Do Large Language Models Resemble Humans in Grammaticality Judgement?

Zhuang Qiu, Xufeng Duan, and Zhenguang G. Cai[*]
The Chinese University of Hong Kong, Hong Kong SAR

Background: Research in artificial intelligence over this decade has witnessed the surge of large language models (LLMs) demonstrating improved performance in various natural language processing tasks[1]. This has sparked significant discussions about the extent to which large language models emulate human linguistic cognition and usage [2,3,4]. While previous research has explored how LLMs process grammatical features such as filler-gap dependencies [5] and garden path sentences [6], not many studies have compared LLMs' nuanced intuition of grammaticality with laypeople's knowledge of various syntactic structures. We report two preregistered research projects (https://osf.io/t5nes; https://osf.io/75dtk) presenting a large-scale investigation of LLMs grammatical intuition, comparing their grammaticality judgement with that of human laypeople and linguists.

Study 1: This study builds upon a previous study [7] that collected laypeople's grammatical judgements for 148 pairwise linguistic phenomena (Table1), categorized by linguists as grammatical, ungrammatical, or marginally grammatical. Our primary focus was to compare ChatGPT with both laypeople and linguistic experts in the judgement of these linguistic constructions. **Exp.1** used a magnitude estimation task, where we presented to ChatGPT a reference sentence that we had pre-assigned an acceptability rating of 100 and asked it to assign a rating (proportional to the reference rating) to a target sentence. Comparing the ChatGPT data with the laypeople data from [7], we found that grammatical sentences were judged to be more acceptable by humans than by ChatGPT, while ungrammatical sentences were judged as less acceptable by humans than by ChatGPT. We also estimated that the ChatGPT-Linguist convergence rate in this experiment was 73% - 91%. In **Exp.2**, we asked ChatGPT to rate the grammatical acceptability of a target sentence on a 7-point scale. We observed a strong correlation in ratings between ChatGPT and laypeople (r = 0.72) and no statistical difference in rating was observed between the two groups. The estimated convergence rate between ChatGPT and linguists was 75% - 95%. In **Exp.3**, we presented to ChatGPT a pair of sentences and asked it to decide which was grammatically more acceptable. We found a modest correlation between ChatGPT's choices with those of humans (r = 0.39) and a higher accuracy rate with ChatGPT than with humans. The estimated ChatGPT-Linguist convergence rate was 88% - 89%. Overall, our findings demonstrate convergence rates ranging from 73% to 95% between ChatGPT and linguists, with an overall point-estimate of 89%.

Study 2: This study collected grammaticality judgement data for over 2400 English sentences with varying grammatical structures from ChatGPT and Vicuna [8], comparing them with human judgement data for the same sentences from [9]. As a partial replication of Study1, the data were elicited from three judgement tasks. The **Binary Judgement Task** required a judgement of sentence as either grammatical or ungrammatical. In this task, a strong correlation (defined as a Pearson r above 0.5 based on [10]) was observed between human and the two LLMs (r = 0.83 for human and chatgpt; r = 0.66 for human and vicuna). The **4-Category Rating Task** required the rating of sentence grammaticality using a 4-point Likert scale. We found a strong correlation between human and ChatGPT (r = 0.84) and a moderate correlation between human and Vicuna (r = 0.49). The **Sliding Scale Rating Task** required a grammaticality judgement using a scale of integers from 1 to 100. We again found a strong correlation between human and the two LLMs (r = 0.81 for human and chatgpt; r = 0.72 for human and vicuna).

Conclusion: We found significant correlations between human and LLMs in both of the studies, though the strength of the correlation varied as a function of the task. We attribute these results to the psychometric nature of the judgement tasks and the differences in the representation of grammatical knowledge between humans and LLMs.

[*] Corresponding author, Zhenguang G. Cai: zhenguangcai@cuhk.edu.hk

References:

[1] Brown et al. (2020). NeurIPS, 33, 1877-1901. [2] Chomsky et al. (2023). The New York Times. [3] Piantadosia. (2023). lingbuzz/007180. [4] Mahowald et al. (2023). arXiv:2301.06627. [5] Wilcox et al. (2023). Linguistic Inquiry, 1-44. [6] Van Schijndel & Linzen (2018). CogSci. 2603-2608. [7] Sprouse et al. (2013). Lingua,134, 219-248. [8] Chiang et al. (2023) https://lmsys.org/blog/2023-03-30-vicuna/. [9] Lau et al. (2017). Cognitive science, 41(5), 1202-1241. [10] Cohen. (1992). Psychological Bulletin, 112, 155–159.

Grammatical Sentences	Ungrammatical Sentences
It seems to him that Kim solved the problem.	He seems to that Kim solved the problem.
It appears to them that Chris is the right person for the job.	They appear to that Chris is the right person for the job.
It seems to her that Garrett should be punished for lying.	She seems to that Garrett should be punished for lying.
It appears to me that Dana is an unsafe driver.	I appear to that Data is an unsafe driver.
It seems to me that Robert can't be trusted.	I seem to that Robert can't be trusted.
It appears to her that Kyle cheated on his homework.	She appears to that Kyle cheated on his homework.
It seems to them that Sandra hates cooking.	They seem to that Sandra hates cooking.
It appears to him that Erin enjoys swimming.	He appears to that Erin enjoys swimming.

Table 1: Examples of experimental items from Sprouse et al. (2013) which were adopted in Study 1.

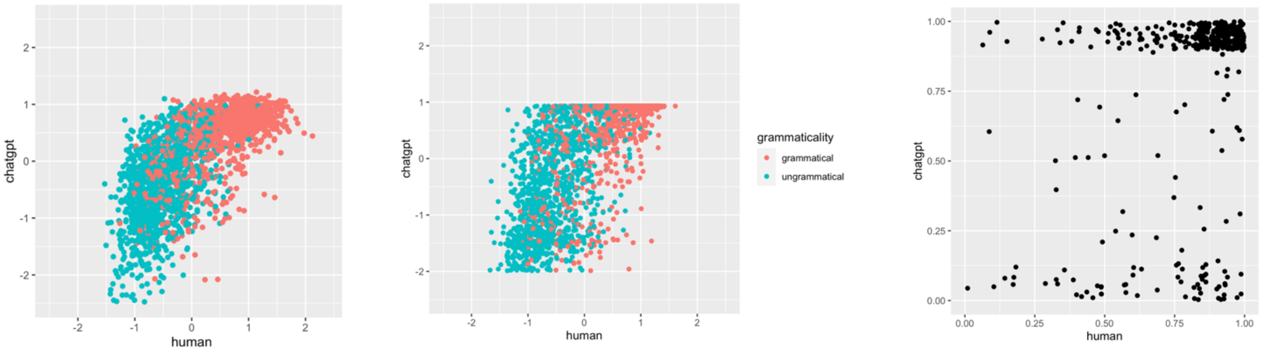

Fig 1: Correlation of acceptability ratings between human participants and ChatGPT in Study 1; each point represents the mean rating score of a sentence. Left panel: Exp1; Middle panel: Exp2; Right panel: Exp3.

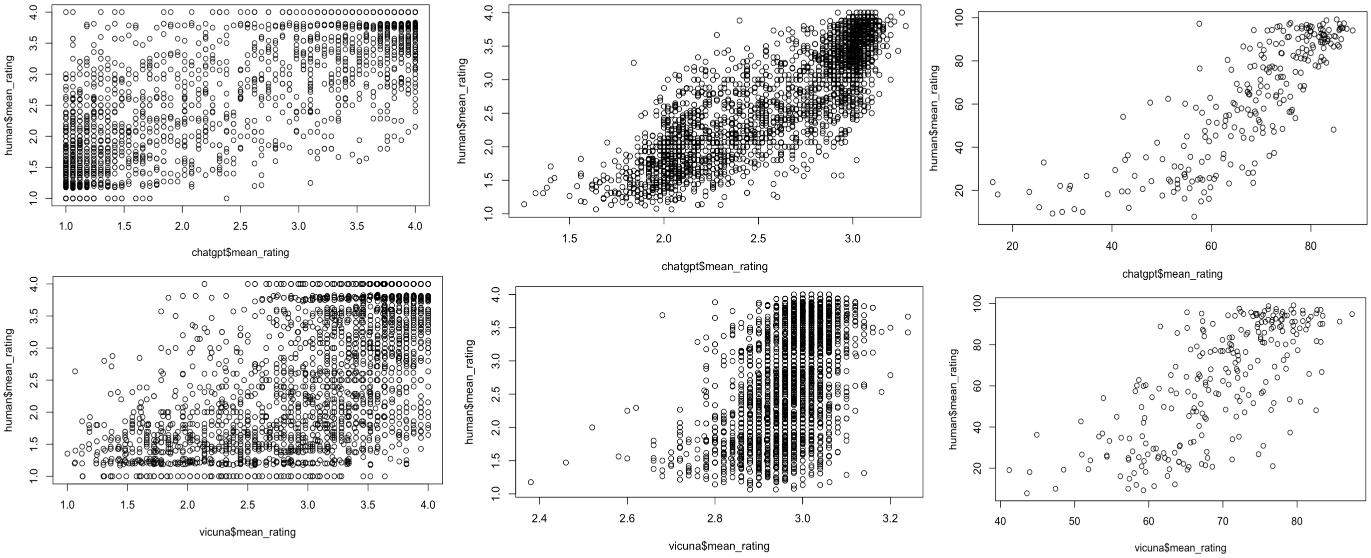

Fig 2: Correlation of acceptability ratings between human participants and LLMs in Study 2; each point represents the mean rating score of a sentence. Left panel: Binary Judgement Task; Middle panel: 4-Category Rating Task; Right panel: Sliding Scale Rating Task.