EXAMINING CHATGPT'S GRAMMATICAL INTUITION: CONVERGENCE WITH LINGUISTIC EXPERTS AND LAYPEOPLE

Zhuang Qiu, Xufeng Duan, Zhenguang Cai (The Chinese University of Hong Kong)
zhuangqiu@cuhk.edu.hk

Large language models (LLMs) have demonstrated exceptional performance across various linguistic tasks [1,2]. However, it remains uncertain whether LLMs have developed human-like fine-grained grammatical intuition. This preregistered study (https://osf.io/t5nes) presents the first large-scale investigation of ChatGPT's grammatical intuition, building upon previous research [3] that examined the grammaticality judgment of 148 linguistic constructions with varying degrees of acceptability for humans (e.g., 1a. "he was the judge" vs 1b. "he was judge"). These constructions were sampled from the journal of *Linguistic Inquiry* published between 2001 to 2010. Linguists had classified these constructions as grammatical, ungrammatical, or marginally grammatical, and their grammatical acceptability was assessed by layman participants.

In this study, our primary focus was to explore ChatGPT's judgments of these constructions in comparison to both layman participants and linguistic experts. The experimental design followed the methodology outlined in [3], with each item being tested across 50 runs. In Experiment 1, we presented to ChatGPT a reference sentence with a pre-assigned acceptability rating (e.g., 100) and asked it to assign a rating (in multiples of the reference rating) to a target sentence. In Experiment 2, we asked ChatGPT to rate the grammatical acceptability of a target sentence on a 7-point scale (1 = "least acceptable" and 7 = "most acceptable"). In Experiment 3, we presented ChatGPT with a pair of sentences (such as 1a and 1b) and asked it to decide which is grammatically more acceptable. The data from ChatGPT was combined with the human data from [3] for subsequent analyses.

Overall, our findings demonstrate convergence rates ranging from 73% to 95% (depending on the experiment and statistical test) between ChatGPT and human linguistic experts, with an overall point-estimate of 89%. This means that, in general, ChatGPT correctly distinguishes grammatical sentences from ungrammatical ones approximately 89% of the time. However, the behaviour patterns of ChatGPT and layman participants varied depending on the specific judgement task. While ChatGPT provided lower rating scores for ungrammatical sentences (β = -0.91, CI = [-0.97, -0.86]) than grammatical sentences in the magnitude estimation task, the differences in rating scores between the two were not as big as those observed among the layman participants. Furthermore, ChatGPT exhibited higher accuracy in the force choice task compared to the layman participants (β = -18.1, CI = [-21.69, -14.98], with ChatGPT being the baseline). Notably, significant differences between ChatGPT's judgments and those of human non-experts were observed in both the magnitude estimation (Experiment 1) and force choice (Experiment 3) tasks, whereas no such difference was found in the Likert scale task (Experiment 2). We attribute these results to the psychometric nature of the judgment tasks and the differences in the representation of grammatical knowledge between humans and LLMs.

1. Ortega-Martín, M., García-Sierra, Ó., Ardoiz, A., Álvarez, J., Armenteros, J. C., & Alonso, A. (2023). Linguistic ambiguity analysis in ChatGPT. *arXiv preprint*, arXiv:2302.06426.
2. Jiao, W., Wang, W., Huang, J. T., Wang, X., & Tu, Z. (2023). Is ChatGPT a good translator? A preliminary study. *arXiv preprint*, arXiv:2301.08745.
3. Sprouse, J., Schütze, C. T., & Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua*, 134, 219-248.