# Language models demonstrate the good-enough processing seen in humans

**Yan Cong (cong4@purdue.edu)**
School of Languages and Cultures, Purdue University

**Julia Rayz (jtaylor1@purdue.edu)**
Department of Computer and Information Technology, Purdue University

## Abstract

Comparative illusions, also called Escher sentences, are comparative sentences that appear acceptable but challenge the boundaries of comprehension. Escher sentences, illusions that are the subject of this paper, require structural reinterpretation to resolve their anomalies, making them ideal phenomena for probing the mechanisms of language processing in humans and machines. Using human behavior as a benchmark for large language models' (LLM) performance, we assessed LLMs' behavior with three methods: prompt, probability measurement, and lexical disturbance. Our results indicate that LLMs, when prompted, display "human-like" behavior, LLMs struggled to reliably rank surprisals, and they were sensitive to both lexical and syntactic cues in Eschers. These results indicate that LLMs seem to manifest the good-enough processing seen in human cognition, constructing quick, shallow representations for efficiency. Although the rational inference processing is also seen in humans, we did not find solid evidence for LLMs, indicating differences from human cognition. Understanding and narrowing this discrepancy could result in AI systems that are more in tune with human reasoning and more interpretable, ultimately enhancing our grasp of sentence processing in humans and machines.

**Keywords:** illusion; comprehension; LLMs interpretability

## Introduction

### Escher sentences: what is being compared?

Comparative illusions, for example, *More people have been to Russia than I have [been to Russia]*, are also called **Escher sentences**. Eschers can be defined as a language illusion that people accept at first glance but struggle to pinpoint its precise meaning (Wellwood, Pancheva, Hacquard, & Phillips, 2018; Kelley, 2018; Zhang, Gibson, & Davis, 2023; Zhang, 2024). In the example sentence it is unclear what is being compared—-people (individual reading) versus the speaker's actions or visits (events reading)—which creates an illusion of meaningfulness but breaks down upon closer analysis. Initially, it seems as though the sentence compares how many times people have visited Russia with how many times the speaker has. However, when the ellipsis is resolved as *been to Russia*, the sentence becomes nonsensical: it implies a comparison between the number of times people have been to Russia (for example, three) and whether the speaker has been to Russia at all (True/False) (Wellwood et al., 2018; Kelley, 2018; Zhang et al., 2023). Our scripts can be found in GitHub `https://github.com/yancong222/EschersLLMs`.

Linguists have studied such illusions to understand how the human brain processes language, and why our minds do not immediately reject them, unlike most unacceptable sentences, namely where they fail to detect the unacceptability (Ferreira & Patson, 2007; Wellwood et al., 2018; Zhang et al., 2023; Zhang, 2024; Ke, Tong, Chen, & Peng, 2024). Research has shown that several factors influence the strength of the illusion, including the plurality of determiner phrases (DPs), the repeatability of the event described by the verb phrase, the subject form in the *than*-clause, the number of that subject, and so on (Wellwood et al., 2018; Kelley, 2018; Paape, 2024). In this study, we use a multi-method approach that complements prior efforts such as in Zhang (2024), focusing on the factor of DPs' plurality, which leads to *Weak* Escher illusions when the DP is singular, and *Strong* illusions when the DP is plural. There is no such illusory effects in canonical acceptable comparatives, as illustrated below:

1. **Weak Escher:** More Brazilians made sandwiches than the American did.
2. **Strong Escher:** More Brazilians made sandwiches than the Americans did.
3. **Canonical comparative:** More Brazilians made sandwiches than Americans did.

Previous research showed that strong Eschers elicited greater illusory acceptability effects than weak Eschers, with controls rated as most acceptable (Kelley, 2018). Strong Eschers refer to the extent to which humans are fooled by illusions and consider a sentence acceptable. Human participants rated acceptability of strong Eschers lower than controls and weak Eschers even lower, suggesting that human participants are sensitive to Eschers with varied strength: they are more likely to be fooled by strong Eschers than weak Eschers. Kelley also found that the extent to which humans are fooled by illusions depends on task demands and sentence complexity.

### Human sentence processing of Eschers

Two major frameworks of human sentence processing are pertaining to our work. First, the *good-enough* framework suggests that human language processing prioritizes efficiency over accuracy, relying on heuristics and partial analyses (Ferreira, Bailey, & Ferraro, 2002; Ferreira & Huettig, 2023; Ferreira & Patson, 2007). Human participants construct surface-level interpretations sufficient for most contexts, even if they overlook critical details (Wellwood et al.,



In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* 

2018; Paape, 2024). In the case of Escher sentences, shallow processing allows them to pass initial plausibility checks before their structural anomalies become apparent. In contrast, the *rational inference* framework argues that comprehenders aim to reconstruct intended meanings by integrating linguistic input with prior knowledge and probabilities of potential errors (Gibson, Bergen, & Piantadosi, 2013; Levy, 2008). This framework assumes an idealized comprehender with access to language statistics and error-correction mechanisms. Applying rational inference to the example Escher sentence involves reinterpreting the input to achieve coherence. For example, comprehenders might infer that the speaker intended to compare their visits to Russia with those of others, despite the literal sentence being unacceptable. This process requires additional cognitive effort, such as rereading or mentally editing the input, to resolve under-specifications and ambiguities (Zhang et al., 2023; Paape, 2024).

The good-enough and rational frameworks represent distinct strategies in sentence processing. Good-enough processing emphasizes cognitive efficiency, often at the cost of precision, while rational inference prioritizes accuracy through resource-intensive error correction. Both mechanisms likely coexist in human cognition, with their activation depending on individual differences, context, and task demands (Brehm, Jackson, & Miller, 2021; Yadav, Paape, Smith, Dillon, & Vasishth, 2022). The coexistence of these strategies indicates the flexibility of human comprehension and the trade-offs between efficiency and accuracy (Paape, 2024).

## Aims and overall hypothesis

Built up on illusion processing studies (Paape, 2024; Kelley, 2018), we propose that humans process linguistic illusions mostly through rational inference, though they adeptly switch to good-enough processing depending on illusory strength and other factors. Empirical results about comparative illusions show a dominance of rational inference in formal tasks requiring error detection, though significant good-enough processing persists (Paape, 2024). Relatedly, behavioral and neuro-linguistic experiments demonstrated that the human identifies Escher sentences as unacceptable, indicating that there is active and supplementary reconstruction involved, a characteristic of the rational inference framework (Kelley, 2018). Human participants also displayed systematic variability and a sensitivity to diverse Escher comparisons, implying that good-enough processing remains in effect.

Our hypothesis is that if the behavior of LLMs in Escher comprehension is compatible with predictions, specifically, if it theoretically aligns with human behavior and can be interpreted through the lens of good-enough or rational inference, then LLMs will adeptly alternate between rational inference and good enough processing, and the dominant processing approach in LLMs will be rational inference. In humans, error correction involves implicit updates via a feedback loop, where cognitive reasoning errors are adjusted on the fly, particularly at the level of making sense of language. Unlike human cognition, LLMs may not adjust for errors at the cognitive reasoning level, operating solely at the surface level. Furthermore, LLMs do not face the same resource limitations as humans, as they primarily rely on probabilistic, data-driven approaches. Note that probabilistic next-token prediction does not equate to full rational reasoning over meanings. Overall, our sub-hypothesis is that LLMs' outputs will more closely align with the "rationalist" view, as they utilize probabilistic modeling in their predictions.

Our aim is to add to the research of the unique (cognitive) strategies underlying human and machine language comprehension. Further, such investigations will inform the development and optimization of computational models that more closely mirror human processing, hence more interpretable and flexible (Millière, 2024; Kelley, 2018; Cai, Duan, Haslett, Wang, & Pickering, 2024; Dentella, Günther, Murphy, Marcus, & Leivada, 2024; Paape, 2023; Zhang et al., 2023; Qiu, Duan, & Cai, 2024). We address these aims by comparing LLMs outputs with human responses to carefully designed comparative illusions items. The findings will enhance our grasp of language processing paradigms in LLMs through the lens of Escher sentences.

## Method

The LLMs we tested include GPT-2 (Radford et al., 2019; Brown et al., 2020), Llama-2-7b-hf and Llama-2-13b-hf (Touvron et al., 2023), Qwen1.5-MoE-A2.7B (Bai et al., 2023), Falcon-7b (Almazrouei et al., 2023; Penedo et al., 2023), mpt-7b (Team et al., 2023), Mistral-7B-v0.1 (Jiang et al., 2023), and GPT-4o-mini (Achiam et al., 2023). Behavioral analyses were conducted using *minicons* (Misra, 2022). Open-source models were accessed via *HuggingFace*, and GPT-4o-mini experiments were run through the OpenAI API (query time: from May 2024 to January 2025). All LLMs are auto-regressive models. The LLMs were analyzed on three experiments: prompt analysis on acceptability and interpretation, probability-based surprisal analysis on both the whole sentence and the target area, and lexical disturbance from nonce words substitution.

### Prompt: materials and design

To examine LLMs' alignment with humans and how the two major processing paradigms account for their behavior, we prompted LLMs on accepting and interpreting comparative illusions. We used previous human findings (Kelley, 2018) for the acceptability survey as a benchmark and conducted our own human experiments for the interpretation survey, which we compared with LLMs.

**Acceptability survey** Acceptability ratings on LLMs were collected on a 7-point Likert scale (1 = unacceptable, 7 = fully acceptable). The stimuli included 120 sentences divided into 60 experimental items and 60 fillers, adapted from Kelley. Experimental items refer to canonical comparatives and comparatives with strong and weak illusory effects, whereas fillers included both acceptable (good) and unaccept-

able (bad) examples to set rating benchmarks. Following Kelley's human experiments, fillers also mixed comparative and declarative sentences to disguise the focus on comparative cases, such that LLMs are prevented from noticing the experimental cases contained only in comparative sentences, hence addressing potential bias seen in human cognition. We used GPT-4o-mini with its default settings. We also crafted the following prompt to better interpret and understand the model's "reasoning": "*If an interpretation exists for the following statement, provide a concise 1-sentence interpretation; otherwise, output 'no interpretation'*". We conducted both quantitative analysis of the ratings and qualitative examination of responses. Note that LLM-generated explanations may not faithfully reflect the model's internal mechanisms. Our prompting method aims to support descriptive alignment between LLM outputs and human interpretations (likely with mixed strategies). Combined with probability-based analyses, this approach contributes to our broader proposal for understanding and evaluating LLM "reasoning", with the caveat that such reasoning is observed behavior, a proxy to and a representation of internal cognitive processes.

We hypothesized that if LLMs align with humans' responses, the two would show similar patterns in the acceptability survey: the control grammatical comparatives are the most acceptable, hence the highest acceptability scores, the strong Eschers are strongly illusory but not as acceptable as the control, hence lower scores in the rating, followed by weak Eschers, which are the least acceptable.

**Interpretation survey** We additionally conducted an interpretation survey involving both humans and LLMs, modifying the items from Kelley to include a subsequent question concerning the DPs for all three conditions: the control, strong Eschers, and weak Eschers. Among the total of 336 items, were 60 items using real stimuli and 52 synthetic control items produced by GPT-4. The synthetic controls were created using the real control items, with the second DP adjusted to form the respective weak and strong stimuli items. Each synthetic item was manually validated for quality. Over *Prolific* (Palan & Schitter, 2018), we recruited 24 adult self-identified native speakers of English currently living in the U.S. to complete a multiple-choice task exemplified below:

- Stimuli item: More Brazilians made sandwiches than ***Weak****: the American/* ***Strong****: the Americans/* ***Control****: Americans* did. Who made fewer sandwiches?
- Interpretation options: A. Americans B. The Americans C. The American D. None of the above.

Items and response options were pseudo-randomized to minimize confounding variables and experimental biases. All participants gave informed consent to participate in the study. The median duration of an experimental session was 20 minutes. Human and LLM were prompted using the same format and items for valid comparison.

We hypothesized that if the rational processing framework is dominant, humans would approach the task by leveraging explicit reasoning and conscious error correction to resolve linguistic illusions. Thus, participants engaging with the stimuli would carefully evaluate each response option, favoring interpretations that align with logical coherence and grammatical correctness. For example, when interpreting the above stimuli item, rational processors, regardless of weak or strong Escher items, would overwhelmingly choose "A. Americans" after rationally resolving the comparison structure, reconstructing (weak or strong) Escher sentences into an acceptable comparative construction, and identifying the implied meaning. Conversely, the good-enough processing framework suggests that humans often rely on shallow processing strategies that prioritize efficiency over precision. In this case, participants might bypass a thorough analysis of the syntactic and semantic nuances, instead relying on heuristics or surface-level cues. Therefore, they might choose whichever option that matches the words following *than* and preceding *did*, which they have just observed. It is also likely that good-enough processing might lead to errors or unexpected responses, based on human participants' familiarity with definite noun phrases or overlooking subtle distinctions in the stimulus structure.

The "rationalist" view of language processing envisions an idealized comprehender with almost flawless knowledge of linguistic statistics (Paape, 2024). As discussed earlier, we expected LLM's responses to align more closely with the rational inference framework, given LLM's extensive reliance on probabilistic modeling. However, LLMs are also susceptible to generating "hallucinated" responses, especially when handling ambiguous or under-specified stimuli, reflecting a form of shallow processing akin to the good-enough framework. This dynamic would likely manifest in cases where GPT-4o-mini generates responses that correspond with the sequence of words in the stimulus item, despite such choices being less plausible upon inspection.

## Surprisal: materials and design

The prompt method may not fully disclose the behavior of LLMs in a comprehensive manner (Hu & Levy, 2023). Thus, to strengthen our investigation and to pinpoint LLMs' sensitivity to anomalies at the semantic-syntactic interface, surprisal scores derived from LLMs were used as proxies for acceptability, with higher surprisals indicating lower acceptability (Michaelov & Bergen, 2022). We accumulated token-level LLM surprisals, adjusted for sentence length, for every sentence within the same collection of 120 sentences used in our first experiment. Taking Kelley's work as an interpretation baseline, we hypothesized that if LLMs show sensitivity to Escher illusions, LLMs should be able to distinguish between strong and weak Eschers consistently, in addition to teasing apart fillers, controls, and Eschers. Translated into LLMs-computed surprisals, if LLMs process Eschers in a "human-like" interpretable pattern, there should be statistical differences in the LLM-derived scores for comparisons where significant results were found in Kelley's human experiments, and crucially, the rank of such scores should resemble Kelley.

Moreover, Kelley's study collected EEG (Electroencephalography) data utilizing a rapid serial visual presentation approach, identifying the critical EEG segment at the sentence's conclusion, beginning 200 milliseconds prior to the target auxiliary verb *did*. Early detection of surprising sequence in Escher constructions was observed despite syntactic mismatches, suggesting rational inference processing. Therefore, besides calculating tokenwise surprisal for the same 120 sentences, we extracted the surprisals for the target word *did*. We hypothesized that if LLMs exhibit "human-like" variability and sensitivity to Eschers of varying strength levels (none, weak, and strong), they should display significant differences in surprisal at the target word *did* across these conditions, mirroring the human rank order of surprisal across conditions as reported in Kelley.

### Disturbance: materials and design

Lastly, to further evaluate the robustness of LLMs processing, we replaced key nouns in the stimuli with nonsense words (Misra, Rayz, & Ettinger, 2023). This approach tested LLMs' reliance on lexical semantics and their ability to process unacceptable structures in the absence of semantic grounding. Specifically, we hypothesized that if LLMs rely solely on syntax when processing Eschers, their behavior should remain consistent across original and substitution conditions; however, any differences would indicate that both lexical and syntactic cues play a role in their processing.

We used the same set of 120 sentences, utilized in the prior experiments. We conducted three types of nonce word replacements to examine the boundaries of LLMs' syntactic processing by introducing various degrees of disturbance: we replaced the noun before the critical auxiliary *did* with *vun/vuns* (s1); we replaced the noun after the first *more* with *kad/kads* (s2); we applied both substitutions (s1s2).

We tested seven nonce words in total: *kad, vun, pilk, malk, mirk, milp, filk* (Keuleers & Brysbaert, 2010; de Varda, Gatti, Marelli, & Günther, 2024). We additionally manipulated parts of speech by capitalizing words based on the original stimuli and context. As a syntactic cue, capitalization signals proper nouns. We calculated GPT-2 surprisals on the target word *did*, and we prompted GPT-4o-mini on the same items. The same pipeline used in the prior prompt and surprisal experiments was applied here.

## Results

### Acceptability survey

Descriptive statistics for the ratings were given in Table1. Similar to humans' ratings, the model assigned lower acceptability to "bad" unacceptable fillers and higher to acceptable fillers. Also, GPT-4o-mini assigned higher acceptability to control, followed by strong and weak. Different from humans' ratings, GPT-4o-mini generally showed less variation across conditions, with a more compressed scale of acceptability judgments. The model tends to "rate" all constructions with relatively high scores, even in cases where human participants showed sharper declines in acceptability.

Table 1: Descriptive statistics of prompting GPT-4o-mini in Eschers acceptability ratings, contrasted with human judgments inferred from histograms in Kelley (2018).

| **GPT-4o-mini** | **Mean (Std)** | **Range** |
|---|---|---|
| Control | 5.550 (0.605) | [4, 6] |
| Filler (Bad) | 3.050 (1.146) | [2, 6] |
| Filler (Good) | 5.425 (0.675) | [4, 7] |
| Strong | 5.500 (0.607) | [4, 6] |
| Weak | 5.100 (0.641) | [4, 6] |
| **Human Ratings** | **Mean (approx.)** | **Range** |
| Control | 6.5 | [1, 7] |
| Strong | 5.5 | [1, 7] |
| Weak | 4.5 | [1, 7] |

Qualitatively, for LLMs' 1-sentence interpretation, there appears to be an observable association between simple interpretations and higher acceptability ratings. For example, interpretations that summarize a sentence in basic terms often lead to higher ratings (e.g. 6 out of 7). This behavior suggested that GPT-4o-mini tends to "rate" sentences as acceptable if it can form a quick, surface-level processing of the sentence's meaning, even if that "impression" may overlook deeper ambiguities or illusions. This is typical of good-enough processing where initial impressions or familiar structures dominate the model's output (Paape, 2024). Conversely, when the model struggled to construct a straightforward 1-sentence interpretation, its ratings dropped, indicating that it finds those constructions more "problematic" or less natural. This is evident when dealing with sentences that require resolving more complex comparative illusions such as strong Escher sentences. The model often defaulted to generic or incomplete responses under these conditions. This behavior mirrors good-enough processing, where an LLM might fail to fully resolve complex sentences but still produce an approximate interpretation based on surface cues.

Overall, the results of prompting indicated LLMs' "human-likeness" in general patterns as well as their limited sensitivity to the syntactic intricacies that affect acceptability for humans, as evidenced by the compressed range of acceptability scores under different conditions. This suggests that LLM might prioritize surface-level patterns or statistical regularities rather than deeply engaging with the underlying structure and actively reconstructing or correcting it.

### Interpretation survey

Humans' and model's choices were summarized in Table2. Human responses indicated a tendency towards good-enough processing, as they often opt for the choice that most closely resembles the DP prior to *did*, regardless of conditions. In the control scenario where the noun phrase before *did* is plural, participants selected the plural noun phrase "NPs" like *Americans* 76.1% of the time. In contrast, "The NPs" like

*The Americans* was chosen 18.9% of the time, and singular "The NP" like *The American* was selected merely 0.07% of the time. Similar trends appeared in both weak and strong Eschers conditions.

Table 2: Comparison between human responses and those from GPT-4o-mini to an Escher-related follow up question, with 112 items for each condition. The number indicates the proportion of choices made.

| Condition | Source | NPs | The NPs | The NP |
|---|---|---|---|---|
| Control | Human | 0.761 | 0.189 | 0.007 |
| | GPT-4o-mini | 0.750 | 0.250 | 0.000 |
| Weak | Human | 0.008 | 0.019 | 0.934 |
| | GPT-4o-mini | 0.069 | 0.181 | 0.750 |
| Strong | Human | 0.040 | 0.907 | 0.015 |
| | GPT-4o-mini | 0.083 | 0.889 | 0.030 |

The model displayed similar patterns as humans, suggesting good-enough processing as a major account for LLM. Rational inference processing was generally not prominent in the task performance of both humans and the model. Interestingly, the option "None of the above" was rarely seen in humans or the model. Further, for the weak condition, humans overwhelmingly chose "The NP", whereas GPT-4o-mini displayed a slightly more distributed response pattern.

## Probability measurements

LLMs-surprisals (Fig.1) results suggested that LLMs can distinguish fillers from Eschers, but they failed to tease apart strong and weak Eschers. Controls and Eschers were also indistinguishable in LLMs. Despite the lack of consistent statistical significance, the rank of LLMs-surprisals was mostly interpretable and "human-like", with unacceptable filler sentences showing the highest surprisals, the acceptable fillers the lowest, and weak Eschers showing higher surprisals than strong Eschers.

Results of LLMs-surprisals on *did* were visualized in Figure 2. All the LLMs showed sensitivities to strong and weak Eschers. Further, Falcon, Llama2, and MPT were able to distinguish controls from strong Eschers. However, the rank of the surprisals on the target word *did* was not consistently interpretable. We found that across LLMs, weak Eschers showed the highest surprisals on target word *did*, followed by controls, and the strong Eschers showed the lowest surprisals. This is different from the rank of acceptability found in humans: control followed by strong and then weak Eschers. LLMs were so strongly "fooled" by strong Eschers that nonsensical sentences were "considered" less surprising than acceptable comparatives. This further suggested that rational reconstruction and correction is not a dominant paradigm in LLMs' processing of Eschers.

Additionally, the results revealed that weak Eschers showed the largest amount of variance across LLMs, somewhat echoing Kelley's findings in humans where the controls had most ratings between 5-7 in a likert scale of 1-7, while strong and weak Eschers were centered around the middle, highlighting the degradation in acceptability as well as *variability* unique to Eschers.

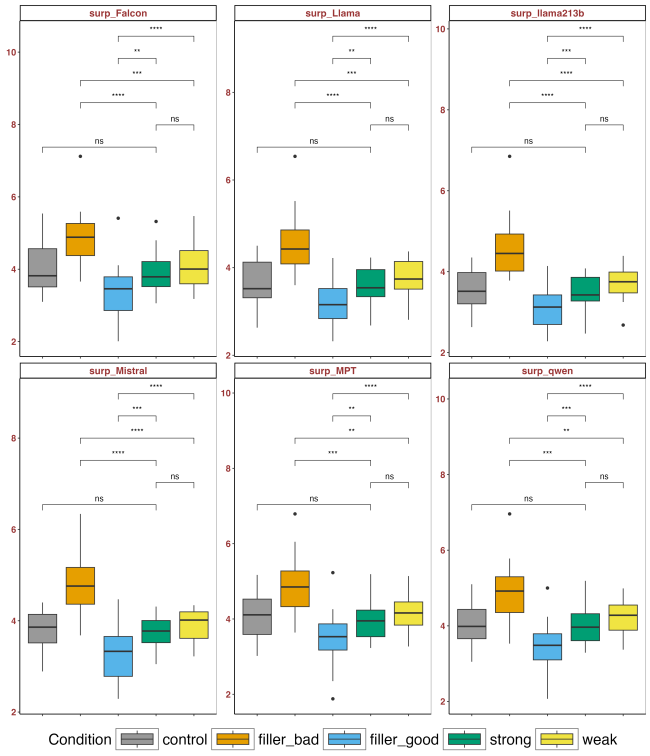


Figure 1: Average tokenwise LLMs-surprisals normalized by sentence length (*surp*). Wilcoxon pairwise comparisons with Bonferroni correction: $*p < 0.05$; $**p < 0.01$; $***p < 0.001$; ns indicates $p \geq 0.05$.

## Lexical disturbance

The prompt experiment showed that LLMs and humans generally aligned when processing Eschers, suggesting the likely absence of rational inference processing and the potential activation of good-enough processing. On the other hand, the probability experiment suggested that LLMs lack the human-level sensitivity to illusory effects. The finding that LLMs were more inclined to be "fooled" than humans indicated a potential (over-)reliance of good-enough heuristics. Our third experiment added one more piece of evidence that Eschers were "good enough" to fool LLMs, even in the scenario where the lexical substitution disturbed LLMs.

Similar to what we found in prior experiments, the prompt method led to a "human-like" rank of acceptability, whereas the surprisal measurement showed a more refined picture. Table3 showed that even with lexical disturbance, the control items remained more acceptable than strong Eschers, followed by weak Eschers. All the ratings were lower than the same items without lexical disturbance, and the range became larger (c.f. Table1), suggesting that the way LLMs handle

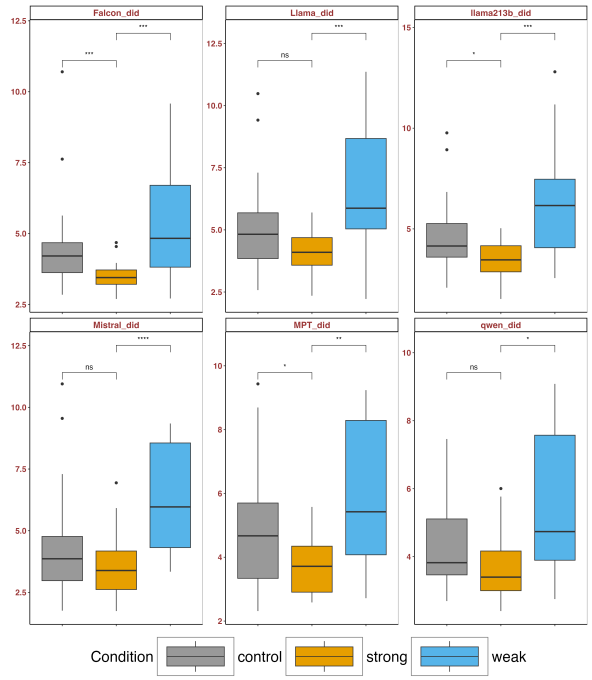


Figure 2: LLMs-surprisals at the target word *did*. Wilcoxon pairwise comparisons with Bonferroni correction: $*p < 0.05$; $**p < 0.01$; $***p < 0.001$; ns indicates $p \geq 0.05$.

comparative illusions is not just based on syntax, as their responses varied between original and substitution scenarios.

Table 3: Descriptive statistics of prompting GPT-4o-mini in Eschers acceptability ratings on nonce words stimuli.

| **Condition** | **Mean (Std)** | **Range** |
|---|---|---|
| Control | 4.444 (0.836) | [3, 6] |
| Strong | 4.308 (0.766) | [2, 6] |
| Weak | 4.167 (0.768) | [2, 6] |

Table4 showed that across different degree of lexical disturbance (s1, s2, s1s2), weak Eschers were the most surprising, hence the least acceptable. Further, LLMs again failed to tease apart strong Eschers from control items, and even "considered" strong Eschers to be significantly more acceptable than control items for the disturbance type s2.

Besides aggregated statistics across seven nonce words (Table4), we conducted *Mann-Whitney U* tests for each individual nonce word. Results showed that the variation in distribution between conditions diverges from the original stimuli without disturbance. This indicates that lexical disturbance affects GPT-2's stability when evaluating the subtler contrasts between strong and weak Eschers. We also analyzed the same items without capitalization on nonce words, with the number of significant comparisons decreased slightly. These imply that both lexical and syntactic clues are involved in LLMs' processing of comparative illusions, and semantic grounding still matters for LLMs processing of unacceptable sentences.

Table 4: Pairwise *Mann-Whitney U* test results for GPT-2 target word *did* surprisal by condition and disturbance type.

| Disturbance Type | Comparison of surprisal | $p$-value |
|---|---|---|
| s1 | Weak > Strong | 0.000 |
| s1 | Weak > Control | 0.006 |
| s1 | Strong < Control | 0.571 |
| s2 | Weak > Strong | 0.000 |
| s2 | Weak > Control | 0.003 |
| s2 | Strong < Control | 0.004 |
| s1s2 | Weak > Strong | 0.643 |
| s1s2 | Weak > Control | 0.732 |
| s1s2 | Strong < Control | 0.750 |

## Discussion

Analyses of prompting showed a consistent pattern between LLMs and human responses, suggesting LLMs can exhibit "human-like" behavior with this method. However, with probability measures, LLMs struggled to reliably rank surprisal values based on the intensity of Escher illusions. For the nonce words method, LLMs' handling of comparative illusions was influenced by both lexical and syntactic cues, as their responses varied between original and substitution cases. Overall, LLMs align with the good-enough processing seen in human cognition, constructing quick, shallow representations for efficiency. Different from our prediction that LLMs might dominantly exhibit rational inference processing, we did not find robust evidence that LLMs can flexibly incorporate rational inference processing, reinterpret structures or resolve linguistic illusions.

The rational inference framework proposes that language comprehension involves integrating various probabilistic cues to construct the most likely interpretation of a sentence. We propose that it aligns more closely with formal linguistic competence (Mahowald et al., 2024), which entails knowledge of linguistic rules and structures. On the other hand, the good-enough framework suggests that individuals often construct superficial or incomplete representations of sentences, sufficient for immediate communicative purposes. We maintain that it aligns more with functional linguistic competence (Mahowald et al., 2024), which focuses on the practical use of language in real-world contexts. Functional competence involves the ability to use language effectively and efficiently, even if it means sometimes accepting interpretations without engaging in deep structural analysis. Our findings indicate that Escher sentences are useful diagnostic tools for probing LLMs' *functional* linguistic capabilities. While LLMs exhibit some "human-like" processing, limitations in robust structure reconstruction and rational error correction highlight key differences in functional competence between humans and current model architectures, which operate under different resource constraints and optimization objectives.

## Acknowledgments

We would like to thank all the Prolific participants for the human experiments (IRB protocol number: IRB-2024-1882). This research was funded by the College of Liberal Arts, Purdue University.

## References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., ... Penedo, G. (2023). Falcon-40B: an open large language model with state-of-the-art performance. *arXiv preprint arXiv:2311.16867*.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., ... others (2023). Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Brehm, L., Jackson, C. N., & Miller, K. L. (2021). Probabilistic online processing of sentence anomalies. *Language, Cognition and Neuroscience*, *36*(8), 959–983.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc.

Cai, Z., Duan, X., Haslett, D., Wang, S., & Pickering, M. (2024). Do large language models resemble humans in language use? In *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 37–56).

Dentella, V., Günther, F., Murphy, E., Marcus, G., & Leivada, E. (2024). Testing ai on language comprehension tasks reveals insensitivity to underlying meaning. *Scientific Reports*, *14*(1), 28083.

de Varda, A. G., Gatti, D., Marelli, M., & Günther, F. (2024). Meaning beyond lexicality: Capturing pseudoword definitions with language models. *Computational Linguistics*, *50*(4), 1313–1343.

Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current directions in psychological science*, *11*(1), 11–15.

Ferreira, F., & Huettig, F. (2023). Fast and slow language processing: A window into dual-process models of cognition.[open peer commentary on de neys]. *Behavioral and Brain Sciences*, *46*.

Ferreira, F., & Patson, N. D. (2007). The 'good enough'approach to language comprehension. *Language and linguistics compass*, *1*(1-2), 71–83.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056.

Hu, J., & Levy, R. (2023). Prompting is not a substitute for probability measurements in large language models. *arXiv preprint arXiv:2305.13264*.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., ... others (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Ke, L., Tong, S., Chen, P., & Peng, K. (2024). Exploring the frontiers of llms in psychological applications: A comprehensive review. *arXiv preprint arXiv:2401.01519*.

Kelley, P. (2018). *More people understand eschers than the linguist does: The causes and effects of grammatical illusions*. Michigan State University.

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior research methods*, *42*, 627–633.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

Michaelov, J. A., & Bergen, B. K. (2022). Collateral Facilitation in Humans and Language Models. In *Proceedings of conll.*

Millière, R. (2024). Language models as models of language. *arXiv preprint arXiv:2408.07144*.

Misra, K. (2022). minicons: Enabling Flexible Behavioral and Representational Analyses of Transformer Language Models. *arXiv preprint arXiv:2203.13112*.

Misra, K., Rayz, J., & Ettinger, A. (2023, May). COMPS: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In A. Vlachos & I. Augenstein (Eds.), *Proceedings of the 17th conference of the european chapter of the association for computational linguistics* (pp. 2928–2949). Dubrovnik, Croatia: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.eacl-main.213 doi: 10.18653/v1/2023.eacl-main.213

Paape, D. (2023). When transformer models are more compositional than humans: The case of the depth charge illusion. *Experiments in Linguistic Meaning*, *2*, 202–218.

Paape, D. (2024). How do linguistic illusions arise? rational inference and good-enough processing as competing latent processes within individuals. *Language, Cognition and Neuroscience*, *39*(10), 1334–1365.

Palan, S., & Schitter, C. (2018). Prolific. ac—a subject pool for online experiments. *Journal of behavioral and experimental finance*, *17*, 22–27.

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., ... Launay, J. (2023). The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*. Retrieved from https://arxiv.org/abs/2306.01116

Qiu, Z., Duan, X., & Cai, Z. (2024, August). Evaluating grammatical well-formedness in large language models: A comparative study with human judg-

ments. In T. Kuribayashi, G. Rambelli, E. Takmaz, P. Wicke, & Y. Oseki (Eds.), *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 189–198). Bangkok, Thailand: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.cmcl-1.16/ doi: 10.18653/v1/2024.cmcl-1.16

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. In *Open-AI Blog*.

Team, M. N., et al. (2023). *Introducing mpt-7b: A new standard for open-source, commercially usable llms.* Accessed.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... others (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wellwood, A., Pancheva, R., Hacquard, V., & Phillips, C. (2018). The anatomy of a comparative illusion. *Journal of Semantics*, *35*(3), 543–583.

Yadav, H., Paape, D., Smith, G., Dillon, B. W., & Vasishth, S. (2022). Individual differences in cue weighting in sentence comprehension: An evaluation using approximate bayesian computation. *Open Mind*, *6*, 1–24.

Zhang, Y. (2024). *The rational processing of language illusions*. Harvard University.

Zhang, Y., Gibson, E., & Davis, F. (2023). Can language models be tricked by language illusions? easier with syntax, harder with semantics. *arXiv preprint arXiv:2311.01386*.