Evaluating LLMs for abstract linguistic generalization using English parasitic gaps

Introduction. Some research shows that large language models (LLMs) show human-like behavior [1,2]. We evaluate twelve LLMs' adequacy [8,9,10,11] as a cognitive model by investigating their abilities to form human-like abstract linguistic generalizations. Specifically, we compare LLMs with humans using structural priming [3], speakers' tendency to reuse structures recently encountered. The human results are taken from two experiments from [4] involving English parasitic gaps (PG [6]). To understand what a PG is, first consider (1c), which doesn't have a PG. This is a wh-question with an object gap after *read* and a co-referential pronoun *it* serving as the object of *criticizing*. In the minimally different yet synonymous (1b), *it* is replaced with a gap, which cannot exist without the gap after *read* and is thus called a parasitic gap (**Which novel did the reviewer read it before criticizing?*). PGs offer a strong test for LLMs' ability to form abstract generalizations because their distribution is difficult to infer from superficial properties of sentences. While LLMs seem to generalize across concrete cases of PGs, this generalization does not align with human generalizations found in previous studies.

Data. Using sentence recall tasks, [4] compared a variety of constructions in their priming effects on PG production. Half of the target sentences contained PGs like (1b); the other half had a pronoun instead (like (1c)). The proportion of PG production (like (1b)) among both PG and pronoun productions (like (1c)) was recorded. The PG priming effect of a construction was measured by the difference in this proportion compared to that of the neutral control.

Evaluation. We artificially generated two sets of 2000 sentences that fit the syntactic patterns in the stimuli from [4]. For each prime, we let the LLMs calculate the log probabilities of concatenations of said prime with the PG target (like (1b)) and with the pronoun target (like (1c)). We took the difference between these two log-probabilities, and compared the differences between each prime and the control prime to measure the PG priming effect of that prime, a metric proposed in [5]. A positive change from the control prime indicates PG priming.

Experiment 1 compares across-the-board movement (ATB [7]; 1a) constructions with PGs (1b). They are superficially similar: both (1a,b) feature two transitive verbs with missing objects with a word in between. For humans, while PGs prime PGs, increasing PG production by 11%, ATBs don't prime PGs. All LLMs show a PG priming effect for both PGs and ATBs (Fig. 1).

Experiment 2 uses PGs hosted in relative clauses. In (2a), the PG is located in an adjunct *after playing in*. In (2b), the PG is located in the embedded relative clause *who plays in*. Thus, while (2a,b) and the recall target (like (1b)) all feature PGs, they appear superficially dissimilar as the PG is hosted in different environments: relative clauses vs. wh-questions. Nonetheless, for humans, the two kinds of PGs in relative clauses both prime PG production in wh-questions, both increasing PG production by 10%. Five LLMs (GPT-Neo 125M, 13B, 27B; DialoGPT small; Mamba 130M) show priming effects for both kinds of PGs in the human-like direction (Fig. 2). However, while the magnitude of the priming effect was comparable between Exp. 1 and Exp. 2 for humans [4], it is reduced for all these five LLMs. Their average PG priming effect for PGs in Exp. 1 was 1.65 (i.e., PG production is 5.22 times as likely as pronoun production), and in Exp. 2 this was reduced to 0.293 (PG / pronoun = 1.34) for adjunct PGs and 0.507 (PG / pronoun = 1.66) for embedded relative PGs. This suggests that LLMs were less sensitive to the abstract similarity between PGs created by wh-question formation and by relativization.

Conclusion. The fact that PGs prime PGs for all LLMs suggests that they form some abstract generalization over specific instances of PGs, but no LLM correctly distinguished PGs from ATB movement (Exp. 1), and LLMs mostly failed to generalize across PGs caused by different processes (wh-question formation vs. relativization); some models failed to capture even the direction of priming effect, and the other models predicted considerable reduction in the priming effect compared to Exp. 1, which was not observed in humans (Exp. 2). Thus, LLMs are capable of generalizing beyond surface-level pattern recognition, but the way they make generalizations is insufficiently sensitive to abstract similarities across constructions and overly sensitive to superficially similar constructions.

(1) Primes used in [4]'s Experiment 1. ATB = across-the-board movement.
a. ATB: Which novel did the reviewer read and criticize?
b. PG: Which novel did the reviewer read before criticizing?
c. Control: Which novel did the reviewer read before criticizing it?

(2) Primes used in [4]'s Experiment 2. Adjunct PG = parasitic gap in adjunct. Embedded relative PG = parasitic gap in embedded relative clause.
a. Adjunct PG: This is the park that every kid loves after playing in.
b. Embedded relative PG: This is the park that every kid who plays in loves.
c. Control: This is the park that every kid who has a dog loves.

Figure 1. LLM results for [4]'s Experiment 1. Error bars indicate +/- one standard error.

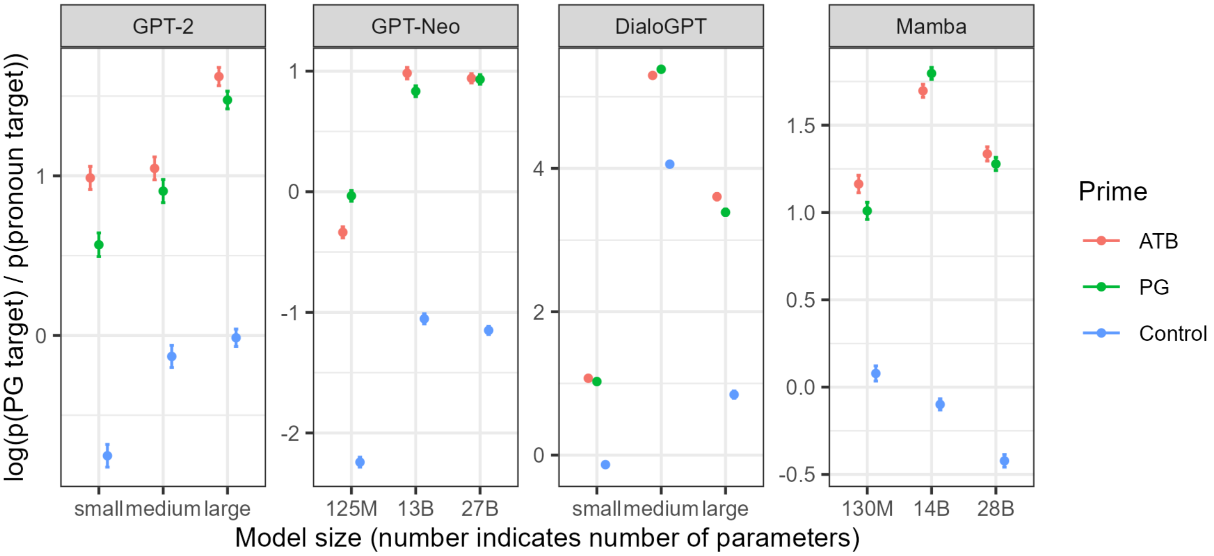
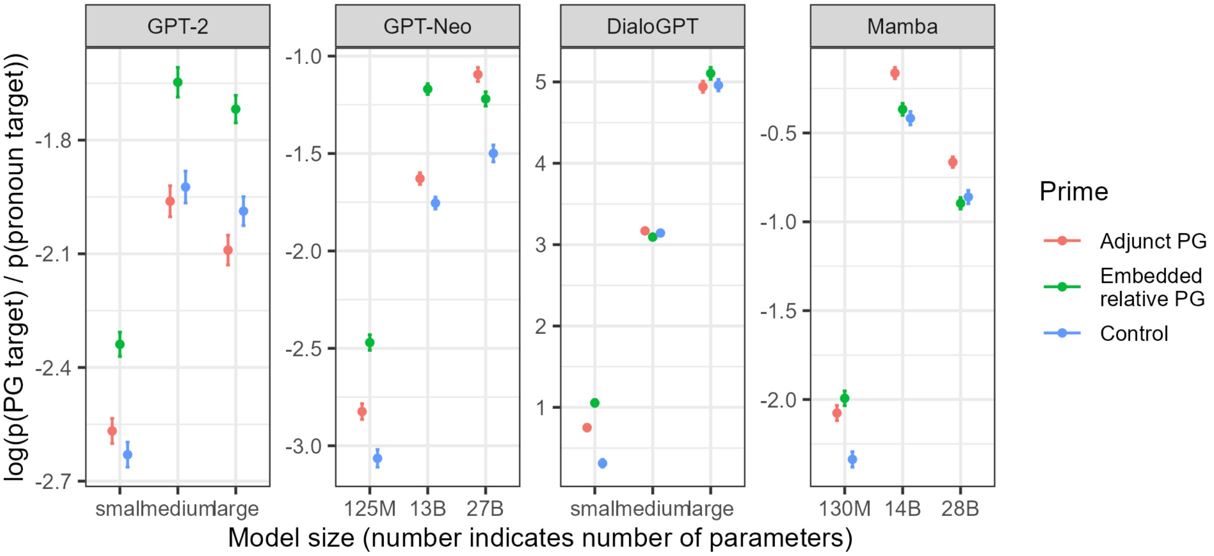


Figure 2. LLM results for [4]'s Experiment 2. Error bars indicate +/- one standard error.



References. [1] Lau et al., 2017. *Cog Sci.* [2] Wilcox et al., 2024. *Linguistic Inquiry.* [3] Bock, 1986. *Cog Psych.* [4] Momma et al., 2024. *Proceedings of NELS 54.* [5] Sinclair et al., 2022. *Transactions of ACL 10.* [6] Engdahl, 1983. *Ling & Phil.* [7] Williams, 1978. *Linguistic Inquiry.* [8] Radford et al. *OpenAI Blog.* [9] Gao et al., 2020. *arXiv 2101.00027.* [10] Zhang et al., 2020. *arXiv 1911.00536.* [11] Dao & Gu, 2024. *arXiv 2405.21060.*