Evaluating Grammatical Well-Formedness in Large Language Models: A Comparative Study with Human Judgments

Zhuang Qiu, Xufeng Duan, Zhenguang G. Cai
Department of Linguistics and Modern Languages
The Chinese University of Hong Kong
Hong Kong SAR
zhuangharveyqiu@gmail.com, xufeng.duan@link.cuhk.edu.hk
zhenguangcai@cuhk.edu.hk

Abstract

Research in artificial intelligence has witnessed the surge of large language models (LLMs) demonstrating improved performance in various natural language processing tasks. This has sparked significant discussions about the extent to which large language models emulate human linguistic cognition and usage. This study delves into the representation of grammatical well-formedness in LLMs, which is a critical aspect of linguistic knowledge. In three preregistered experiments, we collected grammaticality judgment data for over 2400 English sentences with varying structures from ChatGPT and Vicuna, comparing them with human judgment data. The results reveal substantial alignment in the assessment of grammatical correctness between LLMs and human judgments, albeit with LLMs often showing more conservative judgments for grammatical correctness or incorrectness.

1 Introduction

The rise of LLMs has been extraordinary, demonstrating proficiency across numerous linguistic tasks such as resolving ambiguities (Ortega-Martín et al., 2023), addressing queries (Brown et al., 2020), and facilitating multilingual translation (Jiao et al., 2023). Despite not being initially programmed with a human-like hierarchical syntax structure, these models have managed to identify complex syntactic patterns and generate sophisticated syntactic interpretations (Wilcox et al., 2022; Van Schijndel and Linzen, 2018; Futrell et al., 2019). However, the critical question remains: Do LLMs truly mirror human linguistic cognition? Prominent figures such as Chomsky et al. (2023) argue that LLMs and humans process and understand language differently, while others like Piantadosi (2023) suggest that LLMs might indeed reflect genuine human linguistic processes.

Recent empirical research has become key to this debate. Innovative experiments by Binz and Schulz (2023) subjected GPT-3 to a battery of psychological tests originally crafted to understand facets of human thought processes, ranging from decision-making matrices to reasoning pathways. The outcomes were intriguing, with GPT-3 not just mirroring but at times outperforming human benchmarks in specific scenarios. Similarly, Kosinski (2023) assessed the capacity of LLMs to understand and respond to false-belief scenarios, which are utilized to gauge human empathy and comprehension. Here, the responses from ChatGPT echoed the patterns seen in school-going children, though subsequent research from Brunet-Gouet et al. (2023) voiced concerns about the consistency of such responses. Further, Cai et al. (2023) subjected ChatGPT to a range of psycholinguistic tests, revealing significant alignment in language use between the model and humans, although differences such as in word length preference were observed (e.g., Mahowald et al. (2013)). Qiu et al. (2023) assessed ChatGPT's ability to compute pragmatic implicatures and found that ChatGPT did not demonstrate human-like flexibility in switching between pragmatic and semantic processing. Additionally, ChatGPT did not exhibit the effect of communicative context on the rates of computing scalar implicatures, which is a well-established effect for human participants.

When examining LLM-human similarities, it's crucial to assess the extent to which LLMs' representations of linguistic knowledge align with those of humans. Contemporary linguistic theories often distinguish between the inherent mental systems that enable language comprehension and production and the actual use of language–illustrated by distinctions like "Langue vs. Parole" from Saussure (1916) and "Competence vs Performance" by Chomsky (1965). Grammaticality judgement is a central method to assess linguistic representation competence. Chomsky (1986) highlighted that evidence for linguistic theorizing largely depends on

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 189–198

"the judgements of native speakers". While there are other sources of evidence like speech corpus or acquisition sequences (Devitt, 2006), formal linguists typically favor native speakers' grammaticality intuitions. The prevailing assumption is that our language knowledge comprises abstract rules and principles forming intuitions about sentence well-formedness (Graves et al., 1973; Chomsky, 1980; Fodor, 1981).

Our study focuses on the representation of grammatical well-formedness in LLMs. Recent research on the grammatical capabilities of language models has primarily focused on binary grammaticality judgments using minimally different sentence pairs. Marvin and Linzen (2018) evaluated various language models on syntactic phenomena and found that while models handle local dependencies well, they struggle with non-local dependencies. Similarly, Warstadt et al. (2019) introduced the Corpus of Linguistic Acceptability (CoLA) to test neural network models on binary acceptability judgments, revealing that these models still fall short of human performance on complex syntactic structures. Dentella and et al. (2023) further examined GPT-3's performance on less frequent grammatical constructions, highlighting its limitations in understanding underlying meanings. While these studies have provided valuable insights into the grammatical abilities of language models, they primarily relied on binary judgment tasks and focused on specific syntactic phenomena. In contrast, our preregistered study (https://osf.io/75dtk) adopts a more comprehensive approach by incorporating both binary and graded naturalness judgments, allowing for a finer-grained analysis of language model performance. We collected grammaticality judgment data for over 2400 English sentences from ChatGPT and Vicuna, comparing them with human judgment data. Our findings indicate substantial agreement between ChatGPT and humans regarding grammatical intuition, although noticeable differences were also observed.

2 Experiment 1

In this experiment, we presented ChatGPT and Vicuna with English sentences of varying grammaticality and asked them to judge the sentences as either natural or unnatural. We compared the LLMs' judgement data with human judgement data, examining the similarities and differences in their knowledge of sentence grammaticality.

2.1 Method

We did not recruit human participants ourselves; instead, in all experiments reported in this paper, we utilized datasets from Lau et al. (2017) which were made publicly available. In their study, human participants recruited from Amazon Mechanical Turk performed a series of judgement tasks, including the grammaticality judgement of English sentences. We adopted their data from three judgement tasks as a proxy for humans' grammatical knowledge and later compared the human data with the LLM data that we gathered.

The stimuli used for the judgement tasks were adopted from the experimental materials in Lau et al., which consisted of English sentences of graded grammaticality. These sentences were created following an automated procedure in which texts from the British National Corpus were selected and translated into four different languages: Norwegian, Spanish, Chinese, and Japanese. The sentences were then translated back to English, resulting in 2500 English sentences of various degrees of grammaticality. According to Lau and colleagues, this automated procedure created a ranked distribution of relative grammatical well-formedness in English, with Norwegian texts yielding the best results and Japanese texts yielding the most distorted versions (Lau et al., 2014). Table 1 provides a breakdown of the languages from which experimental sentences were derived.

Language	Counts
English	500
Spanish	491
Japanese	500
Norwegian	480
Chinese	498

Table 1: The number of stimuli derived from each language in Exp 1&2.

A set of five related sentences used in the experiment is shown in Table 2. Note that there were 31 duplicated stimuli in Lau et al. due to some translated sentences being identical across languages. Consequently, we only included the 2469 distinct sentences as our experimental items.

Our data collection followed a "one trial per run" procedure where each interaction with the LLMs contained only a singular experimental trial. Unlike the procedure in Lau et al., where each participant

Language	Text
English	This essential motion cannot take place except in a liquid medium.
Norwegian	This required movement cannot take place except in a liquid medium.
Spanish	This fundamental movement cannot take place except in a liquid medium.
Chinese	The necessary motion in addition to the liquid medium does not occur.
Japanese	This exercise is essential cannot take place except for the liquid medium.

Table 2: A set of related stimuli adopted from Lau et al. (2014). The original English sentence was translated into four languages specified in the "Language" column, and then the translated version was translated back to English, resulting in corresponding sentences in the "Text" column.

was given a multi-item survey, our "one trial per run" method minimized potential biases stemming from preceding trials on the current judgment. This approach also circumvented an issue observed in prior projects where LLMs would occasionally lose track of the instructions midway through. Additionally, the shorter sessions characteristic of the "one trial per run" design were less vulnerable to potential server or connectivity problems.

Judgement data from ChatGPT (gpt-3.5-turbo-0613) and Vicuna (vicuna 13b 1.1) were collected separately using the R package MacBehaviour (Duan et al., 2024). In each trial, we presented ChatGPT or Vicuna with an English sentence from our inventory of stimuli and prompted the model to judge whether the sentence was natural or unnatural. The sentences to be judged were the 2469 distinct experimental items from Lau et al. (2014, 2017). Each sentence was randomly selected following the one trial per run procedure, and we conducted 50 runs for each experimental item. A detailed description of the data collection pipeline is available in the project's preregistration report on the OSF website (`https://osf.io/75dtk`). Following Lau et al. (2014, 2017), LLMs' responses were coded as integer scores, with "1" standing for "unnatural" and "4" for "natural". We combined human judgement data with ChatGPT and Vicuna data and performed two sets of analyses to examine the degree of similarity between human and LLMs' judgements. First, we conducted correlational analyses to examine whether sentences judged as grammatical by humans are more likely to be judged as grammatical by LLMs and vice versa. To do this, we calculated the mean rating score of each sentence stimulus for humans, ChatGPT, and Vicuna, and then computed the correlation coefficients between ChatGPT and humans as well as between Vicuna and humans.

To examine how human and LLMs' ratings were influenced by the grammaticality of the stimuli, we recoded the "natural" and "unnatural" response as "1" and "0" respectively and constructed a Bayesian mixed-effects logistic regression model using the R package `brm` (Bürkner, 2017) with default priors. We treated the logit of the "natural" response as a function of participant type (human vs. ChatGPT vs. Vicuna) and the language from which the stimuli sentences were derived (English vs. Norwegian vs. Spanish vs. Chinese vs. Japanese). The predictors were dummy-coded, with the human data in the English condition being the reference level. Random effects structures were constructed, including item intercepts and slopes:

$$\begin{aligned}&\text{Logit of "natural" response} \sim \\ &1 + \text{participant} \times \text{language} \\ &+(1 + \text{participant} \times \text{language} \mid \text{item})\end{aligned}$$

2.2 Results

The correlation between human and LLM judgements of sentence naturalness is shown in Figure 1. There was a significant correlation between human and ChatGPT judgement ($r = 0.83$, 95% CI = [0.82, 0.84], $p < 0.01$), indicating that sentences judged as natural by humans tended to be judged as natural by ChatGPT as well and vice versa. A significant correlation was also found between human and Vicuna judgement ($r = 0.66$, 95% CI = [0.63, 0.68], $p < 0.01$). According to Cohen (2013), a correlation coefficient of 0.5 or larger represents a strong correlation. A strong and significant correlation between humans and LLMs in their naturalness judgement suggested a considerable extent of shared grammatical knowledge. We also noticed that the ChatGPT-human correlation was stronger than the Vicuna-human correlation, as evidenced by their respective 95% confidence intervals (95% CI = [0.82, 0.84] vs. 95% CI = [0.63, 0.68]).

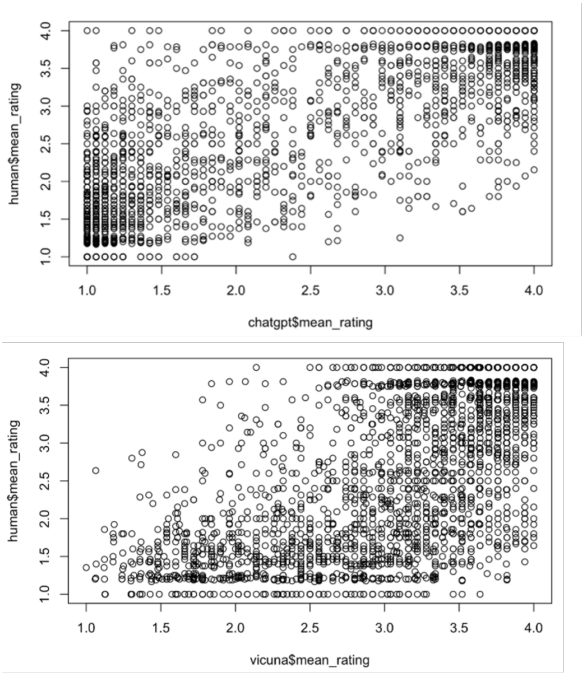

Figure 1: Correlation of naturalness judgement between humans and LLMs in Exp.1. Each point represents the mean rating score of a sentence. Top panel: human vs. ChatGPT. Bottom panel: human vs. Vicuna.

The mixed-effects model, on the other hand, revealed noticeable variations in the naturalness judgement across participant types and the languages from which the stimuli sentences were derived. The baseline for comparison was the human participants' judgement of the original English sentences. Compared with human participants, ChatGPT was more likely to judge the original English sentences as natural ($\beta = 0.39$, 95% CI = [0.19, 0.58]), while Vicuna was less likely to judge the original English sentences as natural ($\beta = -0.32$, 95% CI = [-0.47, -0.18]). For human participants, the probability of a "natural" response decreased for sentences derived from languages other than English, as seen from the negative slopes in the language conditions other than English ($\beta = -1.81$ for Spanish; $\beta = -3.69$ for Japanese; $\beta = -1.55$ for Norwegian; $\beta = -3.06$ for Chinese). Noticeably, this decrease was more dramatic for ChatGPT but reversed for Vicuna. As shown in Figure 2, sentences derived from other languages were rated higher by Vicuna than by human participants.

2.3 Discussion

In this experiment, we investigated the extent to which LLMs share grammatical knowledge with human beings by replicating the binary judgement

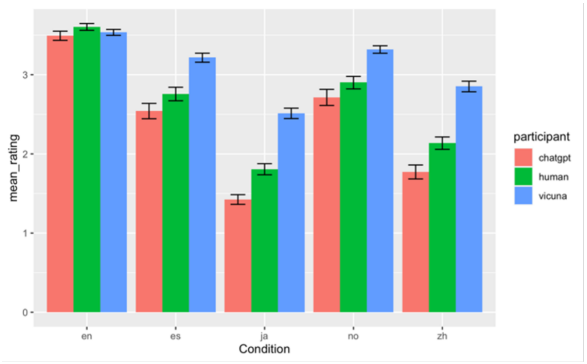

Figure 2: Comparison of mean rating scores across participant types and language conditions in Exp1. An error bar represents the 95% confidence interval of the mean calculated using bootstrapping methods.

task from Lau et al. (2014, 2017), using ChatGPT and Vicuna as participants. We found strong correlations between human and LLM naturalness judgements, with sentences judged to be more natural by human participants generally being judged more natural by LLMs, and vice versa. Adopting the perspective that naturalness judgement is a proxy grammatical knowledge (Lau et al., 2014, 2017), we interpreted this strong correlation as evidence of LLMs and humans sharing a considerable range of knowledge in sentence grammaticality. Though both LLMs' judgements correlated highly with human judgements, the correlation between ChatGPT and humans was stronger than that between Vicuna and humans.

The major difference between human and LLMs lies in their tolerance towards ungrammatical sentences. Compared with human participants, ChatGPT was less tolerant of ungrammaticality, as it gave much lower ratings to sentences auto-translated from languages other than English. On the other hand, Vicuna offered a much higher ratings to those less grammatical sentences than human participants did. This suggests a degree of heterogeneity among current LLMs in that a general label of large language model does not provide detailed information on an individual model's performance in language tasks.

One limitation for this current research design is the response type. The stimuli were created following the procedure that aimed towards a graded profile of sentence grammaticality; however, participants were required to provide binary judgements on the naturalness of the sentences. It is possible that the binary nature of the response type may not be optimal for judging graded grammaticality. We

address this limitation in the second experiment by changing the response type from binary to a ranked measure.

3 Experiment 2

Our second experiment replicated the four-category grammaticality judgement in Lau et al. (2014, 2017) with ChatGPT and Vicuna as the participants. We then compared human performance with that of the LLMs.

3.1 Method

We used the same experimental stimuli as in the first experiment but followed a similar procedure with an important modification: instead of asking LLMs to judge whether a given sentence is natural or unnatural, we instructed them to judge if the sentence is extremely unnatural, somewhat unnatural, somewhat natural, or extremely natural. By employing a four-point Likert scale as the response type, we believe that participants' judgement should be more sensitive to the graded nature of the stimuli. A detailed description of the experimental procedure is available from the project's preregistration report on the OSF website (https://osf.io/75dtk).

We combined LLMs' and human participants' judgements for statistical analysis in which the four-point responses were numerically represented using numbers from one to four. Following the same rationale as the first experiment, we conducted two sets of analyses to compare the grammatical knowledge between humans and LLMs. First, we conducted correlational analyses following the same steps as in Experiment 1. Second, we constructed a Bayesian mixed-effects model that treated the naturalness ratings as a function of participant type (human vs. ChatGPT vs. Vicuna) and the language from which the stimuli sentences were derived (English vs. Norwegian vs. Spanish vs. Chinese vs. Japanese). The predictors were dummy coded with the human data in the English condition serving as the reference level. Random effects structures were constructed, including item intercepts and slopes:

$$\text{judgment} \sim 1 + \text{participant} \times \text{language} + (1 + \text{participant} \times \text{language} \mid \text{item})$$

3.2 Results

There is a significant correlation between human and ChatGPT's judgement ($r = 0.84$, 95% CI = [0.83, 0.85], $p < 0.01$) as well as between human and Vicuna's judgement ($r = 0.49$, 95% CI = [0.45, 0.52], $p < 0.01$). The ChatGPT-human correlation was stronger than the Vicuna-human correlation (Figure 3).

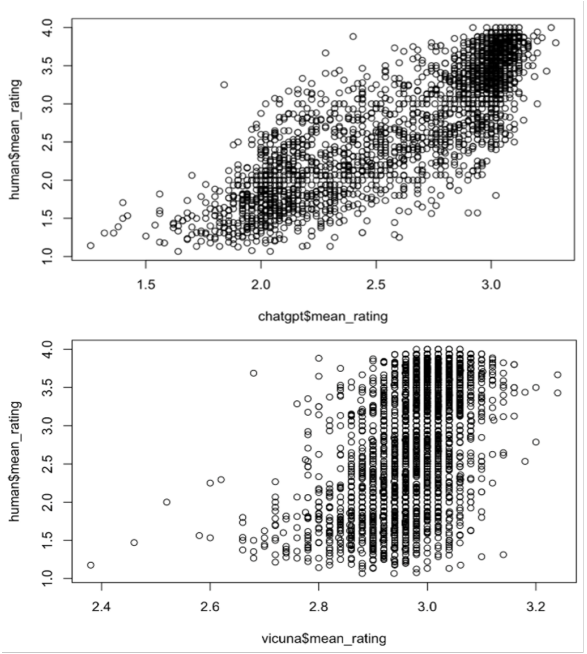

Figure 3: Correlation of naturalness ratings between humans and LLMs in Exp.2. Each point represents the mean rating score of a sentence. Top panel: human vs ChatGPT. Bottom panel: human vs Vicuna.

The mixed-effects model showed that human participants on average judged the original English stimuli (baseline) as between "somewhat natural" and "extremely natural" ($\beta = 3.4$, 95% CI = [3.35, 3.45]). On the other hand, stimuli derived from languages other than English were rated lower than the baseline ($\beta = -0.56$ for Spanish; $\beta = -1.35$ for Japanese; $\beta = -0.42$ for Norwegian; $\beta = -1.06$ for Chinese). Furthermore, for the original English stimuli, human participants' ratings were significantly higher than ChatGPT's ratings ($\beta = -0.44$, 95% CI = [-0.48, -0.40]) and Vicuna's ratings ($\beta = -0.39$, 95% CI = [-0.44, -0.34]).

Additionally, the variation in Vicuna's responses was minimal within a specific language condition and across different language conditions. This is evident in Figure 4 from the small 95% confidence intervals of the mean and from the similar rating scores Vicuna provided across language conditions. Roughly speaking, stimuli sentences were judged as "somewhat natural" (a score of 3) by Vicuna regardless of the actual grammaticality of the sen-

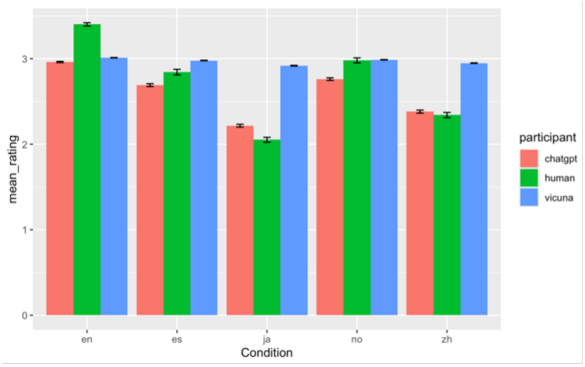

Figure 4: Comparison of mean rating scores across participant types and language conditions in Exp2. An error bar represents the 95% confidence interval of the mean calculated using bootstrapping methods.

tences. This behavior contrasted greatly with ChatGPT, which exhibited a noticeable variation in its naturalness judgement across sentence types. Mimicking the pattern observed among human participants, ChatGPT provided higher naturalness ratings for the original English sentences while provided lower ratings for sentences derived from languages that are typologically further than English such as Japanese and Chinese.

3.3 Discussion

In this study, we replicated Experiment 1 using the same stimuli while modifying the response type to a four-point Likert scale. Major findings of Experiment 1 were successfully replicated. First, we observed strong correlations between LLMs and human participants regarding their ratings of sentences of varying grammaticality. Secondly, significant differences were observed between human participants and LLMs in the ratings of sentences in specific language conditions. These findings suggest that although there is a general agreement between humans and LLMs regarding the relative grammaticality of various sentence structures, the grammatical knowledge of LLMs and human participants differs in terms of the degree of endorsement to specific sentence structures. For sentences deemed very natural by human participants, the naturalness judgements from LLMs were more conservative. Conversely, for sentences judged as "unnatural" by human participants, Vicuna placed them on the "natural" side of the scale. This revealed the heterogeneity among current LLMs previously discussed in Experiment 1. Though both Vicuna and ChatGPT are representative of current LLMs, they nevertheless differed in their performance of naturalness judgement. While ChatGPT closely mimicked human participants in the naturalness rankings of different stimuli categories (en > no/es > zh/ja), Vicuna showed minimal variation in its judgments across stimuli derived from different languages.

4 Experiment 3

This experiment aimed to further our understanding of human and LLMs' knowledge of grammaticality by replicating the previous two experiments using a sliding scale judgement task that was adopted from Lau et al. (2014, 2017).

4.1 Method

Following the design of Lau and colleagues, we instructed our participants, ChatGPT and Vicuna, to rate the naturalness of stimuli sentences with integer scores from 1 (extremely unnatural) to 100 (extremely natural), after which we compared LLMs' judgement data with that of human participants following the same data analysis procedure as the previous two experiments.

The original study of Lau et al. (2017) sampled 250 items from the same inventory of the previous two experiments as the stimuli of the sliding scale judgement task. Two out of the 250 items were duplicated and thus we included 248 unique sentences as the experimental items. Since the experimental items were a subset of the previous experimental items, they were derived from the same automatic procedure as the previous experiments. A breakdown of the languages they were derived from is shown in Table 3.

Language	Counts
English	50
Spanish	44
Japanese	59
Norwegian	45
Chinese	50

Table 3: The number of stimuli derived from each language in Exp 3.

4.2 Results

Consistent with the findings of the previous experiments, we again observed a strong and significant correlation between human and ChatGPT's rating ($r = 0.81$, 95% CI = [0.77, 0.85], $p < 0.01$)

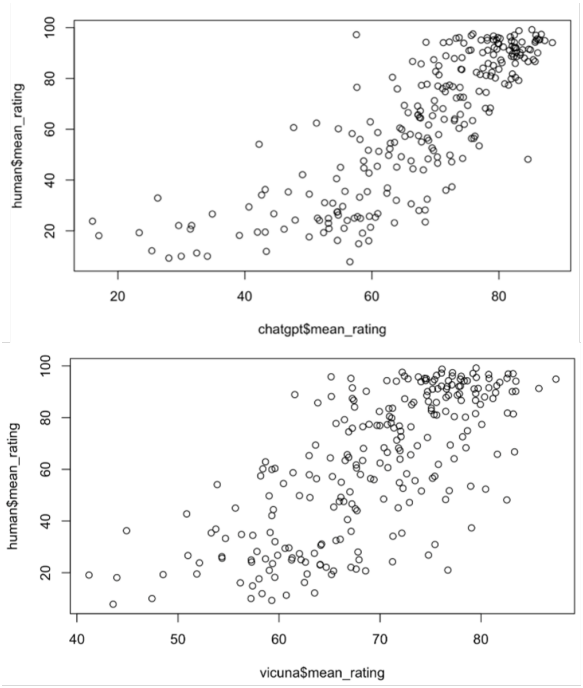

Figure 5: Correlation of naturalness ratings between humans and LLMs in Exp.3. Each point represents the mean rating score of a sentence. Top panel: human vs ChatGPT. Bottom panel: human vs Vicuna.

as well as between human and Vicuna's rating ($r = 0.72$, 95% CI = [0.65, 0.77], $p < 0.01$) in the sliding scale judgement task.

The output of the mixed-effects model was consistent with what we found in Experiment 2. The original English stimuli received a naturalness rating of 86.96 out of 100 from human participants (95% CI = [81.98, 91.85]). Compared with this baseline, the stimuli derived from languages other than English received a lower naturalness rating ($\beta = -15.58$ for Spanish; $\beta = -47.13$ for Japanese; $\beta = -16.76$ for Norwegian; $\beta = -39.33$ for Chinese). Moreover, for the original English stimuli, human participants' ratings were significantly higher than ChatGPT's ratings ($\beta = -9.86$, 95% CI = [-13.6, -6.25]) and Vicuna's ratings ($\beta = -11.46$, 95% CI = [-15.68, -7.14]).

Due to a much smaller number of stimuli adopted in this experiment, the estimates from the mixed-effects model had a larger error term associated with them as compared with the previous experiments. The variation in rating score across different language conditions was specifically noticeable for human participants, as shown from the bootstrapped confidence intervals in Figure 6.

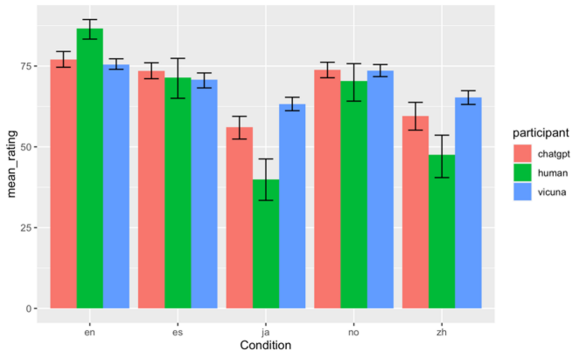

Figure 6: Figure 6 Comparison of mean rating scores across participant types and language conditions in Exp3. An error bar represents the 95% confidence interval of the mean calculated using bootstrapping methods.

4.3 Discussion

In this experiment, we adopted a sliding scale judgement task to elicit finer-grained responses regarding sentence grammaticality. Compared with Experiment 2, the response type of this experiment allowed for more nuanced patterns to occur; nevertheless, major findings of Experiment 2 were replicated. First, human and LLMs shared a considerable amount of grammatical knowledge as evident from the strong correlation in the naturalness rating score. This shared knowledge determines the relative soundness of various sentence structures. For example, the sentence "This essential motion cannot take place except in a liquid medium" is viewed by human and LLMs as more grammatical than the sentence "This exercise is essential cannot take place except for the liquid medium". Second, original sentences from the British National Corpus were rated higher by human participants than by LLMs, while translated sentences, especially those derived from Chinese and Japanese, were rated lower in naturalness by humans than by LLMs. It seems that LLMs are more "conservative" in naturalness judgement compared with human participants. This revealed the differences between human and LLMs in terms of the "distributional knowledge" of sentence grammaticality, which will be further elaborated in the general discussion section.

Compared with Experiment 2, the rating scores in this experiment exhibited larger variations within and across experimental manipulations. For instance, in Experiment 2, Vicuna ratings were largely stable across language conditions; however, in this experiment, we clearly observed a ranked

distribution of Vicuna's judgement. The original English stimuli received the highest ratings, and the Japanese-oriented sentences received the lowest ratings, while the stimuli derived from Norwegian and Spanish received intermediate naturalness ratings. We attributed the increased variation to the reduced size of the stimuli in this experiment, which is only one-tenth of the number of stimuli in the previous experiments.

5 General Discussion

Expanding upon Lau et al. (2014, 2017), our study introduced ChatGPT and Vicuna as LLM counterparts for grammaticality judgment, seeking to determine the extent to which LLMs align with humans in their linguistic knowledge. In general, we observed a strong correlation between human and LLM judgments on the naturalness of sentences, which according to Lau et al. (2014, 2017) suggests a significant overlap in their grammatical knowledge. However, this overlap does not imply equivalence, as our data revealed consistent statistical differences in ratings between humans and LLMs across all three experiments. In the binary judgment task, human participants were more conservative than ChatGPT. Conversely, in the four-point and sliding scale tasks, human participants displayed greater variability in their judgments towards both grammatical and less grammatical sentences compared to ChatGPT. Vicuna's ratings, while generally aligning with those of ChatGPT and humans, exhibited less variation across tasks, suggesting a different processing model.

We posit that a fundamental distinction between human and LLM representations of language lies in the 'distributional knowledge' of sentence grammaticality. Humans acquire an understanding of grammaticality through diverse daily language experiences, enriched by a dynamic array of cognitive and contextual cues. In contrast, LLMs rely predominantly on statistical patterns derived from their training data. This difference in linguistic input is crucial, with human language input being inherently more diverse and dynamic, incorporating a wide array of linguistic registers, dialects, and styles shaped by social interactions and cultural contexts. This exposure enables humans to develop a nuanced and contextually adaptive understanding of language, an aspect of linguistic competence that LLMs with their data-driven learning processes cannot fully replicate (Qiu et al., 2023).

Moreover, human language processing is inherently multi-modal, incorporating auditory, visual, and contextual cues that enhance comprehension and interpretation. This multi-modal integration includes body language, tone, facial expressions, and environmental context, all of which contribute to a rich, intuitive grasp of language nuances and grammaticality. In contrast, LLMs such as ChatGPT and Vicuna process language purely as text tokens, which are sequences abstracted from their communicative contexts. The tokenization process specific to each model's architecture often strips away nuanced information that humans naturally use to infer meaning, leading to potential discrepancies in understanding subtle linguistic cues or complex semantic structures.

Additionally, the cognitive processes in humans, including memory, attention, and inference, dynamically interact during language processing, allowing for a rich contextual interpretation of language that adapts in real-time. This level of cognitive engagement in language processing is not mirrored in current LLM architectures, which primarily rely on recognizing patterns and statistical generalizations from extensive datasets. These fundamental differences imply that the grammaticality of a sentence is judged against different distributions of possible sentence structures by humans and LLMs. Understanding these variations is crucial for recognizing the limitations and potential biases of LLM-generated language assessments. It also underscores the importance of incorporating diverse real-world language data and sophisticated cognitive models into LLM training protocols to improve their linguistic adaptability and judgment accuracy.

6 Conclusion

Our investigation into the alignment of LLMs with human grammaticality judgments has revealed both promising correlations and significant nuances in their linguistic capabilities. While LLMs like ChatGPT and Vicuna can effectively mirror human judgments in broad strokes, discrepancies in sensitivity and the conservativeness of their ratings underscore the importance of careful model selection and calibration for specific linguistic tasks.

7 Limitations

While our study provides valuable insights into the grammatical capabilities of LLMs, it is worth

noting that the experiments were conducted using prompting methods. As Hu and Levy (2023) argued, LLMs may be better judges of grammaticality when evaluated using sentence probabilities rather than prompts. A reviewer suggested that this approach aligns more closely with the langue versus parole (competence vs. performance) distinction. Their findings suggest that using probability measures can yield more accurate grammaticality judgments by LLMs. Future work should replicate our study using probability measures to provide a more comprehensive understanding of LLMs' linguistic capabilities.

References

M. Binz and E. Schulz. 2023. Using cognitive psychology to understand gpt 3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, and D. Amodei. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

E. Brunet-Gouet, N. Vidal, and P. Roux. 2023. Do conversational agents have a theory of mind? a single case study of chatgpt with the hinting false beliefs and false photographs and strange stories paradigms. *HAL Open Science*.

P. C. Bürkner. 2017. brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, 80:1–28.

Z. G. Cai, D. A. Haslett, X. Duan, S. Wang, and M. J. Pickering. 2023. Does chatgpt resemble humans in language use? *arXiv preprint arXiv:2303.08014*.

N. Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press Cambridge MA.

N. Chomsky. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. Praeger Publishers, New York.

N. Chomsky, I. Roberts, and J. Watumull. 2023. Noam chomsky: The false promise of chatgpt. *The New York Times*.

Noam Chomsky. 1980. Rules and representations. *Behavioral and brain sciences*, 3(1):1–15.

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Routledge.

Laura Dentella and et al. 2023. Testing ai performance on less frequent aspects of language reveals insensitivity to underlying meaning. *arXiv preprint arXiv:2302.12313*.

M. Devitt. 2006. Intuitions in linguistics. *British Journal for the Philosophy of Science*, 57:481–513.

Xufeng Duan, Shixuan Li, and Zhenguang G Cai. 2024. Macbehaviour: An r package for behavioural experimentation on large language models. *arXiv preprint arXiv:2405.07495*.

J. Fodor. 1981. Introduction: Some notes on what linguistics is about. *In The Language and Thought Series*, pages 197–207.

R. Futrell, E. Wilcox, T. Morita, P. Qian, M. Ballesteros, and R. Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. *arXiv preprint arXiv:1903.03260*.

C. Graves, J. J. Katz, Y. Nishiyama, S. Soames, R. Stecker, and P. Tovey. 1973. Tacit knowledge. *The Journal of Philosophy*, 70(11):318–330.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

W. Jiao, W. Wang, J. T. Huang, X. Wang, and Z. Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

M. Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

J. H. Lau, A. Clark, and S. Lappin. 2014. Measuring gradience in speakers' grammaticality judgements. *In Proceedings of the annual meeting of the cognitive science society*, 36(36).

J. H. Lau, A. Clark, and S. Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive science*, 41(5):1202–1241.

K. Mahowald, E. Fedorenko, S. T. Piantadosi, and E. Gibson. 2013. Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2):313–318.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

M. Ortega-Martín, Ó. García-Sierra, A. Ardoiz, J. Álvarez, J. C. Armenteros, and A. Alonso. 2023. Linguistic ambiguity analysis in chatgpt. *arXiv preprint arXiv:2302.06426*.

S. T. Piantadosi. 2023. Modern language models refute chomsky's approach to language. *Lingbuzz Preprint lingbuzz/007180*.

Z. Qiu, X. Duan, and Z. Cai. 2023. Does chatgpt resemble humans in processing implicatures? *Proceedings of the 4th Natural Logic Meets Machine Learning Workshop*.

F. Saussure. 1916. Cours de linguistique générale. *Paris: Payot*.

M. Van Schijndel and T. Linzen. 2018. Modeling garden path effects without explicit hierarchical syntax. *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

E. G. Wilcox, R. Futrell, and R. Levy. 2022. Using computational models to test syntactic learnability. *Linguistic Inquiry*, pages 1–88.