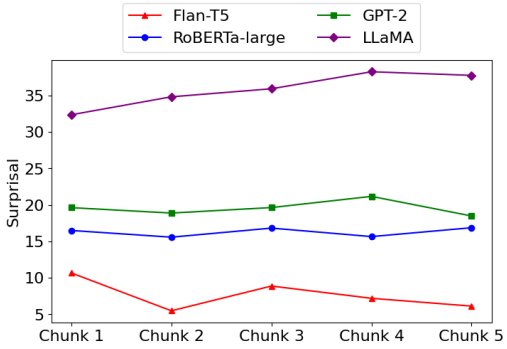

Figure 3: Surprisal in the second task (sentences with the disambiguating comma present) across the four models.

Tracking Semantic Interpretations

The first research question is when incrementally comprehending garden-path sentences, whether LLMs favor the misinterpretation during the ambiguous region (i.e., chunks 2 and 3) and then switch to the correct interpretation at the point of disambiguation (i.e., chunk 4) or afterward (i.e., chunk 5).

Consider the first task, where commas are absent. Figure 4 shows the probability of the misinterpretation (solid lines) and correct interpretation (dashed lines) across the five chunks for each of the four models. The first prediction is that the misinterpretation will be favored during the ambiguous region spanning chunks 2 and 3. This was the case for Flan-T5 and LLaMA, and by a smaller margin for GPT-2: the models assign a higher probability to "yes" to the verification probe (1), which is consistent with the misinterpretation, than to "no". The second prediction is that at the point of disambiguation, chunk 4, the pattern will reverse and the models will assign a higher probability to "no". This was *not* the case; all models continued to endorse the misinterpretation. However, for LLaMA the probability of "no" decreased on chunks 4 and 5, which is a promising trend.

By contrast, the models were more successful for the second task, where a disambiguating comma appears between chunks 1 and 2, suggesting that the noun phrase in chunk 2 should *not* be misinterpreted as the direct object of the matrix verb in chunk 1; see Figure 5. Specifically, LLaMA favors the

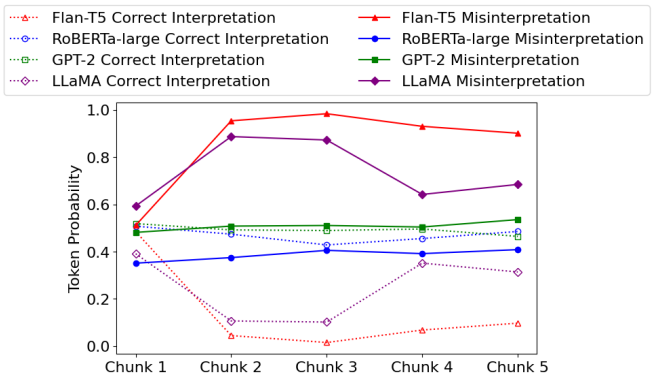

Figure 4: Semantic tracking of the mis- and correct interpretation in the first task (sentences with comma absent) across the four models. Critically, the probability of misinterpretation remains high even after the point of disambiguation.

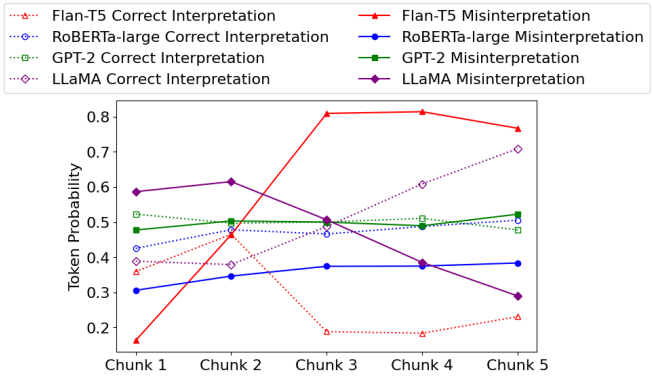

Figure 5: Semantic tracking of the mis- and correct interpretation in the second task (sentences with the disambiguating comma present). In the presence of this extra-syntactic information, the probability of misinterpretation decreases after the point of disambiguation for LLaMA, aligning with human performance.

misinterpretation during the ambiguous region (i.e., chunks 2 and 3). However, beginning at the point of disambiguation, the probability assigned to the misinterpretation decreases (a trend that continues in the final chunk of the sentence). This drop was statistically significant for LLaMA ($p = 0.003$) and also for GPT-2 (with $p = 0.05$). However, only LLaMA successfully switched to the correct semantic interpretation.

Turning from incremental semantic interpretation to the final semantic judgment, Figure 6 shows the percentage of garden-path sentences for which the probe question (1) corresponding to the misinterpretation was correctly rejected (i.e., had a "yes" response probability less than 50%) at the end of the sentence for the two tasks and four models. Performance on the first task (i.e., sentences with comma absent) is shown in light blue, and on the second task (i.e., sentences with the comma present) in dark blue. Also shown is the human performance on the two tasks in the Christianson et al. (2001) study. LLaMA, GPT-2, and Flan-T5 show human-like performance in being garden-pathed for sentences with-

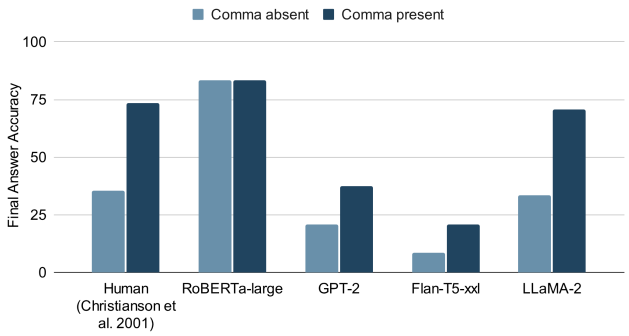

Figure 6: Final answer accuracy across models

out the comma but capitalizing on extra-syntactic information when it is present and recovering from the garden path. For all three of these models, accuracy is significantly greater on the comma-present vs. absent sentences ($p < 0.05$). LLaMA is notable in also showing human-like accuracies on the comma-present and comma-absent. By contrast, RoBERTa-large offers a poor fit for human performance.

Incrementally Extracting Parse Trees

The second research question concerns syntactic reanalysis: when LLMs reach the disambiguation point of a garden-path sentence, do they reanalyze the implicit parse tree they are constructing and shift to one consistent with the correct interpretation? Figure 7 shows the incremental parse trees across chunks 1-5 of the example sentence in Figure 1. These were extracted from the RoBERTa-large model. Note that the model initially incorrectly attaches *the rocket* in chunk 2 as the direct object of the main verb *photographed* in chunk 1. However, by the final chunk, it has shifted to the correct attachment, as the subject of the second verb *sat*. (More generally, many of the extracted parse trees identify the correct dependencies of the words and subordinate phrases.)

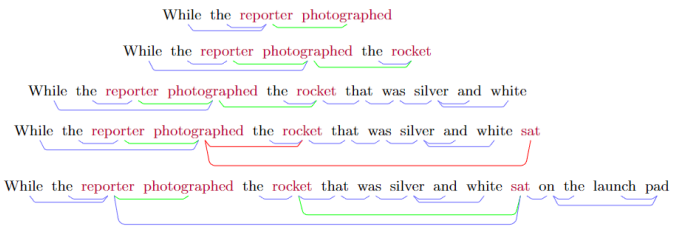

Figure 7: Example of incremental parse tree construction in RoBERTa-large.

The performance of GPT-2 and RoBERTa-large across the 24 garden-path sentences is summarized in Table 1. The table also contains the human data from the Christianson et al. (2001) and Patson et al. (2009) studies. First, consider the model and human performance on the first task, where the comma is absent. Following chunk 4, the point of disambiguation, especially RoBERTa-large begins to shift to the correct parse tree. By the end of the sentence (i.e., chunk 5), it has computed the correct parse tree for 45.83% of the sentences. This is comparable to the performance observed in especially the original Christianson et al. (2001) study. These findings coincide with those of Slattery et al. (2013), who found that reinterpretation of garden-path sentences spills over to the few words after the point of disambiguation.

Table 1: Proportion of sentences where the LLM makes the correct structural assignment shift for garden path sentences

LLM / Human	Comma Absent		Comma Present	
	Chunk 1-4	Chunk 1-5	Chunk 1-4	Chunk 1-5
GPT-2	12.50	16.67	45.83	50.00
RoBERTa-large	29.17	45.83	50.00	62.50
Christianson et al. (2001)	-	35.40	-	73.40
Patson et al. (2009)	-	21.00	-	62.00

Consider the second task, where sentences contain a disambiguating comma between chunks 1 and 2. As Table 1 shows, GPT-2 and RoBERTa-large shift to the correct parse tree after chunk 4, the point of disambiguation, for 45.83% and 50% of the sentences, respectively. The percentages increase to 50% and 62.5%, respectively, by the end of the sentence (i.e., chunk 5). Again, the accuracy of RoBERTa-large is comparable to the human participants in both studies. To summarize, RoBERTa-large performs most similarly to humans.

We confirmed that these descriptive patterns are also statistically significant by t-tests. For GPT-2, the presence of a comma had a significant impact on the proportion of sentences on which the model switched to the correct parse tree at the point of disambiguation, i.e., after chunks 1-4 ($p < 0.01$), as well as at the end of chunk 5 ($p < 0.01$). The results were similar for RoBERTa-large: the presence of a comma had a significant impact on switching to the correct parse tree after chunks 1-4 ($p < 0.1$). However, the percentages at the end of chunk 5 are comparable ($p > 0.1$). On the other hand, and collapsing across the two tasks, having seen chunk 5 had a significant impact on switching to the correct parse tree for RoBERTa-large ($p < 0.01$). But it does not make much difference on the performance of GPT-2 ($p > 0.1$).

Visualizing Attention Weights

The final research question is whether the attention mechanism of transformer models is sensitive to the point of disambiguation when processing garden-path sentences. Because LLaMA-2 and RoBERTa-large showed the strongest alignment with human performance in tracking semantic interpretations and extracting parse trees, respectively, we focus on these models.

We defined sensitivity as follows. LLaMA-2 is composed of 40 layers x 40 attention heads per layer; RoBERTa-large is composed of 24 layers x 16 attention heads. After a model processes a garden-path sentence, we can quantify the sensitivity of an attention head to the correct interpretation. Pos-

itive evidence is given by the attentional weight between the noun phrase in chunk 2 (e.g., *deer*) and the verb in chunk 4 (e.g., *ran*) representing the correct interpretation/attachment. Negative evidence is given by the attentional weight between the same noun phrase and the verb in chunk 1 (e.g., *hunted*) representing the misinterpretation/incorrect attachment. Then, positive evidence minus the negative evidence gives us the desired index: more positive values indicate good sensitivity, and more negative values have poor sensitivity.

Figure 8 shows the heatmap of the sensitivities across the attention heads of LLaMA-2 (left column) and RoBERTa-large (right column) for the comma-absent sentences in the top row. The heatmaps for the comma present sentences are in the middle row, and the thresholded difference of the middle row minus the top row is in the bottom row. The bright cells in the top and middle rows indicate the attention heads sensitive to the point of disambiguation, and thus the correct interpretation/parse, for the sentences of the first and second tasks. The bright cells in the bottom row indicate the attention heads that show particularly increased sensitivity in the context of the disambiguating comma of the second task.

These findings are exploratory in nature, and thus valuable in the new questions they raise. For example, one is whether the same attention heads that are sensitive to disambiguating information for the garden-path sentences studied here are also sensitive to disambiguating information in other temporarily ambiguous sentence structures.

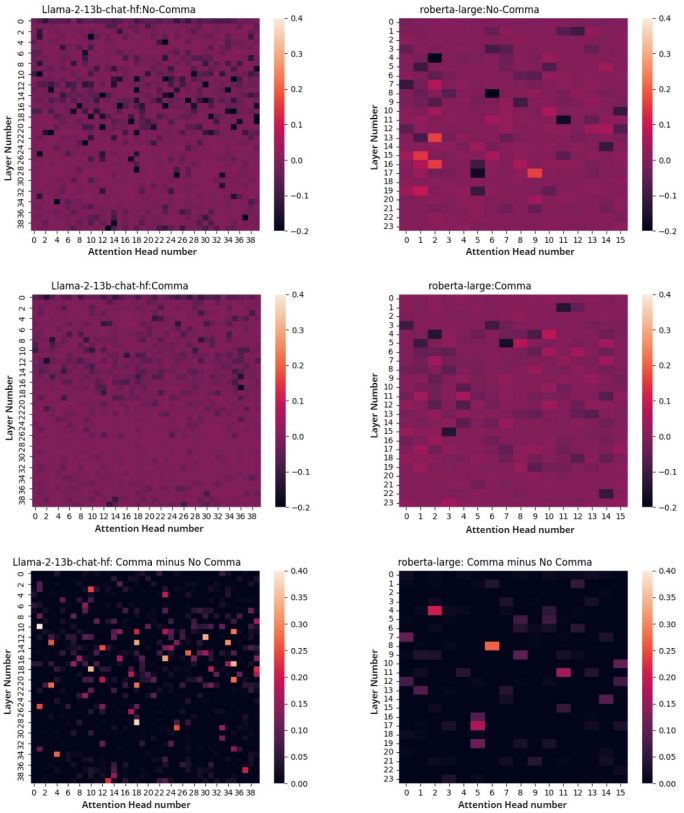

Figure 8: Sensitivity of the attention heads towards the correct interpretation in the presence of the comma.

Discussion

Cognitive scientists are increasingly investigating the value of LLMs as scientific models of the human sentence parser (Marvin & Linzen, 2018; Wilcox et al., 2021, 2023). The current study investigated the alignment between LLMs and humans in processing garden-path sentences. Christianson et al. (2001) and Patson et al. (2009) showed that when people process such sentences, misinterpretations can linger past the point of disambiguation, and people will verify probe questions consistent with them; see Figure 1 for an example. However, when the sentence is disambiguated early by adding a comma between chunks 1 and 2, people are less likely to do so.

We first showed that the classic surprisal metric shows some sensitivity to the point disambiguation. The first experiment, on semantic interpretation, found that RoBERTa, GPT-2, Flan-T5, and LLaMA-2 are garden-pathed on the first task, where the sentences have no commas, and misinterpretations lingered. However, on the second task, when the disambiguating comma was present, the larger models (i.e., GPT-2 and especially LLaMA-2) showed evidence of shifting to the correct interpretations. LLaMA-2, in particular, approximated human performance in its overall accuracy. The second experiment, on the incremental extraction of parse trees, found that GPT-2 and especially RoBERTa-large were sometimes able to switch to the correct parse tree at the point of disambiguation (i.e., chunk 4), and were even more successful afterward (i.e., chunk 5), especially when a comma was present. Here, RoBERTa-large approached human performance. The third experiment was more exploratory, showing that some of the attention heads of the LLaMA-2 and RoBERTa-large models are sensitive to the shift from the misinterpretation/incorrect attachment to the correct interpretation/attachment, as well as to the disambiguating information carried by the comma in the second task. Taken together, these results add to the growing evidence for LLMs as viable psycholinguistic models.

There are limits to our research that should be addressed in future studies. First, the four models we used span a range of architectures, training approaches, and sizes. However, they are far from exhaustive, and future work should evaluate a larger set of LLMs. Second, larger models tend to show emergent abilities (Wei et al., 2022). It is therefore important to run these experiments on larger models, both in upscaling the current models like LLaMA-70B and with newer LLMs like LLaMA-3 (AI@Meta, 2024) or Mixtral (Jiang et al., 2024).

A final limitation is that the set of materials used in our experiments is small, containing only 24 garden-path sentences. To increase the robustness and generality of our conclusions, it would be beneficial to repeat the experiments on a larger number of garden-path sentences for which human data is available, as well as on a broader range of temporary (syntactic) ambiguities.

References

AI@Meta. (2024). Llama 3 model card. Retrieved from https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

Bhardwaj, K., Shah, R. S., & Varma, S. (2024). *Pre-training llms using human-like development data corpus.*

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc.

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive psychology*, *42*(4), 368–407.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., ... Wei, J. (2022). Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does BERT look at? an analysis of BERT's attention. *arXiv preprint arXiv:1906.04341*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding.*

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Second meeting of the north American chapter of the association for computational linguistics.*

Ivanova, A. A. (2023). *Running cognitive evaluations on large language models: The do's and the don'ts.*

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., ... Sayed, W. E. (2024). *Mixtral of experts.*

Jurayj, W., Rudman, W., & Eickhoff, C. (2022). Garden path traversal in GPT-2. In *Proceedings of the fifth blackboxnlp workshop on analyzing and interpreting neural networks for nlp* (pp. 305–313). Association for Computational Linguistics. doi: 10.18653/v1/2022.blackboxnlp-1.25

Kallini, J., Papadimitriou, I., Futrell, R., Mahowald, K., & Potts, C. (2024). Mission: Impossible language models. *arXiv preprint arXiv:2401.06416*.

Koubaa, A. (2023). GPT-4 vs. GPT-3.5: A concise showdown.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126-1177. doi: https://doi.org/10.1016/j.cognition.2007.05.006

Liu, A., Wu, Z., Michael, J., Suhr, A., West, P., Koller, A., ... Choi, Y. (2023). We're afraid language models aren't modeling ambiguity. In *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 790–807). Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.51

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological review*, *101*(4), 676.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*(48), 30046–30054.

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 1192–1202). Association for Computational Linguistics. doi: 10.18653/v1/D18-1151

OpenAI. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Patson, N. D., Darowski, E. S., Moon, N., & Ferreira, F. (2009). Lingering misinterpretations in garden-path sentences: evidence from a paraphrasing task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(1), 280.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.

Shah, R. S., Marupudi, V., Koenen, R., Bhardwaj, K., & Varma, S. (2023). Numeric magnitude comparison effects in large language models. In *The 61st annual meeting of the association for computational linguistics.*

Slattery, T. J., Sturt, P., Christianson, K., Yoshida, M., & Ferreira, F. (2013). Lingering misinterpretations of garden path sentences arise from competing syntactic representations. *Journal of Memory and Language*, *69*(2), 104–120.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., ... Christiano, P. (2022). *Learning to summarize from human feedback.*

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... Lample, G. (2023). *Llama: Open and efficient foundation language models.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vemuri, S., Shah, R. S., & Varma, S. (2024). Evaluating typicality in combined language and vision model concept representations. In *Under review.*

Vig, J. (2019). A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... Fedus, W. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wilcox, E., Futrell, R., & Levy, R. (2023). Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, 1-44. doi: 10.1162/ling{_}a{_}00491

Wilcox, E., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020b). *On the predictive power of neural language models for human real-time comprehension behavior.* arXiv. doi: 10.48550/arXiv.2006.01912

Wilcox, E., Gauthier, J., Hu, J., Qian, P., & Levy, R. P. (2020a). On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd annual meeting of the cognitive science society* (p. 1707–1713).

Wilcox, E., Levy, R., & Futrell, R. (2019). *What syntactic structures block dependencies in rnn language models?* arXiv. doi: 10.48550/arXiv.1905.10431

Wilcox, E., Vani, P., & Levy, R. (2021). A targeted assessment of incremental processing in neural language models and humans. 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)* (pp. 939–952). Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.76

Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., ... Huang, X. (2023). A comprehensive capability analysis of GPT-3 and GPT-3.5 series models. *arXiv preprint arXiv:2303.10420*.