Predicting Sentence Acceptability Judgments in Multimodal Contexts

Hyewon Jang[1], Nikolai Ilinykh[1]
Sharid Loáiciga[1], Jey Han Lau[2], Shalom Lappin[1,3,4]
[1]Dept. of Philosophy, Linguistics and Theory of Science, University of Gothenburg,
[2]The University of Melbourne, [3]Queen Mary University of London, [4]King's College London
{hyewon.jang, shalom.lappin}@gu.se

Abstract

Previous work has examined the capacity of deep neural networks (DNNs), particularly transformers, to predict human sentence acceptability judgments, both independently of context, and in document contexts. We consider the effect of prior exposure to visual images (i.e., visual context) on these judgments for humans and large language models (LLMs). Our results suggest that, in contrast to textual context, visual images appear to have little if any impact on human acceptability ratings. However, LLMs display the compression effect seen in previous work on human judgments in document contexts. Different sorts of LLMs are able to predict human acceptability judgments to a high degree of accuracy, but in general, their performance is slightly better when visual contexts are removed. Moreover, the distribution of LLM judgments varies among models, with Qwen resembling human patterns, and others diverging from them. LLM-generated predictions on sentence acceptability are highly correlated with their normalised log probabilities in general. However, the correlations decrease when visual contexts are present, suggesting that a higher gap exists between the internal representations of LLMs and their generated predictions in the presence of visual contexts. Our experimental work suggests interesting points of similarity and of difference between human and LLM processing of sentences in multimodal contexts.

Keywords: sentence acceptability prediction, LLMs, visual context, multimodal effects on sentence acceptability

1. Introduction

There has been a considerable amount of work on using DNNs to predict human sentence acceptability judgments (Lau et al., 2017; Bernardy et al., 2018; Warstadt et al., 2019; Lau et al., 2020; Qiu et al., 2024). Some previous studies (Bernardy et al., 2018; Lau et al., 2020) consider the effect of document (textual) context on human acceptability judgments. We examined the impact of visual contexts on both human and LLM sentence ratings.[1] The primary question we ask in this work is *whether the presence and relevance of visual contexts affect sentence acceptability judgments in humans and LLMs*. This question is motivated from previous work that reports a compression effect for human ratings in document contexts, relative to null contexts (Bernardy et al., 2018; Lau et al., 2020). This involves raising acceptability at the lower end of the acceptability scale, and lowering them at the higher end. Two explanations have been suggested for this effect, one being due to cognitive load, and the other being due to discourse coherence effect. In the cognitive load account, acceptability ratings gather around the middle because humans experience cognitive load when faced with more input to process, making them more conservative in their judgment. In the discourse coherence account, the ratings for bad sentences are raised because they look less bad when placed in the relevant discourse.

In our experiments using visual contexts instead of document contexts, we observe no such effect for human acceptability judgments, although there is some indication of a slight raising effect at the lower end, in relevant visual image contexts. We discuss the implications of this finding in §3.

We selected recent text from the Internet in a variety of genres to prevent data contamination. We follow Lau et al. (2020) in performing round trip machine translation on sentences, through several languages, using an older statistical MT system, *Moses*, to introduce a variety of syntactic, semantic, and lexical infelicities into the English output.

We employed Prolific[2] crowd sourcing to obtain native speaker human ratings on these sentences, together with a subset of the original English sources, in the context of preceding visual images. GPT-5 generated images for the original English versions of the sentences in our test data. We looked at contexts where the images were relevant to the content of the sentences, contexts where they were irrelevant, and null contexts where no visual images appear. We then tested closed LLMs

[1]Code and data are available at: https://github.com/GU-CLASP/multimodal-sentence-acceptability.

[2]https://www.prolific.com/

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 25–34
16 May 2026.

and open-source LLMs to predict the human ratings in these contexts. Four of the seven LLMs that we tested score well on predicting human sentence ratings (0.8-0.9 for Spearman correlations). However, in contrast to humans, two out of four of these models exhibit a clear compression effect for the visual contexts, which closely resembles the one reported in Lau et al. (2020) for human judgments in document contexts (see the three regression graphs in Figure 3). Overall, our models' predictions are much higher in the absence of visual contexts. We also find varying patterns in the distribution of LLM judgments, with only one of them, Qwen2.5-7B, resembling the human rating clusters to a degree.

Our experiments suggest that while current LLMs have greatly improved in their ability to identify levels of sentence acceptability relative to earlier DNNs, including first generation transformers, they process sentences in multimodal contexts differently than humans do.

2. Related Work

Prior studies employ different settings for sentence acceptability judgment. Lau et al. (2017); Warstadt et al. (2019); Qiu et al. (2024) use DNNs to predict human sentence acceptability ratings independently of context. Lau et al. (2020) and Bernardy et al. (2018) address the effect of document contexts on sentence acceptability judgment by both humans and DNNs. They employ various probabilistic methods to approximate acceptability judged by DNNs, such as applying normalising score functions to raw probabilities to filter out the effects of lexical frequency and sentence length.

Different approaches have been employed in previous work for collecting sentence data. Lau et al. (2020) and Bernardy et al. (2018) crawl natural sentences from Wikipedia and introduce infelicities by round-trip machine translation. Warstadt et al. (2019) construct a test corpus (CoLA) consisting largely of linguists' example sentences in minimal pairs, with ratings given by linguists.

We follow Lau et al. (2020) and Bernardy et al. (2018) in using naturally occurring text (modulated through round trip MT), crowd sourced non-expert human rating, and assessment relative to context. We focus on the impact of visual image contexts on both human and LLM acceptability ratings. We also consider the similarities and differences in rating distributions between humans and models.

More recent studies have been reported that are relevant to our approach in this work. Qiu et al. (2024) apply GPT-3 to Lau et al. (2017)'s human rated sentences. They show that it performs well in predicting these ratings. But they do not address the possibility of contamination in which GPT-3 may have been trained on some of this data, which has been publicly available since 2017. To minimise the influence from data contamination we include sentences published online after the training dates of the models for acceptability prediction. Furthermore, Qiu et al. (2024) only report results based on prompting experiments. Our work provides more robust findings following Kauf et al. (2024), who compare prompting and logprobs for human judgments of semantic plausibility. They report that logprobs are a good predictor of these judgments.

3. Human Acceptability Judgments

We collected 75 English sentences between the length of 25 and 40 words, from news (the Guardian, CNN, Washington Post, Wallstreet Journal, BBC), books (Google Books) and Wikipedia. To minimize bias from data contamination, we selected sentences from 2025, with the exception of Wikipedia, for which it is hard to extract new sentences only. We subjected these original English sentences to a round-trip translation using publicly available *Moses* models (Koehn et al., 2007),[3] to introduce lexical, syntactic, and semantic infelicities to the sentences, following Lau et al. (2020). We used *Moses* because most decoder only LLM-based MT systems generate fluent well-formed text. We employed three non-English languages as the pivot language (i.e., en→cs→en, en→fr→en, en→de→en) to create 75 × 3 round-trip translated sentences (see Table 1 for examples). We generated images for 75 original sentences, using GPT-5 with the prompt "*Generate an image that describes or is relevant to this sentence*". We inspected the generated images and observed that the quality was high in terms of their relevance to the original sentence (see Table 2).[4] The total of 300 sentences (75 original, 225 modified) were split into multiple batches, so that each human participant would provide sentence acceptability ratings for 20 sentences (5 original, 15 modified) on a scale from 1 (very unnatural) to 4 (very natural).[5] The participants also indicated how concrete/abstract they found each sentence on a scale of 1 (very concrete) to 4 (very abstract). The sentences were presented in three

[3] https://www.statmt.org/moses-release/RELEASE-3.0/

[4] We acknowledge that the AI-generated images may contain hallucinations and imperfections. However, we opted for this approach — as opposed to using existing image captioning dataset with natural images — to mitigate the risk of data contamination. Additionally, generating the images allowed us to ensure tight congruence between the textual and visual modalities, so that the text is directly relevant to the image and vice versa.

[5] Following Lau et al. (2017), we asked participants to evaluate *naturalness* rather than *acceptability*, a non-expert-friendly term.

Book	Orig	One piece of felt, folded in the corner, will be spliced, unfurled and dangled from the ceiling like a canopy.
	Mod	A piece in a corner, spliced were put forward and believe that the upper ceiling as a canopy.
News	Orig	But the answer seems to be no: Reeves lets it be known she requests no costings on raising the three forbidden taxes.
	Mod	But the reply does not seem: reeves suggests that it would not cost to raise the three banned taxes.
Wiki	Orig	In response, the US deployed an additional 170,000 troops during the 2007 troop surge, which helped stabilize parts of the country.
	Mod	In response, we deployed with further 170 000 soldiers in 2007 increase troops, which help to stabilize parts of the country.

Table 1: Examples of sentences used in our experiment, sourced from books, news, and Wikipedia in original form (Orig) and modified (Mod) through round-trip machine translation with Moses 3.0.

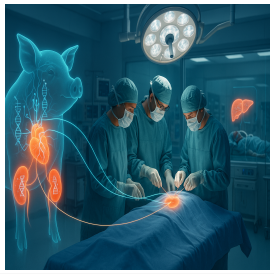
Original Sentence	**Relevant Image**	**Irrelevant Image**
One piece of felt, folded in the corner, will be spliced, unfurled and dangled from the ceiling like a canopy.		

Table 2: An example sentence with relevant and irrelevant image as context.

different conditions (*null*, *relevant*, *irrelevant*). In the *null* condition, participants rated each sentence in terms of naturalness without any preceding visual context. In the *relevant* and *irrelevant* condition, participants rated each sentence after having seen a relevant or irrelevant visual context, respectively. The irrelevant images were paired with the sentences by a simplified permutation design (e.g., s_1-im_2, s_2-im_3, s_3-im_1). To make sure that participants pay attention to the images, we asked them to select the most foregrounded object in the image from multiple choices, before rating the sentence. We batched the sentences such that no participant would see duplicate sentences or images. Each participant only rated sentences in one of three conditions. We added two attention check questions, where a prompt to choose a specific answer was embedded in a usual sentence rating task. Every sentence was rated in all three conditions (e.g., s_1-*null*, s_1-*relevant*, s_1-*irrelevant*). We recruited 20-25 native English-speaking participants for each condition and batch (gender-balanced). For quality control, we discarded data from participants who a) failed the attention check questions, b) rated original sentences as unnatural more than 40% of the time, or c) selected incorrect answers to the image question more than 25% of the time. As a result of such filtering criteria, 10% of the total participants were removed from the collected data. Participants were 45 years old on average ($\pm$ 13), with 71% British, 15% American, and 14% English speakers from other countries. The participants were paid GBP 9/hour.

3.1. Results

	Original			**Modified**		
	N	R	I	N	R	I
mean	3.54	3.54	3.53	1.96	2.02	1.94
sd	0.76	0.75	0.77	1.05	1.03	1.01

Table 3: Descriptive statistics for human ratings on sentence acceptability (1 - 4) in *null* (N), *relevant* (R), and *irrelevant* (I) conditions.

Table 3 shows the average acceptability ratings for original and modified sentences. The average ratings for original sentences are higher than for modified sentences, showing that modified sentences are infelicitous, as expected. We calculated

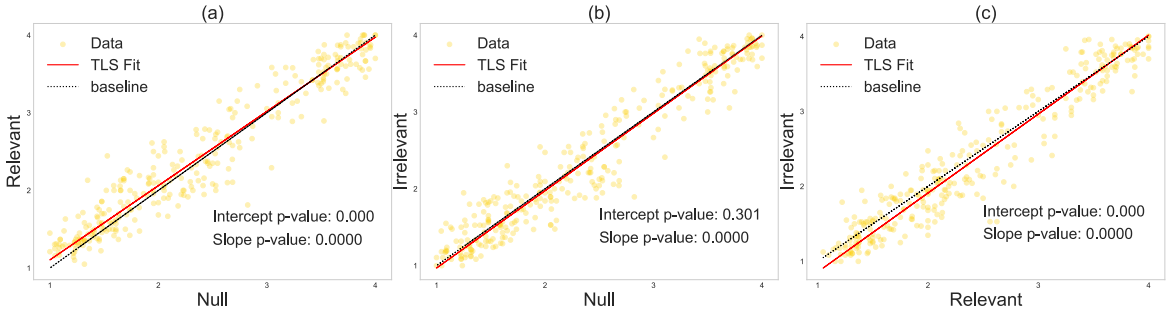


Figure 1: Human ratings of sentence acceptability in different conditions with regression lines generated from total least square regression. P-value alpha: 0.017 (Bonferroni correction).

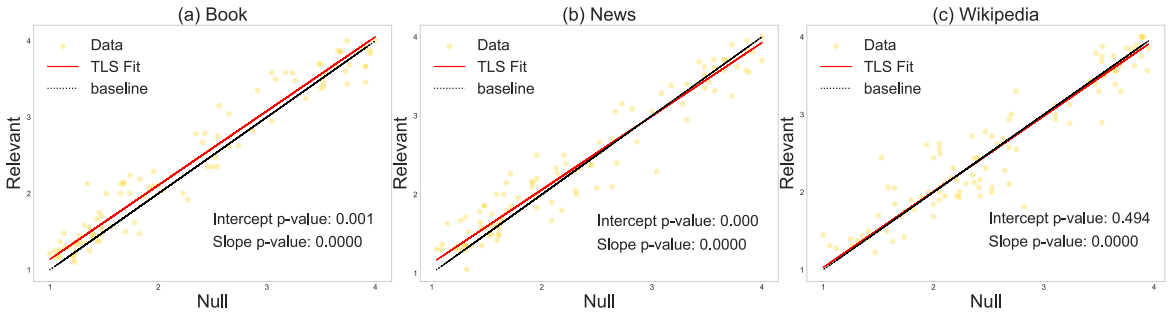


Figure 2: Human ratings of acceptability of sentences from different genres (books, news, Wikipedia) for *null-relevant* condition pair.

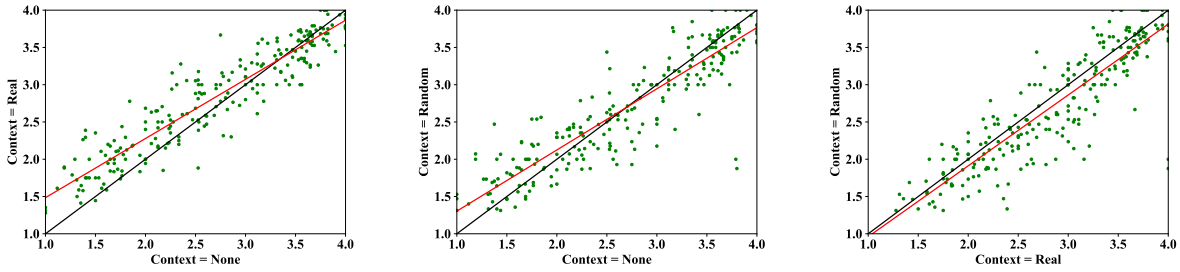


Figure 3: Human ratings of sentence acceptability in different **textual** context. "Context = None" means null context, "Context = Real" means relevant context and "Context = Random" means irrelevant context. Image reproduced from Lau et al. (2020).

mean sentence acceptability scores for each sentence across multiple human raters in each of the three conditions. We ran total least square regressions (TLR) for these average scores, with a pair of conditions at a time. Since there are no dependent variables serving as the ground truth between the condition pairs, TLR is a suitable regression analysis for this task. It accounts for errors in both X and Y variables, by minimising orthogonal distances rather than vertical ones when fitting a regression line. Figure 1 shows the average human ratings for each sentence in different conditions. Each data point (yellow dot) represents the average acceptability rating for a sentence. The red line in each pane represents the total least square regression fit for the data points in two out of three conditions at a time (*null-relevant*, *null-irrelevant*, *relevant-irrelevant*). In (a), relative to the *null* condition, we observe a slight raising effect in the *relevant* condition, where sentence acceptability scores in the lower end are higher. This indicates that human raters tend to judge a bad sentence slightly less bad when presented with relevant visual context prior to the judgment. We observe no such effect for *irrelevant* contexts (b), with the ratings for a sentence being almost identical in both conditions. The comparison between relevant and irrelevant in (c) confirms this, as the ratings for unnatural sentences are higher in relevant conditions, which mirrors (a).

Our results generally do not exhibit the compres-

sion effect for humans, reported in Lau et al. (2020) (Figure 3), where sentence acceptability ratings on both ends were compressed, causing bad sentences to be judged less bad, and good sentences to be judged less good, when a textual context precedes the target sentence.[6] In our results, we observe a raising effect only for the lower end (bad sentences) when relevant visual contexts are presented (see Figure 1(a)). This would seem to indicate that a discourse coherence effect is operative in visual contexts for human processing, as it works only to (slightly) raise lower end judgments when the preceding image renders the text more accessible. On the discourse coherence account, infelicitous sentences look more natural when placed within a context that makes them appear more coherent (see §1 for discussion regarding the cognitive load and discourse coherence accounts of compression).

Genre is also relevant. One possibility is that the characteristics of the sentences we tested, represented by the genre (news, books, Wikipedia) produces different effects for preceding contexts on human sentence processing. We looked at the effect of genre of sentences. Figure 2 shows the same information about human sentence acceptability judgment, on sentences from books, news, and Wikipedia, respectively. For sentences sourced from books, we observe a more uniform raising effect, while for sentences from news we observe a compression effect, though the magnitude is much smaller than the one reported in Lau et al. (2020). Visual contexts for sentences of distinct genres may affect acceptability judgmens differently. Sentences from books are rated as more abstract than sentences from news or wikipedia (see Table 4). Sentences from books are rated as more natural when relevant visual contexts are presented prior to the sentences, because images may render the content of ill-formed sentences more accessible. Sentences that are more concrete show a pattern comparable to the compression effect, which was manifest in textual contexts, but this effect is minimal. We do not observe it in irrelevant visual contexts, unlike Lau et al. (2020), where it is apparent in both non-null textual environments.

Another possible factor conditioning the absence of the compression effect involves cognitive load. According to this account, interpreting textual context puts additional informational processing demands on humans, forcing the acceptability ratings to be more conservative, clustering towards the middle range. This mechanism does not appear to operate in the case of visual images in conjunction with sentences. It is easier for humans to ignore them if they find no connection between an image and a sentence. Ignoring *textual* context, relevant or irrelevant, is presumably more difficult, as it is interpreted through the same processing mode. This may explain why humans exhibit a clear compression effect only with textual contexts. As a general note, the characteristics of visual contexts and textual contexts may differ in informational role. A visual context can be a depiction of a sentence's content, while a textual context is a preceding set of sentences that provides a prologue to the target sentence.

[6] Lappin (2021) describes a total least square regression test which shows that the compression effect that Lau et al. (2020) report is actual, and it cannot be reduced to regression to the norm.

	Books	News	Wikipedia
mean	1.95	1.63	1.70
sd	1.03	0.86	0.92

Table 4: Mean abstractness scores reported by human participants on original English sentences from books, news, and Wikipedia.

The observations made in this experiment need further reinforcement in order to reach a clear attribution for the presence and absence of compression effect. We leave this to future work, where we will directly compare the influence of a document context and visual context for a given sentence.

4. LLM Acceptability Judgments

We conducted two types of analyses with LLMs. First, we prompted LLMs (§4.1) to provide sentence acceptability judgments in the identical setup as the human experiment in §3. We then looked at the internal representations of the open-source LLMs by calculating normalised logprob values (§4.2), based on token probabilities, following prior work (Lau et al., 2020; Kauf et al., 2024). We used five open-source models that can take image and text as input (vision and language models) – InternVL3-1B, InternVL3-8B (Chen et al., 2024), Qwen2.5-3B, Qwen2.5-7B (Team, 2025), and llava-1.5-7b (Liu et al., 2024) – and two closed models – gpt-4o & gpt-4o-mini (OpenAI et al., 2024). InternVL3-1B, InternVL3-8B, have 1B and 8B parameters, respectively. Qwen2.5-3B, Qwen2.5-7B have 3B and 7B, respectively, and llava-1.5-7b has 7B parameters. We also made sure to select models released in early 2025 at the latest, to minimize the possibility of data contamination, although it should be noted that the chance of contamination is high for sentences from Wikipedia. We did our probability-based experiment (§4.2) on open-source models, because closed models do not allow for the extraction of logprob values for their output. Our experiments with open-source LLMs were conducted on 1 NVIDIA Tesla A40 GPU with 48GB RAM. The

experiments lasted about 30 hours in total.

4.1. Prompting-based analysis

We prompted the LLMs in a zero-shot setting to judge the naturalness of each of the 300 sentences used in §3 (75 original sentences + 75 × 3 modified sentences), under three different conditions. In the *null* condition, we presented only the sentences to the LLMs. In the *relevant* and *irrelevant* conditions, we first prompted the LLMs to see an image and solve the same task given to the human participants in §3. We then prompted them to judge the naturalness of the subsequent sentence after feeding their response to the next generation call (i.e., in the second call the input concatenates the image attention check task and response, and the sentence acceptability task instruction). We did not provide any explicit instructions to connect the image with the following sentence. We used 10 initialization seeds for each model, and we averaged the sentence acceptability ratings across the seeds. We used hyper-parameter settings of temperature=0.7, top-p=1.0, top-k=50 for all open-source models, and the default settings for closed models.[7]

		All	N	R	I
LLM ratings ∼ Human ratings	gpt-4o	0.87	**0.89**	0.87	0.88
	gpt-4o-mini	0.89	**0.89**	**0.89**	0.88
	InternVL3-1B	0.32	**0.66**	0.26	0.16
	InternVL3-8B	0.83	**0.88**	0.85	0.85
	Qwen2.5-3B	0.65	**0.70**	0.61	0.66
	Qwen2.5-7B	0.78	**0.84**	0.75	0.76
	llava-1.5-7b	0.25	**0.39**	0.21	0.15

Table 5: Spearman (ρ) correlations between average human sentence acceptability ratings (1-4) and LLM-prompted ratings (1-4) in null (N), relevant (R) and irrelevant (I) conditions. All correlations significant ($p<0.001$). Highest correlations per row marked in bold.

4.2. Probability-based analysis

For this analysis we used the sentence probability as estimated by the LLMs directly (Ide et al., 2025; Hu et al., 2024; Kauf et al., 2024). We passed the images and sentences through the open-source models (forward pass without generation) and extracted logits for only the input sentence in all three conditions (*null*, *relevant*, *irrelevant*). We calculated the average log probabilities of a sentence, normalised by the number of tokens (MeanLP), by summing the log-probabilities of each token in a sentence, conditioned on the preceding tokens (Lau et al., 2020; Kauf et al., 2024). We considered the MeanLP as a proxy for the LLM's internal representation of its sentence acceptability judgment, following prior work. We experimented with normalising functions employed in Lau et al. (2020), and found them to be unnecessary based on the high correlations of most LLMs with the human ratings.

		All	N	R	I
MeanLP ∼ Human ratings	InternVL-1B	0.70	**0.72**	0.70	0.70
	InternVL-8B	0.70	**0.72**	0.71	0.70
	Qwen-3B	0.61	**0.70**	0.61	0.57
	Qwen-7B	0.73	**0.79**	0.76	0.71
	llava-7B	0.74	0.74	**0.76**	0.73

Table 6: Spearman (ρ) correlations between Mean Logprobs from open-source LLMs and average human sentence acceptability ratings. All correlations significant ($p<0.001$). Highest correlations per row marked in bold.

4.3. Results

Table 5 shows correlations (Spearman ρ) between average human sentence acceptability ratings and LLM-generated ratings (§4.1).[8] The two GPT models show overall correlations of over 0.87, with little to no variations across conditions. Of the open-source LLMs, models with higher parameter size (7B-8B) perform almost equally well, with the exception of llava-1.5-7b. The poor performance of the smaller models (InternVL3-1B, Qwen2.5-3B) and llava-1.5-7b might be attributed to their reduced capability to follow instructions (llava-1.5-7b in particular is not trained to follow instructions). The open-source LLMs also seem to perform better without any visual context (N).

Table 6 presents correlations between average human sentence acceptability ratings and probability measures from the LLMs described in §4.2. Compared to the prompting results (Table 5), MeanLP generally performs on par or better for the open-sourced LLMs. Most interestingly, the performance of these models *across conditions* (N, R, I) also appears to be much more consistent — a contrast compared to the prompting results in the case of llava-1.5-7b.

To better understand the similarity between MeanLP versus prompted model ratings, we present their correlation in Table 7. The larger models (InternVL3-8B and Qwen2.5-7B) have better correlations (both 0.70 for *All*), while the others have much lower correlations. Either way, these results show that these two approaches produce

[7] We ran the experiments with varying temperatures posthoc (0.5 to 1.0 in the increments of 0.05), and the results maintained the same patterns.

[8] We only report Spearman ρ, as Pearson r yielded very similar results. The same applies to Tables 6 and 7.

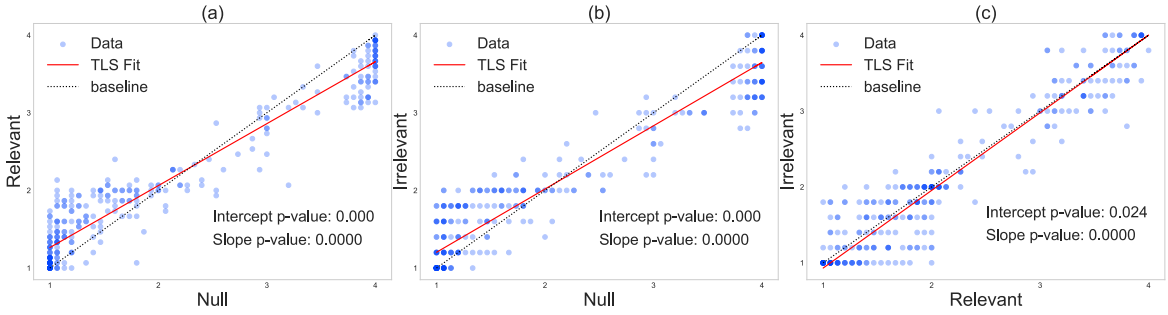


Figure 4: Sentence acceptability ratings generated by gpt-4o.

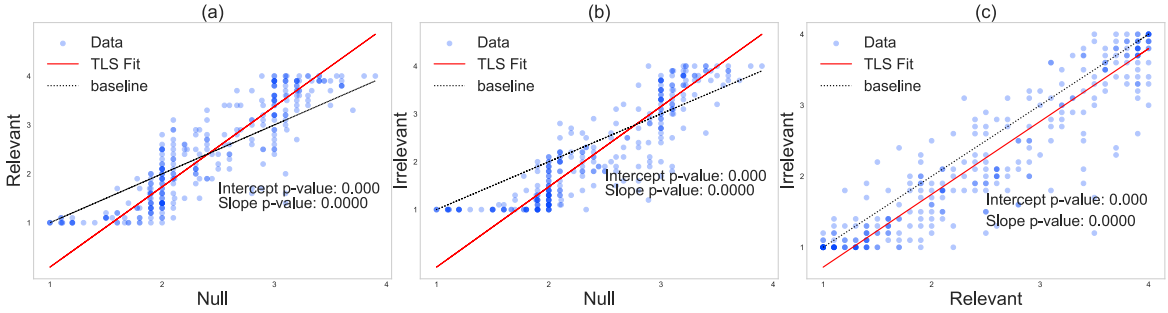


Figure 5: Sentence acceptability ratings generated by InternVL3-8B.

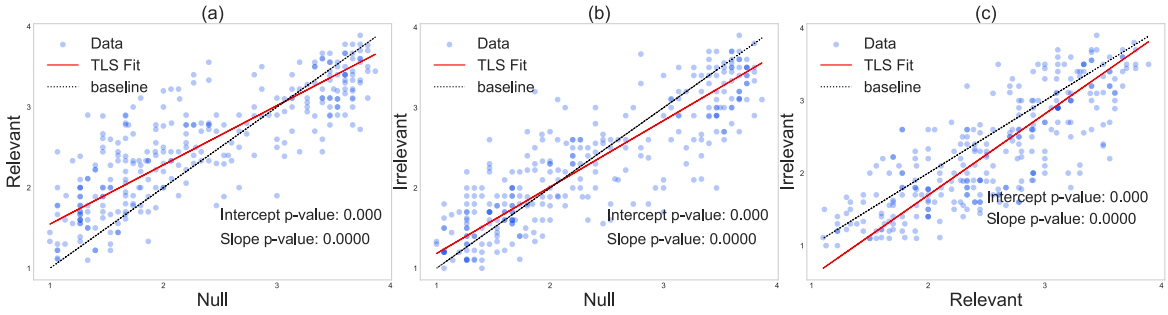


Figure 6: Sentence acceptability ratings generated by Qwen2.5-7B.

somewhat distinct predictions of acceptability ratings. But, Tables 5, 6, and 7 all suggest that correlations are higher in the *null* condition compared to the non-*null* conditions (with the exception of one case involving llava-1.5-7b).

		All	N	R	I
MeanLP ∼ Model ratings	InternVL-1B	0.30	**0.58**	0.30	0.17
	InternVL-8B	0.70	**0.75**	0.74	0.71
	Qwen-3B	0.52	**0.56**	0.48	0.45
	Qwen-7B	0.70	**0.76**	0.71	0.68
	llava-7B	0.28	**0.35**	0.24	0.22

Table 7: Spearman (ρ) correlations between Mean Logprobs from open-source LLMs and model prompted ratings. All correlations significant ($p<0.001$). Highest correlations per row marked in bold.

Based on the correlation analyses, we consider Qwen2.5-7B, InternVL3-8B, and gpt-4o to have yielded the most reliable results. In Figures 4, 5, and 6 we present the scatterplots of the sentence acceptability ratings produced by these models in each condition pair (*null-relevant*, *null-irrelevant*, *relevant-irrelevant*) for more detail, parallel to the human results in §3. Our intention is to examine these scatterplots to understand the qualitative nature of the model rating distribution, and also to make sure the correlation numbers are not skewed by a few outliers. In general, we see that the distributions of sentence acceptability ratings produced by the LLMs appear quite different from those of humans. gpt-4o -generated ratings cluster around top and bottom ends, and InternVL3-8B -generated ratings cluster around 2 and 3. Qwen2.5-7B shows

distributions that are most similar to humans,[9] but with a much higher variance than humans. Interestingly, we observe the compression effect for both gpt-4o and Qwen2.5-7B. Lau et al. (2020) show a very similar compression effect for human sentence ratings in textual contexts, as we see in Figure 3. But we observe the opposite of compression for InternVL3-8B. We do not have an explanation for this contrast. We will be exploring it in future work. The general divergence in the distributions of model-generated ratings from human-generated ratings may be due to the different mechanisms by which LLMs process sentences in comparison to humans. LLMs have less memory interference, and they retain perfect access to previous words (Oh and Linzen, 2025).

5. Discussion

The work that we present here indicates that LLMs achieve a high degree of accuracy in predicting human sentence acceptability judgments. It also shows that the normalised logprob values that these models assign to sentences are a reliable predictor of human ratings for sentence acceptability. Despite the strong convergence of LLM logprob scoring and human naturalness rating, most of the LLMs that we consider in this work display different data clustering patterns than humans in the distributions of these judgments (cf. Figures 1 and 5 for two illustrative examples). In particular, gpt-4o exhibits a more polarised pattern, while Qwen2.5-7B approaches human distributions (Figure 4 vs 6). However, although Qwen2.5-7B most closely approximates the human distributions, it still differs from them, both in our experiments and in those reported by Lau et al. (2020) (cf. Figures 1, 3, and 6). Due to their greater memory capacity compared to humans, Oh and Linzen (2025) argue that LLMs' superhuman next-token prediction ability makes them unsuitable as cognitive models of human linguistic prediction. However, the tasks differ. Oh and Linzen (2025) focus on linguistic continuation tasks which effectively involve next-token prediction. In contrast, our experiments examine acceptability judgments, where LLMs predict evaluation scores and not upcoming tokens. It therefore remains an open question whether the differing patterns we observe reflect the LLMs' superhuman predictive capacity or instead stem from fundamentally different processing strategies.

Our experiments, viewed from the perspective of the work reported in Bernardy et al. (2018); Lau et al. (2020), suggest an interesting difference in the way that humans and LLMs process information in different modalities. Textual context, whether relevant or not to the following sentence, influences the way in which humans assess the naturalness of sentences, both well-formed ones, and those containing infelicities. Interpreting these contexts requires processing resources that influence judgments concerning the naturalness of following sentences.

By contrast, for humans, interpreting images seems to proceed through alternative processing mechanisms, which permit the image to be discarded, or suppressed, when rating the naturalness of a following sentence. This effect is particularly clear when the image has no clear relation to the sentence. However, LLMs, at least the vision and language models we used in this work, incorporate images into their sentence processing environment. They include the images when assigning probability values to sentences, and in rating them for naturalness. They do not discard them, even when they are irrelevant. Cognitive load appears to affect their judgments across visual and textual modalities. This is not surprising since their architectures are generally designed to maximise contexts (Oh and Linzen, 2025). The vectors that encode these contexts include elements of all modalities that the model attends to.[10]

6. Future Work

In future work we will examine the role of different types of text genre more closely in determining the impact of both textual and visual contexts on human sentence acceptability rating. We will also explore the mechanisms through which humans suppress images, but not preceding text, when processing sentences. Our objective here is to obtain a clear sense of the way in which cognitive load is conditioned by different sorts of context, relative to distinct text genres. For this, we will also consider other linguistic tasks than sentence acceptability judgment and experiments with larger data for better generalizations.

Another issue that we will explore is how LLMs can be modified to converge on human processing with respect to the suppression of images in the assignment of logprob values to sentences, without losing the content of these images. This would cause the modified model to more closely approximate observed human sentence processing in multimodal environments.

[9] This might be an artifact of Qwen2.5-7B producing more varied ratings with different initialisation seeds than the other LLMs.

[10] Our experimental results are broadly consistent with the view of the relation between LLMs and human processing presented in Oh and Linzen (2025). Simlarly, our suggestions for future work align well with their approach.

Acknowledgements

The work reported in this paper was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. The computation and data storage for our experiments were supported by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

Limitations

The work reported in this paper only addresses English. We tried to minimise data contamination issues by experimenting with sentences from 2025, but there is a high chance of data contamination with sentences from Wikipedia, which are not as new. We reported that genre effects influence sentence acceptability judgment by humans, but the explored genres in this work are limited to three (news, books, wikipedia). We did not directly compare visual contexts and document contexts for the same target sentence, which may render the results inconclusive. Lastly, we did not fully address the characteristic differences of visual versus textual contexts in this work. We will follow up on these limitations in our future work.

Ethics statement

All our human participants took part in the experiments voluntarily, and received due compensations. We recognized that we might expose crowdworkers to disturbing/offensive images and texts. To minimize this risk, we manually examined the 75 sentences before administering the human experiments. GPT-5 also has its own guardrails for image generation to keep such risks low (https://cdn.openai.com/gpt-5-system-card.pdf). Some of our modeling experiments were done with closed models, which may render the results not entirely transparent. We did not rely on any AI-assistant tools for the generation of our article text.

Jean-Philippe Bernardy, Shalom Lappin, and Jey Han Lau. 2018. The influence of context on sentence acceptability judgements. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 456–461, Melbourne, Australia.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

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, Albuquerque, New Mexico. Association for Computational Linguistics.

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A Ivanova. 2024. Log probabilities are a reliable estimate of semantic plausibility in base and instruction-tuned language models. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 263–277, Miami, Florida, US. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Shalom Lappin. 2021. *Deep Learning and Linguistic Representation*. CRC Press, Taylor & Francis, Boca Raton, London, New York.

Jey Han Lau, Carlos Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. 2020. How furiously can colorless green ideas sleep? sentence acceptability in context. *Transactions of the Association for Computational Linguistics*, 8:296–310.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive Science*, 41:1202–1241.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Byung-Doh Oh and Tal Linzen. 2025. To model human linguistic prediction, make LLMs less superhuman. *arXiv preprint arXiv:2510.05141*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Zhuang Qiu, Xufeng Duan, and Zhenguang Cai. 2024. Evaluating grammatical well-formedness in large language models: A comparative study with human judgments. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 189–198, Bangkok, Thailand. Association for Computational Linguistics.

Qwen Team. 2025. Qwen2.5-vl.

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