Linguistically-Informed Evaluation of LLMs on Acceptability Judgments in a Forced-Choice Paradigm

Ziyue Liu and Nils Reiter
Department of Digital Humanities
University of Cologne
ziyue.liu@uni-koeln.de
nils.reiter@uni-koeln.de

Abstract

Evaluating the grammatical abilities of large language models (LLMs) is important for both NLP and linguistic theory. We investigate the ability of large language models (LLMs) to perform acceptability judgments in a forced-choice paradigm. We evaluate a subset of LLMs on 150 minimal sentence pairs sampled from Linguistic Inquiry and categorized using BLiMP linguistic phenomena. Our results show that while LLMs approximate human judgments, performance varies across models and phenomenon types, with stronger alignment on morphosyntactic phenomena than on linguistically and semantically demanding phenomena. Prompting strategies have minimal impact.

1 Introduction

Recently, Large Language Models (LLMs) have shown remarkable capabilities in NLP tasks including generating human-like texts (Devlin et al., 2019; Brown et al., 2020; Touvron et al., 2023; Argyle et al., 2023; Naveed et al., 2025). Yet whether LLMs have internalized human-like grammatical knowledge remains contested, with studies reporting both close alignment with human judgments (Hu et al., 2024; Qiu et al., 2025) and systematic divergences driven by response biases and inconsistencies (Dentella et al., 2023). This disagreement is partly attributable to the multidimensional nature of linguistic evaluation: LLM performance varies systematically across linguistic phenomena (Warstadt et al., 2020; Dentella et al., 2023), model families (Hu et al., 2024), and prompt formulations (Pichler et al., 2025).

Previous work has largely evaluated LLMs on linguistic tasks without a theoretically grounded or linguistically-informed framework. Most notably, Qiu et al. (2025) replicated Sprouse et al.'s three formal judgment paradigms with ChatGPT, finding 73 – 95% alignment with linguists but notably weaker performance on the forced-choice (FC) task. Critically, however, these studies are limited to a single model under a fixed prompt, leaving a question about how grammatical alignment varies across models, prompts, and linguistic domains. We address this gap by adopting the FC paradigm as our primary experimental method, and extending previous study in three main aspects: (1) linguistically-informed taxonomy; (2) multiple large language models; (3) prompting strategies (linguist vs. layperson; detailed task instruction vs. brief-core task instruction; user-prompt vs. user-system prompt).

The major contributions of this study include:

1. We present a categorization of the experimental items gathered by Sprouse et al. (2013) adopting the categories employed in the BLiMP (Warstadt et al., 2020), thereby allowing for linguistically-informed LLM evaluation datasets. We analyse the LLM responses and performance with respect to these categories.

2. We extend the experiments conducted by Qiu et al. (2025) to newer and larger models, under strategically designed prompt templates.

3. We document and employ an evaluation framework that not only compares general tendencies, but takes into account the whole distribution of human responses. For example, a 'good' LLM not only produces the same mean response as a group of humans, but also the same distribution of responses.

2 Related Work

2.1 Acceptability Judgment Tasks

Acceptability judgment tasks are widely used as a paradigm in empirical linguistic. It is a type

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 1177–1189
July 2-7, 2026 ©2026 Association for Computational Linguistics

of rating task: participants are asked to rate the acceptability of a sentence according to their subjective sense. In contrast to grammaticality as a property of grammar, acceptability reflects participants perceptions of the naturalness and comprehensibility of linguistic forms (Chomsky, 2014). Previous work has measured acceptability using rating-based paradigms (e.g., Likert-scale judgments) as well as forced-choice designs (Sprouse et al., 2013). Following their work, this study adopts forced-choice (FC) tasks due to: (1) FC tasks can provide granular details that are invisible in Likert Scale experiments (Stadthagen-González et al., 2018). (2) FC tasks reduce individual bias on scale use (Schütze et al., 2014). (3) FC tasks are easy to deploy (Schütze et al., 2014). (4) FC tasks are more in line with language processing (Cowart, 1997).

2.2 Simulating LLMs As Participants

Recently, large language models (LLMs) have been increasingly used as simulated participants in various research fields (Dillion et al., 2023; Amouyal et al., 2024; Gao et al., 2025; Lin, 2025). This paradigm has also been adopted in linguistics, where LLMs are treated as in silico participants and probed with controlled experiments to assess their language use (Cai et al., 2024). In particular, judgment-based settings such as acceptability (and plausibility) judgments have been widely used to evaluate models grammatical knowledge (Amouyal et al., 2024; Qiu et al., 2025; Ide et al., 2025). While models often produce human-like responses, it remains unclear to what extent such behavior reflects human-like linguistic cognition or supports reliable generalization (Katzir, 2023; Harding et al., 2024). However, LLM-based simulations can be sensitive to models, prompting, and linguistic phenomena (Pichler et al., 2025; Warstadt et al., 2020). These limitations motivate more systematic evaluations of LLMs' performance as participants in linguistic experiments.

2.3 Human Acceptability Judgments and BLiMP

Sprouse et al. (2013) compared informal and formal methods for collecting acceptability judgments across multiple experimental formats, including magnitude estimation, Likert scales, and forced-choice tasks. Their stimuli were sampled from papers published in *Linguistic Inquiry* between 2001 and 2010, spanning 150 linguistic phenomena. For each phenomenon, the original authors extracted one grammatical sentence and one ungrammatical counterpart. In addition, seven further variants were created for each phenomenon, resulting in a total of 2,400 sentences.

Building on this dataset, Qiu et al. (2025) replicated the study using ChatGPT-3.5 and found that the model's judgments aligned with human responses on the majority of phenomena. However, their analysis did not examine these phenomena in depth, nor did it apply a linguistically informed taxonomy. In the present study, we reclassify the stimuli into a set of linguistically informed phenomenon types based on the BLiMP taxonomy (Warstadt et al., 2020), enabling a more linguistically informed analysis of model behavior across categories (see Appendix D for details).

3 Methodology

3.1 Experiments Design

We conducted two experiments to assess LLMs linguistic knowledge. Experiment 1 evaluates how well different LLMs align with human acceptability judgments on the dataset of Sprouse et al. (2013) and is an extension of the experiments conducted by Qiu et al. (2025). Experiment 2 examines how LLM performance is affected by prompt design, comparing role-based prompting (linguist vs. lay participant), instruction-controlled prompts, and usersystem formats. Across two experiments, we applied the same evaluation framework and experimental settings, while varying the prompt instructions.

The human acceptability judgments used in Experiment 1 were taken from Sprouse et al. (2013), who recruited 307 native English speakers through Amazon Mechanical Turk (AMT). For the model evaluations, we collected responses from several LLMs such as ChatGPT-3.5 Turbo, ChatGPT-4o, LlaMa 4 Maverick, and Claude Sonnet-4.5 via one API key by OpenRouter, a third-party provider, and ChatGPT-OSS-120B via API key by University of Cologne. These models will henceforth be referred to as GPT-3.5, GPT-4o, GPT-OSS, Llama, and Claude, respectively.

3.2 Evaluation Metrics

We coded judgments as 1 for selecting the more acceptable sentence and 0 for selecting the less acceptable sentence, excluding trials where participants responded "equally acceptable" or "un-

known". We evaluated human judgments and models' performance using the accuracy rate, defined as the proportion of trials in which the participant selected the more acceptable sentence. In each sentence pair, the more acceptable sentence was determined based on judgments reported in publications in *Linguistic Inquiry*.

Accuracy Rate For each model and human, and each item, we computed the accuracy rate as:

$$R_{\text{item}} = \frac{1}{N}\sum_{s=1}^{N}\mathbf{1}[J(s)=1] \quad (1)$$

where R_{item} denotes the proportion of selecting more acceptable sentence for a given item, s indexes individual sessions, $J(s) \in \{0,1\}$ is the judgment for session s, and N is the total number of sessions. This yielded one proportion per participant per item, reflecting how consistently a given participant produced the more acceptable for that item across repeated sessions.

Aggregated Accuracy Rate Item-level proportions were further aggregated to the phenomenon-type level by computing the mean accuracy rate across all items within each phenomenon type for each participant:

$$R_{\text{phenomenon}} = \frac{1}{N}\sum_{i=1}^{N} R_{item} \quad (2)$$

where $R_{\text{phenomenon}}$ is the proportion aggregated over all N items, and R_{item} is as defined in Equation 1. This yielded one mean proportion per participant per phenomenon-type, which was used to compare performance across participants.

Wassertein Distance There are several metrics used to compare distributions. We avoided the KullbackLeibler divergence (Kullback and Leibler, 1951), which is sensitive to zero probabilities (Cover, 1999), and instead used the Wasserstein distance (WD; also known as Earth Movers Distance; Vaserstein 1969) to compare distributions of accuracy rate on phenomenon-type level and item level.

$$WD(P,Q) = \int_{-\infty}^{\infty} |F_P(x) - F_Q(x)|\, dx \quad (3)$$

Intuitively, Earth Mover's distance measures how much dirt has to be moved to transform one pile into one with another shape (but with the same total mass). As a proper metric, $WD \geq 0$, with $WD = 0$ if and only if the two distributions are identical. The upper bound depends on the specific distributions under consideration. In our setting, where accuracy rates are concentrated within the interval $(0.75, 1.0)$, the theoretical maximum of WD is 0.25, which is attained only when the two distributions place all probability mass at opposite ends of this interval. Based on this range, we interpret $WD \leq 0.05$ as indicating highly similar distributions, $0.05 < WD < 0.10$ as indicating moderately similar distributions, and $WD \geq 0.10$ as indicating substantially diverged distributions.

4 Experiment 1

4.1 Experimental Setup

The experimental items in our experiment were adapted from the materials used by Sprouse et al. (2013). Qiu et al. (2025) reused this data set and reported that two of the phenomenon pairs were duplicated. Following their observation, we removed these two duplicated pairs from our materials; thus, there are 148 phenomenon pairs, in a total of 1184 pairs of sentences. To enable a linguistic theoretical evaluation of LLMs' linguistic knowledge, we used two LLMs (ChatGPT-5 and Claude Sonnet-4.5) and one PhD student to classify the 150 phenomena from Sprouse et al.'s data set followed the classification scheme used in the BLiMP benchmark (see more details in Appendix B and Appendix D).

In this experiment, all of LLMs were carried out 50 sessions for each item, indicating that each query was independent and has been sent 50 times, following Qiu et al. (2025). Temperature was set to 0.5 for all models. For all models, we reused the original prompt template provided by Qiu et al. (P1). The models were asked to act as a "linguist" with detailed and long instructions (see the details of prompts in Appendix A).

4.2 Results

4.2.1 Overall Alignment

Figure 1 shows aggregated accuracy rate. Human judgments center around 0.93, with Ellipsis ($\approx$ 0.80) emerging as the most challenging linguistic phenomenon. Most LLMs exceed the human baseline, suggesting a general tendency to-

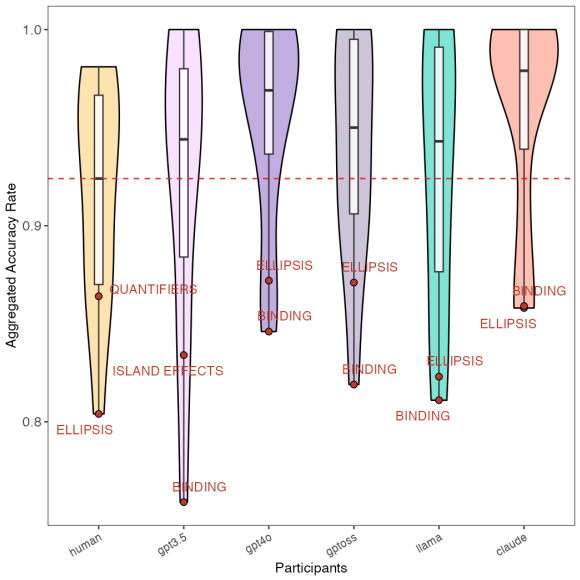

Figure 1: The aggregated accuracy rates of human and models. Red dots indicate each model's worst-performing linguistic phenomena. The red dashed line marks the human median as a reference baseline.

ward over-acceptance. Across models, binding is consistently the weakest phenomenon. GPT-3.5 stands out with overall lower and more variable performance, whereas GPT-OSS and Llama track human patterns more closely. This is further confirmed by the Wasserstein distances (see Table 1), where Llama (WD = 0.019) shows the closest alignment to human responses, with GPT-OSS (WD = 0.021) following. Notably, some phenomena that are problematic for LLMs are not reflected in human acceptability judgments. This mismatch suggests that model-human agreement at an aggregate level may mask qualitatively different sensitivities to linguistic phenomena, raising questions about what kind of alignment these models actually achieve. To investigate this further, we examine item-level response variation within each phenomenon.

4.2.2 Item-level Analysis

Figure 2 highlights item-level variation within phenomena. Human responses, while generally high on average, span a wide range across items, indicating considerable heterogeneity within several categories. Llama shows a similar pattern, whereas GPT-4o, Claude, and GPT-OSS exhibit strong ceiling effects, with responses compressed near 1.0, suggesting less discriminating behavior compared to human judgments. GPT-3.5 stands out with highly dispersed responses across the full range, reflecting low consistency overall. Variability is particularly pronounced for phenomena such as Binding and Ellipsis, where both humans and some models show wide item-level spread.

Model	WD to Human
GPT-3.5	0.023
GPT-4o	0.038
GPT-OSS	**0.021**
Llama	0.019
Claude	0.042

Table 1: Wasserstein Distance (WD) between each LLM's response and human judgments, computed on the distribution of accuracy rates across phenomenon types.

Figure 3 reports item-level Wasserstein Distance between human judgments and each model across phenomenon types. Compared to the overall WD values in Table 1, item-level WD values are generally lower, consistent with the narrower range observed at finer granularity. Across most linguistically informed phenomena, models exhibit high to moderate similarity to human judgments ($WD < 0.10$). The main exceptions are Binding and Quantifiers, where several models show substantially diverged distributions ($WD \geq 0.10$). In particular, Binding yields relatively high divergence for GPT-3.5 ($WD = 0.15$), GPT-4o ($WD = 0.12$), GPT-OSS ($WD = 0.12$), Llama ($WD = 0.12$), and Claude ($WD = 0.11$). Similarly, Quantifiers shows diverged distributions for GPT-4o ($WD = 0.13$), GPT-OSS ($WD = 0.13$), Llama ($WD = 0.11$), and Claude ($WD = 0.14$). Taken together, these results indicate that LLMs broadly approximate human grammatical knowledge; however, they diverge from humans in specific linguistic phenomena. These findings raise the question of whether such differences persist across prompt conditions a question addressed in the following experiment.

5 Experiment 2

5.1 Experimental Setup

This experiment extends Experiment 1 by introducing different prompting strategies to examine the influence of prompt phrasing and assess the robustness of large language models to prompting strategies. Other settings are the same as Experi-

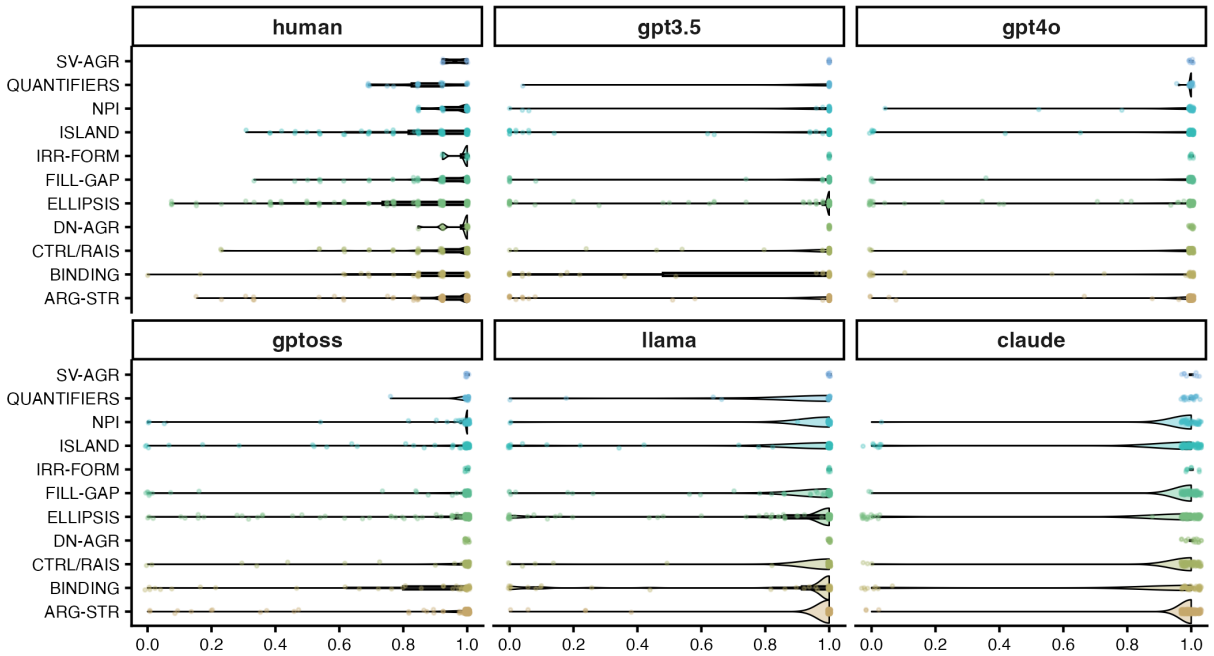

Figure 2: Item-level distributions of accuracy rate across 11 linguistic phenomena for human participants and five LLMs under the linguist prompt. Each dot represents a single item.

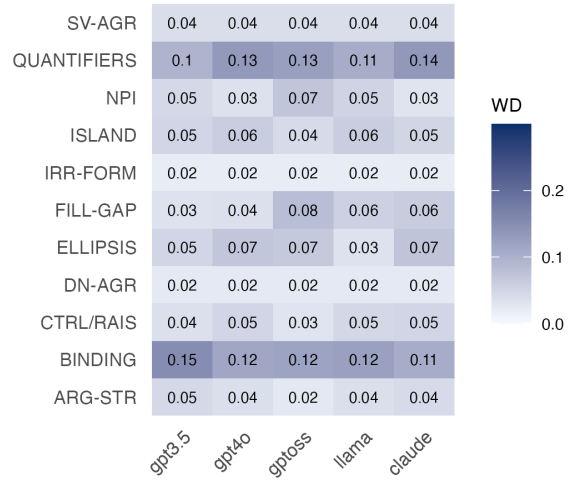

Figure 3: Wasserstein Distance (WD) between each LLM's response and human judgments, computed on the distribution of accuracy rates across items.

ment 1.

Layperson Prompt (P2) Compared to P1 (based on Qiu et al., 2025) in Experiment 1, we streamlined the task instruction by removing non-essential explanations and reducing overall length, while preserving the core task description. Furthermore, we re-framed "linguist" with "participant" to see the effects of laypeople in the judgments.

System-User Prompt (P3) This version refined P2 by incorporating system-user prompt. In this variant, the task instruction and sentences pair are included in the system prompt while the role specification is provided in the user prompt. Such formatting is commonly used to organize prompts with multiple components, and may improve consistency and accuracy in model behavior.

5.2 Results

5.2.1 Overall Prompting Effects

Figure 4 shows model alignment with human acceptability judgments across three prompt variants. Among LLM participants, Llama produced medians near the human baseline (red dashed line $\approx$ 0.92), Claude consistently yields higher medians, often approaching ceiling levels. Models also differ markedly in distributional spread. GPT-3.5 and Llama exhibit substantially wider distributions, with lower tails extending well below the human range, indicating less stable and more variable behavior across items. These patterns are further reflected in the Wasserstein Distances reported in Table 2. Llama achieves the lowest divergence across all conditions (*WD* = 0.015–0.019), with best performance under the linguist-framed prompt (P3), whereas Claude remains the most divergent (*WD* = 0.042–0.052) despite its high median proportions.

In short, prompt effects are strongly model-dependent. GPT-3.5 and Claude align most closely

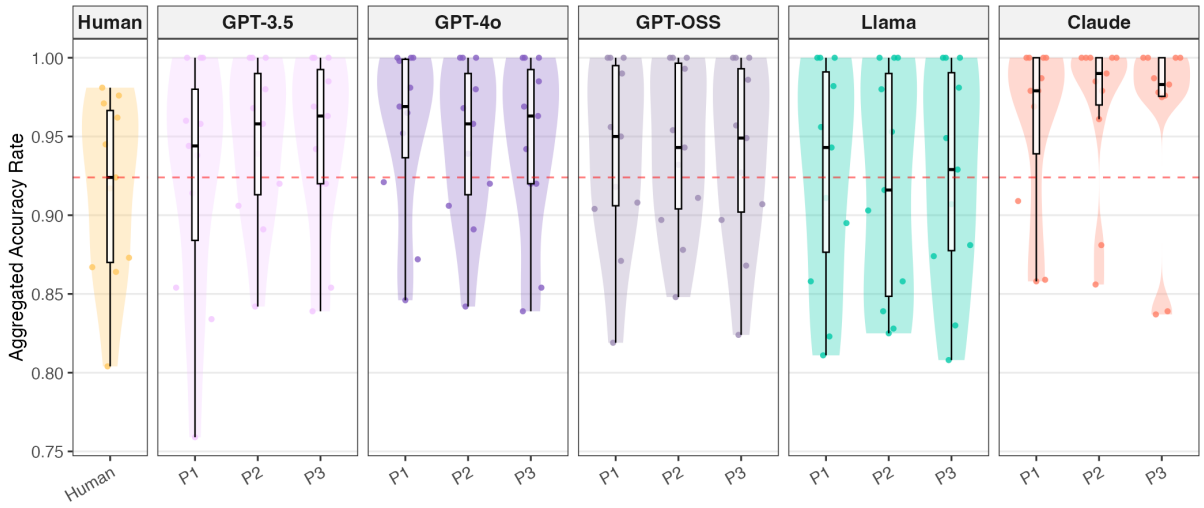

Figure 4: Model performance across three prompt variants (P1: linguist role with detailed instructions; P2: layperson role with minimal instructions; P3: refined P2 with system and user prompt) compared to the human median (red dashed line) at the phenomenon type level. Each dot represents a single phenomenon type.

Model	P1	P2	P3
GPT-3.5	0.023	0.029	0.030
GPT-4o	0.038	0.029	0.030
GPT-OSS	0.021	0.025	0.021
Llama	**0.019**	**0.020**	**0.015**
Claude	0.042	0.052	0.050
Best Model	*Llama*	*Llama*	*Llama*

Table 2: Wasserstein distance (WD) between each model's response distribution and the human baseline, aggregated at the phenomenon-type level across three prompt conditions (P1–P3).

Model	P1	P2	P3
GPT-3.5	0.73**	0.82**	0.78***
GPT-4o	0.73**	0.82**	0.78***
GPT-OSS	0.63***	0.66***	0.67***
Llama	**0.92*****	**0.88****	**0.89****
Claude	0.79**	0.78**	0.71**

Table 3: Pearson correlations between each LLM and Human across prompting strategies on the phenomenon type level

* $p < .05$, ** $p < .01$, *** $p < .001$

with human judgments under P1, GPT-4o perform best under P2, Llama align most closely with human judgments under P3, both P2 and P3 work best for GPT-4o, suggesting that linguist role and detailed instruction don't improve alignment consistently. To gain deeper insight into the relationship between human judgments and model performance, we further examine theory-driven linguistic phenomena across models and prompting conditions in the following analysis.

5.2.2 Theory-Driven Linguistic Analysis

Table 3 reports Pearson correlations between human and model judgments. Correlations are consistently strong across models and prompt conditions (r = 0.67 – 0.92), indicating substantial alignment with human judgments. They vary little across prompts (maximum within-model difference < 0.10) and remain stable across aggregation levels, with all phenomenon-level correlations reaching significance ($p < 0.001$). Llama exhibits the strongest correlation with human (r = 0.88 – 0.92) across all prompting strategies. Notably, differences between models are minimal, suggesting that this alignment is a general property rather than model-specific. Additional item level analysis can be found in table 4 (see Appendix C). Although correlations indicate alignment, they do not capture differences in the overall distribution of judgments. Thus, we addressed this question as follows.

Figure 5 presents performance broken down by linguistic phenomenon. In line with Table 2, Llama and GPT-OSS track human proportions most closely. Prompt condition has little impact at this level, with estimates clustering tightly across conditions. Large language models closely match the human baseline on simpler morphosyntactic phenomena, such as determiner-noun agreement and subjectverb agreement. In contrast, they ex-

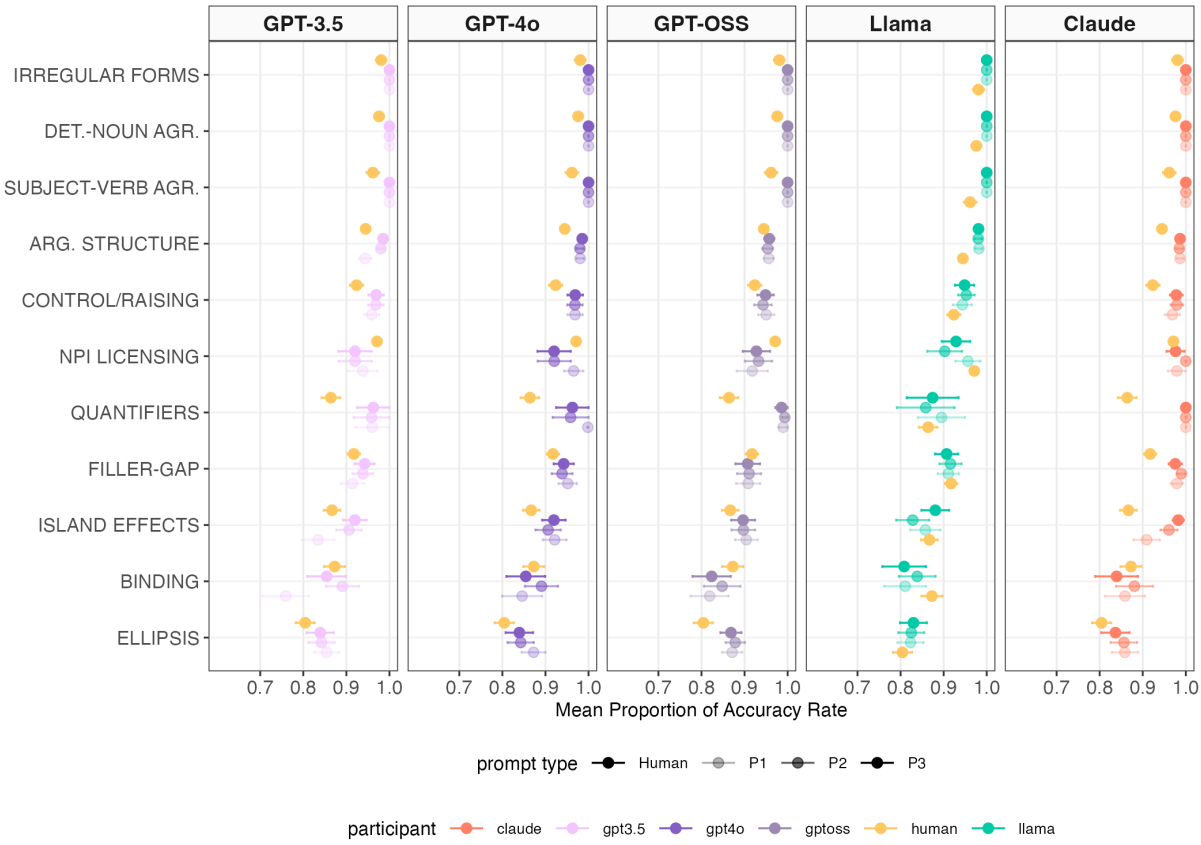

Figure 5: Mean proportions of accuracy rate across phenomenon types. Dots indicate mean proportions, with error bars representing ± 1 standard error. Phenomenon types are ordered by human judgments trends (high $\rightarrow$ low). Colors denote participants, and transparency indicates prompt conditions.

hibit substantially larger error ranges on more complex phenomena such as Binding and Quantifiers. LLMs perform both lower accuracy and less stability in Binding. Quantifiers stands out: Human judgments on Quantifiers were comparatively consistent, suggesting a stable and shared interpretive mechanismin contrast, some LLMs showed a higher overall performance but larger variance, pointing to a lack of stable compositional generalization in quantifier processing. We deeply identified the item and discussed in section 6 (see the examples in (1) and (2)).

In short, these findings indicate that divergence between model and human behavior varies across theory-driven linguistic phenomena. LLMs exhibit less stable performance on linguistically complex cases and limited compositional generalization, suggesting that different phenomena involve different underlying sources of difficulty.

6 Discussion

Our results show that large language models broadly align with human acceptability judgments in the forced-choice paradigm, but this alignment varies across models and evaluation dimensions. Correlation analyses indicate generally strong associations with human responses, yet distributional metrics reveal meaningful differences in how closely models approximate human-like behavior. Llama shows the closest match to human judgments, with low divergence and stable performance across prompt conditions. Claude exhibit ceiling effects; however, this pattern of outperforming does not necessarily indicate good alignment with humans, since human performance is not perfect. Prompt strategies have model-specific effects, with no systematic improvement from "linguists" role and more detailed instructions.

Despite overall alignment, models and humans differ in which linguistic phenomena are challenging. Some phenomena are consistently well-handled across models, while others reveal notable divergence from human behavior. For instance, LLMs generally show high alignment,

with narrow error ranges in simple morphosyntactic phenomena such as subject-verb agreement and determiner-noun agreement, which is inline with previous findings (Warstadt et al., 2020). Notably, some models (e.g., Claude, GPT-4o, GPT-OSS) outperform humans on island effects, contrasting with earlier work reporting that LLMs struggle with such phenomena (Warstadt et al., 2020). However, LLMs exhibit substantially greater variability in quantifier interpretation and lower performance in more complex phenomena such as binding and ellipsis, which are discussed in detail below.

The instability in quantifier interpretation observed in LLMs aligns with previous work (Cui et al., 2022; Gupta, 2023). This may stem from the fact that quantifier interpretation requires complex abilities such as sensitivity to scope relations licensed by syntactic-semantics interface and logical entailments (May, 1985; Heim and Kratzer, 1998), and compositional meaning construction (Barwise and Cooper, 1981; Partee et al., 2012). Specifying on the items example (see (1a) and (1b)), sentence (1a) is acceptable because **all** appears after the verb, separated from its associated noun phrase *the men*. This is a well-known construction called quantifier float, where the quantifier is pronounced in a position left behind when the subject moves (Sportiche, 1988). Sentence (1b) contains two occurrences of **all** – one inside the subject noun phrase and another after the verb, yielding redundancy. This finding further reveal the gap between human and LLMs: unlike human, LLMs primarily approximate linguistic competence via distributional pattern matching (Bender and Koller, 2020).

(1a) The men seem to have all eaten supper.

(1b) *All the men seem to have all eaten supper.

Here, * indicates unacceptable or less acceptable sentence, and without * indicates acceptable or more acceptable sentence.

Structure of (1a):

[[The men] [seem [[have [all [eaten supper]]]]]]

Structure of (1b):

[[All the men] [seem [[have [all [eaten supper]]]]]]

Binding stands out as challenging for LLMs, showing low performance and high variability. Thus, we examine the stimuli with the poorest performance in Binding (see (2a) and (2b)). These items involve a structurally complex dependency with control and embedding, where the reflexive is bound by an implicit subject (PRO) introduced by control and co-referential with Jack (Landau, 2013). Under Binding Theory (Chomsky, 1981), the correct antecedent is determined by hierarchical structure rather than linear proximity, requiring the reflexive to be linked to Jack rather than the linearly closer Sally. Additional difficulty arises from a mismatch between structural relations and surface prominence, as Sally is more recently mentioned and more salient in the discourse. The observed errors suggest that models tend to fall back on surface-level cues such as linear order and prominence, which leads to incorrect antecedent assignment in structurally complex configurations.

(2a) Jack asked Sally to be allowed to take care of himself.

(2b) *Jack asked Sally to be allowed to take care of herself.

Here, * indicates unacceptable or less acceptable sentence, and without * indicates acceptable or more acceptable sentence.

Structure of (2a):

Jack_i asked Sally_j [to be allowed [to take care of himself_i]]

Structure of (2a):

Jack_i asked Sally_j [to be allowed [to take care of herself_j]]

In summary, this study provides a measurement framework for evaluating LLMs as simulated participants in linguistic experiments. Comparing model and human judgments across theory-driven linguistic phenomena reveals that LLM behavior varies substantially across phenomena and does not always generalize robustly to linguistically demanding cases. More broadly, our findings highlight the value of linguistically grounded evaluation for understanding the strengths and limitations of contemporary language models.

Conclusion

Inspired by the grammatical evaluation of LLMs, this study offers a linguistically informed categorization, enabling a theory-driven analysis of model behavior. We show that, despite generally high correlations with human responses, distributional analyses reveal meaningful divergences, including ceiling effects, increased variability, and phenomenon-dependent instability. In particular, while LLMs closely approximate human judgments on relatively morphosyntactic phenomena, they exhibit less stable and less systematic behavior on linguistically and semantically demanding phenomena such as Binding, Ellipsis, and Quantifiers. Finally, we demonstrate that model performance is less sensitive to prompting strategies. Together, these findings provide a more nuanced picture of LLM behavior and highlight a persistent gap between surface-level fluency and robust semantic generalization.

Limitations

In this study, we focus on a subset of large language models (LLMs). While many LLMs can approximate human acceptability judgments, their performance varies across models, and future work should evaluate a broader set. Our stimuli were sampled from *Linguistic Inquiry* and may not reflect everyday language, potentially limiting the models exposure and affecting their judgments. Moreover, we did not balance the difficulty across different linguistically-informed phenomenon types, which may influence the observed patterns. These factors suggest caution in generalizing our findings to other models or linguistic contexts.

Acknowledgments

The research for this paper was carried out within the collaborative research centre SFB 1252 Prominence in Language (Project-ID 281511265) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) at the University of Cologne. We would also like to thank Mark Ellison and Janis Pagel for their feedback and suggestions.

References

Samuel Amouyal, Aya Meltzer-Asscher, and Jonathan Berant. 2024. Large language models for psycholinguistic plausibility pretesting. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 166–181.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Jon Barwise and Robin Cooper. 1981. Generalized quantifiers and natural language. In *Philosophy, language, and artificial intelligence: Resources for processing natural language*, pages 241–301. Springer.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zhenguang Cai, Xufeng Duan, David Haslett, Shuqi Wang, and Martin Pickering. 2024. Do large language models resemble humans in language use? In *Proceedings of the workshop on cognitive modeling and computational linguistics*, pages 37–56.

Noam Chomsky. 1981. Lectures on government and binding, foris, dordrecht. *ChomskyLectures on Government and Binding1981*.

Noam Chomsky. 2014. *Aspects of the Theory of Syntax*. 11. MIT press.

Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.

Wayne Cowart. 1997. *Experimental syntax*. Sage.

Ruixiang Cui, Daniel Hershcovich, and Anders Søgaard. 2022. Generalized quantifiers as a source of error in multilingual nlu benchmarks. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 4875–4893.

Vittoria Dentella, Fritz Günther, and Evelina Leivada. 2023. Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, 120(51):e2309583120.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.

Rena Wei Gao, Xuetong Wu, Tatsuki Kuribayashi, Mingrui Ye, Siya Qi, Carsten Roever, Yuanxing Liu, Zheng Yuan, and Jey Han Lau. 2025. Can llms simulate l2-english dialogue? an information-theoretic analysis of l1-dependent biases. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4355–4379.

Akshat Gupta. 2023. Probing quantifier comprehension in large language models: Another example of inverse scaling. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 56–64.

Jacqueline Harding, William DAlessandro, NG Laskowski, and Robert Long. 2024. Ai language models cannot replace human research participants. *Ai & Society*, 39(5):2603–2605.

Irene Heim and Angelika Kratzer. 1998. Semantics in generative grammar.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.

Yusuke Ide, Yuto Nishida, Justin Vasselli, Miyu Oba, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2025. How to make the most of llms grammatical knowledge for acceptability judgments. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7416–7432.

Roni Katzir. 2023. Why large language models are poor theories of human linguistic cognition. a reply to piantadosi (2023). *Manuscript. Tel Aviv University. url: https://lingbuzz. net/lingbuzz/007190*.

S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Idan Landau. 2013. *Control in generative grammar: A research companion*. Cambridge University Press.

Zhicheng Lin. 2025. Large language models as psychological simulators: A methodological guide. *arXiv preprint arXiv:2506.16702*.

Robert May. 1985. *Logical form: Its structure and derivation*, volume 12. MIT press.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72.

Barbara BH Partee, Alice G Ter Meulen, and Robert Wall. 2012. *Mathematical methods in linguistics*, volume 30. Springer Science & Business Media.

Axel Pichler, Janis Pagel, and Nils Reiter. 2025. Evaluating llm-prompting for sequence labeling tasks in computational literary studies. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 32–46.

Zhuang Qiu, Xufeng Duan, and Zhenguang G Cai. 2025. Grammaticality representation in chatgpt as compared to linguists and laypeople. *Humanities and Social Sciences Communications*, 12(1):1–15.

Carson T Schütze, Jon Sprouse, Robert J Podesva, and Devyani Sharma. 2014. *Research methods in linguistics*. Cambridge University Press Cambridge.

Dominique Sportiche. 1988. A theory of floating quantifiers and its corollaries for constituent structure. *Linguistic inquiry*, 19(3):425–449.

Jon Sprouse, Carson T Schütze, and Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. *Lingua*, 134:219–248.

Hans Stadthagen-González, Luis López, M Carmen Parafita Couto, and C Alejandro Párraga. 2018. Using two-alternative forced choice tasks and thurstones law of comparative judgments for code-switching research. *Linguistic Approaches to Bilingualism*, 8(1):67–97.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Leonid Nisonovich Vaserstein. 1969. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

A Prompts

Some prompt examples are stated as follows:

Baseline Prompt (P1)

```
Dear ChatGPT, I would like you to serve
as a linguist and assess the grammatical
 acceptability of some sentences. I want
 you to read pairs of sentences, and for
 each pair, you need to tell me which of
 the two sentences is grammatically more
 acceptable. For each pair of the
sentences, the first sentence you read
is sentence 1 and the second sentence
you read is sentence 2.

sentence1: There has been a man
considered sick.
sentence2: There has been considered a
man sick.

Which of the two sentences is
grammatically more acceptable, Sentence
1 or sentence2? Please just say "
sentence 1" or "sentence 2" without
other words.
```

Layperson Prompt (P2)

```
You are a participant in a linguistic
experiment. Your task is to choose the
more grammatically acceptable sentence
from the sentences pair.

sentence1: There has been a man
considered sick.
sentence2: There has been considered a
man sick.

Do not explain.
Respond with only one of:
sentence 1
sentence 2
```

System-User Prompt (P3)

```
System prompt is:
You are a participant in a linguistic
 experiment. Your task is to choose the
 more grammatically acceptable sentence
 from the sentences pair.

Do not explain.
Respond with only one of:
sentence 1
sentence 2

User prompt is:
sentence1: There has been a man
 considered sick.
sentence2: There has been considered a
 man sick.
```

B The definition of BLiMP 12 categories

The category definitions are adapted from Warstadt et al. (2020), who applied 12 broader categories covering linguistic phenomena in morphology, syntax, and semantics. In this study, we additionally introduce an *Other* category to capture instances that do not fall into any of the predefined categories.

Argument structure: the ability of different verbs to appear with different types of arguments. For instance, different verbs can appear with a direct object, participate in the causative alternation, or take an inanimate argument.

Binding: the structural relationship between a pronoun and its antecedent. All paradigms illustrate aspects of Chomskys (1981) Principle A. Because coindexation cannot be annotated in BLiMP, Principles B and C are not illustrated.

Control/raising: syntactic and semantic differences between various types of predicates that embed an infinitival VP. This includes control, raising, and tough-movement predicates.

Determiner-noun agreement: number agreement between demonstrative determiners (e.g., this/these) and the associated noun.

Ellipsis: the possibility of omitting expressions from a sentence. Because this is difficult to illustrate with sentences of equal length, our paradigms cover only special cases of noun phrase ellipsis that meet this constraint.

Filler-gap: dependencies arising from phrasal movement in, for example, wh-questions.

Irregular forms: irregular morphology on English past participles (e.g., broken). We are unable to evaluate models on nonexistent forms like *breaked because such forms are out of the vocabulary for some LMs.

Island effects: restrictions on syntactic environments where the gap in a filler-gap dependency may occur.

NPI licensing: restrictions on the distribution of negative polarity items like any and ever limited to, for example, the scope of negation and only.

Quantifiers: restrictions on the distribution of quantifiers. We cover two such restrictions: superlative quantifiers (e.g., at least) cannot embed under negation, and definite quantifiers and determiners cannot be subjects in existential-there constructions.

Subject-verb agreement: subjects and present tense verbs must agree in number.

Other: instances that cannot be assigned to any of the predefined BLiMP categories.

C Additional Results

Table 4 reports the correlation between human judgments and LLM performance at the item level. Compared to the correlations at the phenomenon-type level (see Table 3), the correlation coefficients at the item level are lower. This suggests that the correlation between human judgments and model performance is influenced not only by the models themselves, but also by the specific items.

Model	P1	P2	P3
GPT-3.5	0.39***	**0.46***	0.44***
GPT-4o	**0.48***	0.46***	0.44***
GPT-OSS	0.56***	**0.58***	0.57***
Llama	0.52***	0.46***	**0.53***
Claude	0.53***	**0.54***	0.46***

Table 4: Pearson correlations between each LLM and Human across prompting strategies on the item level

* $p < .05$, ** $p < .01$, *** $p < .001$

D Stimuli Examples List

The following table counts the number of items within each linguistic phenomenon. These items are sampled from *Linguistic Inquiry* by Sprouse et al. (2013), they additional generated 7 variants for each item. The variants are not included in this table.

Phenomenon Type	N	Grammatical Sentence	Ungrammatical Sentence
Arg. structure	28	*There has been a man considered sick.*	*There has been considered a man sick.*
Binding	7	*Last night there was an attempt to shoot me.*	*Last night there was an attempt to shoot oneself.*
Control/raising	12	*John believes Mary to be sick.*	*John believes to be sick.*
Det.-noun agr.	2	*This is a table.*	*This is table.*
Ellipsis	16	*Sandy plays the guitar better than Betsy does.*	*Sandy plays the guitar better than Betsy the harmonica.*
Filler-gap	12	*Sherry met a man who she found herself very fond of.*	*Sherry met a man very fond of whom she found herself.*
Irregular forms	1	*I would have been elected.*	*Me would have been elected.*
Island effects	11	*You wonder which picture of Marge is on sale.*	*Who do you wonder which picture of is on sale?*
NPI licensing	6	*Someone better leave town.*	*Anyone better leave town.*
Quantifiers	3	*John didn't give Mary a red cent.*	*John didn't give every charity a red cent.*
Subject-verb agr.	1	*Some frogs and a fish are in the pond.*	*Some frogs and a fish is in the pond.*
Other	49		—

Table 5: Linguistic phenomena with grammatical and ungrammatical example sentences. Phenomenon Types are adopted from Warstadt et al. (2020). The grammatical and ungrammatical sentences show the examples from Sprouse et al. (2013). N indicates the number of items per phenomenon type.