LLMs flunk garden paths like humans, but excel at memory-heavy structures: A sentence comprehension comparison

Samuel Amouyal (samueljoseph@mail.tau.ac.il), Jonathan Berant, & Aya Meltzer-Asscher
Tel Aviv University

Recent work has shown parallels between language processing in humans and large language models (LLMs): LLMs were shown to successfully predict human outcomes like reading times and eye-gaze, and to produce responses similar to humans on plausibility and acceptability judgments [1,5,7-10]. Yet surprisingly few studies examine a core outcome of processing: actual comprehension. In this study we test whether LLMs of various architectures and sizes show difficulty in answering comprehension questions about complex sentences, similarly to humans.

Methods. Comprehension was examined in seven structures known to be challenging for humans [2-4,6], relative to their baselines: four types of garden path (GP) structures, sentences with similarity-based interference, doubly center-embedded (DCE) sentences, and depth-charge sentences (Table 1). Forty sentence sets of each type in English were created (except for interference, where we used 24 sets from [4]). The human experiment used a single-trial design: each participant saw two simple practice sentences, followed by one experimental sentence and comprehension question. Each sentence was seen by 10 participants, yielding a total of 5,380 data points. The same task was given to 31 LLMs from 5 different families (Qwen, Llama, OpenAI, Gemma, Olmo), with sizes from 0.5B parameters to hundreds of billions. In the prompt to the LLMs we provided examples of the task, but none of the examples included complex structures.

Results. Human performance confirmed the difficulty of the structures. Average accuracy on target sentences was 28.3% (Subj/Obj: 13.3%, NP/S: 29.7%, NP/VP: 18.5%, reduced relative: 41.7%, DCE: 32.3%, Interference: 36.9%; Depth charge: 28.0%). While LLMs outperformed humans overall, their accuracy remained low, ranging from 23.6% to 74.5%.

Two notable patterns emerged from the data. GP vs. memory load: Although strong models achieve near-perfect accuracy on non-GP structures (e.g. 93.7% for GPT-5), they still struggle on GP sentences (e.g. 46.8% for GPT-5). As a result, LLM error rates on GP sentences are closer to human levels (average absolute difference between models and humans ≈ 0.17) than for memory-heavy structures, i.e. interference structures and DCE (difference ≈ 0.37) (Figure 1). The "sweet spot" of model size: Only models within an intermediate capacity range reliably mirror the human pattern of difficulty, namely that baseline sentences are easier than target sentences. Weak models yield poor performance on both baseline and target, and very strong models perform well on both. We validated our "sweet spot" rule by counting for each structure the proportion of violations of this rule, namely cases where similarity to humans decreases and then increases with model size. Violation rates varied between 0 and 15%.

Discussion. Our findings show that, bar the strongest models, LLMs struggle in answering comprehension questions about structures that are challenging for humans. Importantly, we found a dissociation in LLM processing: models seem to successfully handle structures which require ample working memory resources, i.e. center-embedding and interference sentences, presumably as their "working memory" is not limited like humans'; However, they fail at garden-path sentences, which require inhibition of an initial parse and reanalysis. Thus, by testing models, which differ from humans in specific "cognitive" components, we can shed light on why different structures are difficult for humans. The results also show for the first time comprehension difficulties for humans in several structures which were not tested for comprehension before.

Type	Target Sentence	Baseline Sentence
Subj/Obj (GP)	While the man hunted the deer ran into the woods.	The deer ran into the woods while the man hunted.
	Question: Did the man hunt the deer? Correct answer: No	
NP/S (GP)	The policeman saw the lights were off.	The policeman saw that the lights were off.
	Question: Did the policeman see the lights? Correct answer: No	
NP/VP (GP)	The complex houses married soldiers.	The complex housed married soldiers.
	Question: Are there complex houses? Correct answer: No	
RR (GP)	The chef hired last month worked overtime.	The chef who was hired last month worked overtime.
	Question: Did the chef hire someone? Correct answer: No	
DCE	The man that the teacher that the student liked called sat.	The student liked that teacher that called the man that sat.
	Question: Who did the student like? Correct answer: The teacher	
Depth charge	No head injury is too trivial to be ignored.	Every head injury is severe enough to be ignored.
	Question: Can you ignore head injuries? Correct answer: No	
Interference	The banker that the barber praised climbed the hill.	The banker that you praised climbed the hill.
	Question: Did the barber/you climb the hill? Correct answer: No	

Table 1: Example sentences for structures tested in the experiment

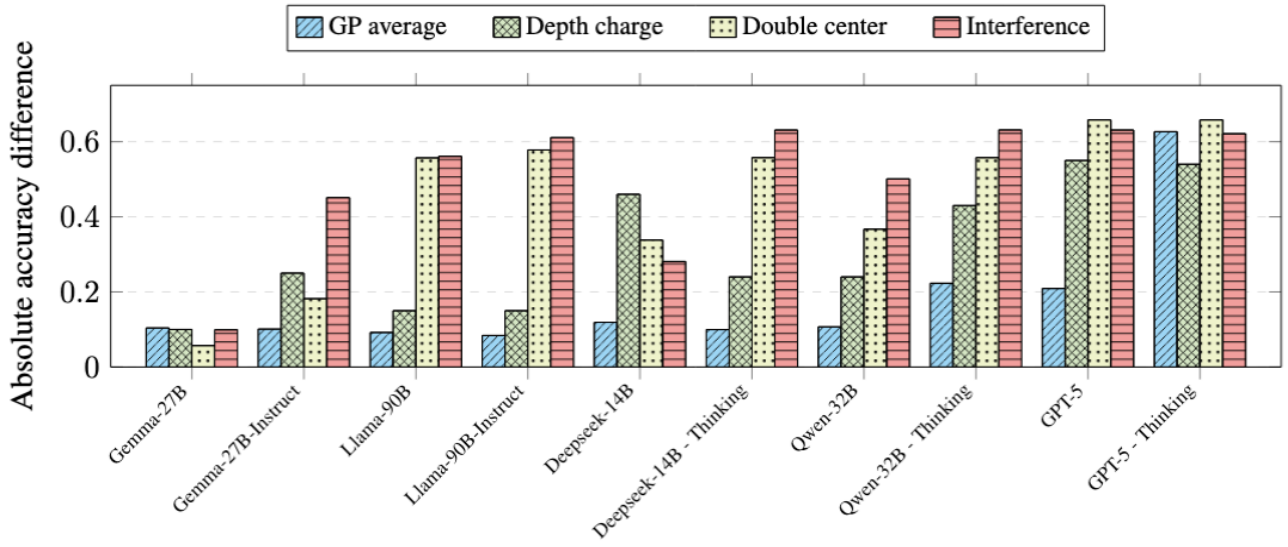

Figure 1: Difference between LLM and human accuracy on target conditions

References

[1] Amouyal, S., Meltzer-Asscher, A., and Berant, J. (2024). Findings of ACL [2] Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Cognitive psychology [3] Frazier, L. (1987). Attention and performance [4] Gordon, P. C., Hendrick, R., & Johnson, M. K. (2001). Journal of experimental psychology. Learning, memory, and cognition [5] Hu, J., Gauthier, J., Qian, P., Wilcox, E., and Levy, R. (2020). ACL [6] Kizach, J., Christensen, K. R., and Weed, E. (2016). Journal of Psycholinguistic Research [7] Kuribayashi, T., Oseki, Y., Taieb, S. B., Inui, K., & Baldwin, T. (2025). ArXiv preprint, abs/2502.01615 [8] Linzen, T., Dupoux, E., and Goldberg, Y. (2016). TACL [9] Rego, A. L., Snell, J., and Meeter, M. (2024). PLOS Computational Biology. [10] Warstadt, A., Singh, A., and Bowman, S. R. (2019). TACL