Can Language Models Be Tricked by Language Illusions? Easier with Syntax, Harder with Semantics

Yuhan Zhang
Linguistics
Harvard University
yuz551@g.harvard.edu

Edward Gibson
Brain & Cognitive Sciences
Massachusetts Institute of Technology
egibson@mit.edu

Forrest Davis
Computer Science
Colgate University
fdavis@colgate.edu

Abstract

Language models (LMs) have been argued to overlap substantially with human beings in grammaticality judgment tasks. But when humans systematically make errors in language processing, should we expect LMs to behave like cognitive models of language and mimic human behavior? We answer this question by investigating LMs' more subtle judgments associated with "language illusions" – sentences that are vague in meaning, implausible, or ungrammatical but receive unexpectedly high acceptability judgments by humans. We looked at three illusions: the comparative illusion (e.g. "More people have been to Russia than I have"), the depth-charge illusion (e.g. "No head injury is too trivial to be ignored"), and the negative polarity item (NPI) illusion (e.g. "The hunter who no villager believed to be trustworthy will ever shoot a bear"). We found that probabilities represented by LMs were more likely to align with human judgments of being "tricked" by the NPI illusion which examines a structural dependency, compared to the comparative and the depth-charge illusions which require sophisticated semantic understanding. No single LM or metric yielded results that are entirely consistent with human behavior. Ultimately, we show that LMs are limited both in their construal as cognitive models of human language processing and in their capacity to recognize nuanced but critical information in complicated language materials.

1 Introduction

Linguistic evaluations of language models use human language processing data (e.g. human norming data (Nair et al., 2020; Zhang et al., 2022), acceptability judgments (Linzen et al., 2016; Marvin and Linzen, 2018), behavioral or neural measures of language processing (Schrimpf et al., 2021; Kauf et al., 2022)) as benchmarks to investigate whether LMs possess knowledge of language. This assumes that human-produced data correctly instantiates abstract rules of a language and that humans fully utilize their linguistic knowledge in laboratories and everyday life. However, this assumption is an oversimplification. Humans make consistent errors during language processing (Gross, 1983). Under these circumstances, should we expect language models to behave the same as humans? Or should they circumvent human limitations and achieve error-free performance?

Consider, for example, the well-studied case of subject-verb agreement. While we expect an LM of Standard American English to prefer "the key to the cabinets **is** on the shelf" to "the key to the cabinets **are** on the shelf" (as discussed in Linzen et al., 2016), a wealth of psycholinguistic research has systematically documented that humans can ignore errors and accept globally ungrammatical strings (stemming from Bock and Miller, 1991). Should LMs follow the ideal grammar or mimic human's (sometimes) errorful behavior?[1]

We add to this discussion by investigating three language illusions. Basic examples of each are given in (1): the comparative illusion (1-a), the depth-charge illusion (1-b), and the negative-polarity item (NPI) illusion (1-c). All three in (1) are literally unnatural English sentences, despite the fact that humans often find them surprisingly acceptable.

(1) a. More people have been to Russia than I have.
b. No head injury is too trivial to be ignored.
c. The hunter who no villager believed to be trustworthy will ever shoot a bear.

In this paper, we relied on minimally different strings springing out from the basic illusion sentences that are either (a) considered fully **acceptable** by human participants, (b) considered fully

[1]For additional critiques of the role of ideal grammatical knowledge in evaluations of LMs, see Pannitto and Herbelot (2020); Weissweiler et al. (2023).

unacceptable by human participants, or (c) rated **surprisingly acceptable** by humans (i.e. instances of the relevant illusion). We explored whether language models capture the basic contrast between acceptable and unacceptable strings, whether they rate illusion sentences as better than their unacceptable counterparts, and finally, whether models capture nuanced linguistic manipulations that influence human judgments of the illusion material. Further, we compared two ways of measuring models' preferences, one over the whole sentence (*perplexity*) and another of a privileged position in the sentence (*surprisal*).

If LMs pattern like human comprehension behavior that involves errors, we expect to derive measures that similarly rate illusion sentences as more acceptable than typical unacceptable sentences. If, on the other hand, LMs align with ideal grammatical judgments, illusion sentences should be rated as unacceptable. Our findings indicate that none of the language models we investigated consistently exhibited illusion effects or demonstrated overall human-like judgment behaviors. Nor do they possess the necessary linguistic knowledge for error-free, literal sentence processing. These findings add more insights into the discussion of LMs' emulation of human behavior and their construal as cognitive models of human language processing.

2 Related work

2.1 LMs' linguistic abilities

We draw insights from evaluation work relying on acceptability tasks. The construction of minimal pairs has been used to evaluate models for a variety of linguistic processes, including subject-verb agreement (e.g. Linzen et al., 2016), filler-gap dependency (e.g. Wilcox et al., 2018), control (e.g. Stengel-Eskin and Van Durme, 2022), and binding (e.g. Davis, 2022). This basic template has been expanded into a variety of benchmarks, both for investigations of English (e.g. Warstadt et al., 2020), but also, other languages (e.g. Chinese (Song et al., 2022); Russian (Mikhailov et al., 2022); Japanese (Someya and Oseki, 2023)). While aggregated results suggest that models overlap with human acceptability judgments in a variety of cases (e.g. Hu et al., 2020), LMs can behave in distinctly non-human-like ways in capturing the intricacies of grammatical phenomenon (e.g. Lee and Schuster, 2022), the interaction between linguistic processes (e.g. Davis and van Schijndel, 2020), and in generalizing knowledge to infrequent items (e.g. Wei et al., 2021).

In our experiments, we are interested in cases where human interpretations and behaviors differ from what is expected given the literal content of the entire string. *Garden path* sentences are a classic example of this basic phenomenon. Strings like "The horse raced past the barn fell" are often difficult for humans on first reading because the word *raced* is misparsed as a main verb (e.g. *the horse raced past*) rather than a reduced relative clause (e.g. *the horse that was raced past the barn fell*). LMs have been shown to similarly misprocess these sentences (van Schijndel and Linzen, 2021), though they fall short of capturing the magnitude of the processing cost (Arehalli et al., 2022). Here we expand these investigations to language illusions that similarly trigger errorful acceptable judgments in humans while being unnatural and unacceptable. We find that LMs do not pattern like humans in all cases.

2.2 Language illusions

Language illusions refer to ungrammatical, semantically vague, or pragmatically implausible sentences that receive higher than expected acceptability by humans (Phillips et al., 2011). We study three language illusions in particular: comparative illusion (Montalbetti, 1984) (Section 4), depth-charge illusion (Wason and Reich, 1979) (Section 5), and NPI illusion (Xiang et al., 2009) (Section 6). Existing human research has found that the illusion effects for both the comparative and the depth-charge illusion are robust and overwhelming but the NPI illusion effect only appears during speeded judgment tasks or word-by-word online paradigms (Parker and Phillips, 2016; Wellwood et al., 2018; Paape et al., 2020; Orth et al., 2021).

For human sentence processing, it has been suggested that language illusions provide evidence for rational inference of error-prone strings which integrates heuristics and available context information during processing (Ferreira et al., 2002; Levy, 2008; Gibson et al., 2013; Futrell et al., 2020; Hahn et al., 2022; Zhang et al., 2023a). These phenomena raise fundamental questions like what is the role of our grammatical knowledge in comparison to other cognitive resources when it comes to assigning a specific interpretation to a linguistic string, and how we can model their interactions to make better predictions about human sentence processing.

Studying LMs' processing of language illusions provides a way to explore whether they can be viewed as cognitive models of human sentence processing. As large language models like ChatGPT improve at generating grammatically appropriate strings, it becomes ever more important to investigate whether they are comparable to human language processing behavior at all (see Mahowald et al., 2023, for a review). From there, we can reason about what characteristics in the training of LMs, the architecture of LMs, and the "abilities" of LMs enable them to carry out either literal interpretations and detect the anomaly, or to fall into the illusion rabbit hole.

3 Methods

3.1 Models and Measures

We analyzed four models, two masked language models, and two autoregressive models: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020). BERT, RoBERTa, and GPT-2 were accessed via HuggingFace (Wolf et al., 2020), and GPT-3 via OpenAI's API.[2] We used two measures, sentence level perplexity and surprisal of specific target words. For autoregressive models, the surprisal of a specific word[3] is given by the following equation:

$$\text{Surp}(\text{w}_i) = -\log \text{Prob}(\text{w}_i|\text{w}_1 ... \text{w}_{i-1}) \quad (1)$$

Perplexity for a sentence of N words is:

$$2^{\frac{1}{N}\sum_{i=1}^{N} \text{Surp}(\text{w}_i)} \quad (2)$$

For bidirectional models, we calculated the surprisal of a word in a context by using the masking technique in Kauf and Ivanova (2023), which corrects for words that are subworded.[4] Further, we used this masking technique to calculate the pseudo-perplexity of a sentence (Salazar et al., 2020).

[2]We used 'bert-base-cased', 'roberta-base', 'gpt2', and 'text-davinci-003'. Code for replicating the results, statistical tests, and figures can be found at https://github.com/forrestdavis/LanguageIllusions.git .

[3]For words that are subworded, the joint probability was calculated.

[4]For example, consider the word 'souvenir'. This is subworded by BERT into 'so', '##uven', and 'ir'. Rather than MASK each subpart, one at a time, (e.g. 'so' [MASK] 'ir'), the right context of the target subword is always masked (e.g. 'so' [MASK] [MASK]).

3.2 Evaluation procedure

We treated LMs as psycholinguistic research subjects to generate both whole-sentence perplexity and surprisals at critical words for carefully controlled minimal pairs for each illusion (following, Futrell et al., 2019). Assuming these two scores are correlated to human acceptability judgments (Lau et al., 2017), we constructed mix-effects linear regression models from the R package $lme4$ to test whether LMs were also sensitive to reported manipulations that affect human judgments. For each scoring metric, we took it as the dependent variable and coded the manipulation condition representing a certain hypothesis into the independent variable. We read the estimated coefficient(s) of the tested condition variable(s) to infer whether LMs show sensitivity to the effect of that condition manipulation on the scoring metric. We evaluated language models in three broad aspects: acceptability differentiation, illusion effect, and sensitivity to manipulations.

- **Acceptability differentiation** We first asked whether language models could distinguish acceptable sentences from unacceptable sentences that humans have no trouble dealing with.[5] Models with relevant knowledge should assign lower perplexity/surprisal to acceptable sentences versus unacceptable ones.

- **Illusion effect** We took the results from the acceptability differentiation task as the foundation to test the illusion sentences. Here, we hypothesized that language models should either (i) align with humans' illusionary judgments, reflected by models' generating a lower perplexity/surprisal for illusion sentences than the unacceptable controls, or (ii) deviate from human behavior and show hints of being a literal processor, reflected by models' generating a higher or similar perplexity/surprisal score compared to the unacceptable condition. If models behave like humans, then we expected (i) to be the models' consistent behavior. If models conform to (ii), we take this as evidence of non-human-like behavior.

- **Sensitivity to manipulations** Lastly, we assessed whether language models were sensi-

[5]According to finer-grained linguistic criteria, acceptable sentences are those that are grammatical, plausible, and felicious. Please refer to Tonhauser and Matthewson (2015) for detailed definitions and review.

Illusion type	item	BERT		RoBERTa		GPT-2		GPT-3	
		PPL	Surp	PPL	Surp	PPL	Surp	PPL	Surp
Comparative	32	-0.36	-0.001	-0.56	-0.09	-0.22	-0.05	-0.30	-0.25
Depth-charge	32	-0.37	-0.15	-0.61	-0.45	-0.12	-0.41	-0.37	-0.98
NPI	32	-0.26	-2.46	-0.71	-2.60	-0.21	-1.73	-0.29	-2.55

Table 1: Estimated coefficients of the main effect (acceptable sentence condition vs. unacceptable condition (reference)) for each statistical model. If LMs rate acceptable sentences as more acceptable, the coefficients for perplexity or surprisal should be significantly negative. Cells color-coded in blue represent statistical significance level ($p < .05$) in the expected direction. White cells represent an insignificant main effect. In other words, blue cells indicate the statistical model output supports LMs' ability to distinguish sentences based on linguistic acceptability.

tive to illusion-specific linguistic manipulations that affect human judgments. A greater degree of sensitivity indicates that the corresponding linguistic knowledge and how the knowledge affects sentence acceptability could be encoded in or learned by LMs. This allowed us to draw a fine-grained comparison between humans and LMs. If language models are insensitive, that indicates a difference between humans and LMs.

4 Comparative illusion

A canonical comparative illusion surfaces in sentences like "More people have been to Russia than I have". People accept it at first glance but have trouble pinning down the exact meaning (Montalbetti, 1984) one of which could be that the number of the group of people who've been to Russia is greater than the number of "me". Potential rational nonliteral inference could be "people have been to Russia more times than I have" or "people have been to Russia but I haven't" (O'Connor, 2015; Christensen, 2016). Psycholinguistic research has found that various factors modulate the strength of the illusion, including the repeatability of the event described by the verb phrase, the subject form of the than-clause subject (e.g. "... than the student has" vs. "...I have"), as well as the number of that subject (e.g. "I have" vs. "we have")(Wellwood et al., 2018). There is also a claim arguing that the processing mechanism follows the noisy-channel predictions under an information-theoretic account (Zhang et al., 2023b).

We adapted the experimental materials with 32 items from Zhang et al. (2023b).[6] An example is in (2) where (2-a) is the canonical comparative illusion, (2-b) is the acceptable control, and (2-c) is the unacceptable one.[7]

(2) a. (?) More teenagers have used Tiktok than I have. (illusion)
b. Many teenagers have used Tiktok more than I have. (acceptable)
c. (#) Many teenagers have installed Tiktok more than I have. (unacceptable)

4.1 Acceptability differentiation

We first ensured that LMs distinguish acceptable neighbors (2-b) of the illusion sentence from unacceptable ones (2-c). We ran statistical mixed-effects linear regression models on whole-sentence perplexity and the surprisal at the word *have* for the four language models. Either the perplexity or the surprisal was taken as the dependent variable with the condition "acceptability" as the fixed effect (reference level = the unacceptable condition, with a nonrepeatable verb phrase vs. the acceptable condition, with a repeatable verb phrase) and the random intercept of each item as the random effect.[8]

Table 1 shows the estimated coefficient for the main effect of each mixed-effect model for each LM and each illusion phenomenon. A significant negative estimated coefficient suggests that acceptable sentences received lower perplexity/surprisal compared to the unacceptable ones, indicating that LMs distinguish sentences based on acceptability. Except for surprisal values from BERT and GPT-2, the other six statistical models indicate that the LMs capture the acceptability difference of baseline sentences for the comparative illusion.

4.2 Illusion effect

This task investigated whether language models pattern with humans in demonstrating illusion effects

[6] See Table 3 in the Appendix for the full paradigm.

[7] The repeatability of the verb phrase is responsible for this contrast, as it is more natural to say "use Tiktok more often or frequent" compared with "install Tiktok more often" when the action typically takes place once (in a while).

[8] The model syntax in R was `PPL/SURP ~ acceptability + (1|item)`.

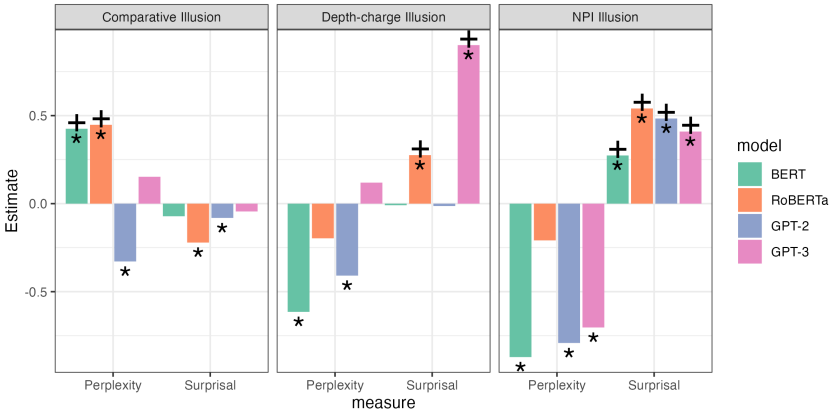

Figure 1: The y axis shows the coefficient estimates which represent the increase in perplexity/surprisal when the sentence is unacceptable compared to the illusion case, crossing three language illusions and four LMs. "+" marks a human-like behavior, in this case, an illusion effect where the unacceptable condition receives significantly higher perplexity/surprisal values than the illusion condition. "*" means that the estimated coefficient is significant.

with the basic comparative illusion construction. The contrast involves the illusion condition (2-a) with existing control conditions ((2-b) and (2-c)). The standardized metrics of the four LMs are displayed in Figure 6 in the Appendix. To evaluate whether LMs capture an illusion effect, we constructed another suite of statistical models across the four LMs and two metrics where the main effect has three levels – the illusion condition (reference), the acceptable condition, and the unacceptable condition – and the random effect included a random intercept for items.[9]

We analyzed the coefficient estimates of the main effect of the unacceptable condition compared with the illusion condition.[10] An illusion effect would appear with higher perplexity/surprisal for the unacceptable condition compared to the illusion case. In other words, the estimated coefficients for the unacceptable condition should be significantly positive.

Figure 1 and Table 2 (in Appendix) display the estimated coefficients for the unacceptable condition compared with the illusion condition. For the comparative illusion, only BERT and RoBERTa measured by perplexity show a human-like illusion effect. Other LM-metric combinations indicate that the illusion condition was rated either the same or worse than the unacceptable condition (contrary to humans).

[9]The model syntax in R was `PPL/SURP ~ condition + (1|item)` where `condition` had three levels.

[10]The coefficients for the acceptable condition generate similar conclusions. Further, no illusion sentences were rated better than acceptable ones.

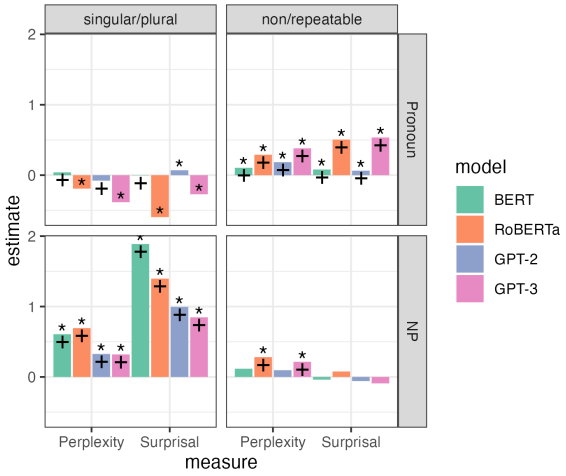

Figure 2: Estimated coefficients for critical linguistic manipulations in **comparative illusion**. The y axis shows the estimated coefficients for the increase in perplexity/surprisal with respect to singular vs. plural than-clause subjects, or nonrepeatable vs. repeatable verb phrases, respectively. "*" means statistically significant contrasts; "+" means human-like results.

4.3 Sensitivity to manipulations

In this step, we evaluated whether language models were sensitive to sentence manipulations that affect human judgments. Three factors were investigated: (1) than-clause subject structure (pronoun vs. NP), (2) subject number (singular vs. plural), and (3) verb repeatability (repeatable vs. nonrepeatable). For humans, plural than-clause subjects are more acceptable than singular ones only in the NP case. Overall, repeatable verbs are more acceptable than nonrepeatable ones (O'Connor, 2015; Wellwood et al., 2018; Zhang et al., 2023b).

Figure 2 displays the estimated coefficients for the main effects from the statistical models.[11] As for the subject number, when the than-clause subject was a pronoun, only BERT and GPT-2 (with perplexity) aligned with human-like behavior: there is no difference between singular and plural than-clause subjects. When it comes to NP subjects, all four LMs with both metrics showed human-like behavior where the singular NP subject was more unacceptable than the plural NP subject. As for repeatability, all four LMs captured this distinction in the pronoun condition but in the NP condition, only RoBERTa and GPT-3 achieved human-like results with perplexity.

In general, we only found partial overlap between LMs and humans. This indicates that even though LMs show some knowledge of acceptability for comparative structures, they might operate differently from humans when processing more subtle differences. None of the language models fully captured all the manipulations.

5 Depth-charge illusion

Consider the most famous depth-charge sentence *No head injury is too trivial to be ignored* (Wason and Reich, 1979). People overwhelmingly interpret it as meaning "no matter how trivial head injuries are, we should not ignore them", while the literal meaning is the opposite as "we should ignore them".

To understand the depth-charge sentence requires knowing meaning composition rules, multiple negation processing (Wason and Reich, 1979), adequate world knowledge reasoning (Paape et al., 2020), and the neighboring constructions of *too...to* such as *so...that*, *so...as to* and *enough to...* (Zhang et al., 2023a). Since existing research already shows that language models are quite limited in processing negation (e.g. Kassner and Schütze, 2019; Ettinger, 2020), we speculate that LMs might encounter difficulty in the more complicated case of depth-charge sentences.

The evaluation materials were adapted from Zhang et al. (2023a) with 32 items. An example is (3) where we take the surprisal of the sentence-final word for comparison.

(3) a. (?) No head injury is too trivial to be ignored. (depth-charge sentence)
b. Some head injury is too severe to be ignored. (plausible, acceptable)
c. (#) Some head injury is too trivial to be ignored. (implausible, unacceptable)

5.1 Acceptability differentiation

Utilizing the same methodology as the comparative illusion, we found, as depicted in Table 1, that all combinations of LMs and metrics, except GPT-2 (perplexity), captured the acceptability difference between ((3-b)) and ((3-c)) with a significantly lower perplexity/surprisal for the acceptable sentences like (3-b).

5.2 Illusion effect

Next, we studied if LMs "experience" the illusion effect by assigning lower perplexity/surprisal scores to the depth-charge sentence (3-a) compared to the unacceptable one (3-c).

Our statistical results show, in Figure 1 and Table 2 (Appendix), that only RoBERTa and GPT-3 demonstrated an illusion effect (for surprisal) by assigning a significantly higher score to the unacceptable control sentences. This means that it is not easy to "trick" LMs with the depth-charge illusion. Similar results have led concurrent work to suggest that LMs are better at deriving the literal meaning of a sentence, which is in sharp contrast with the overwhelming illusion effect from humans (Paape, 2023, a.o.).

5.3 Sensitivity to manipulations

This task tested LMs' sensitivity to the plausibility contrast of three near-neighbor pairs of the depth-charge sentence. These pairs differ by the degree quantifier construction (*too...to* vs. *so...as to* vs. *too...to not*).[12] Competent language models should differentiate plausible sentences from implausible ones.

Figure 3 displays estimated coefficients of statistical models' main effect. We expect implausible sentences to receive higher perplexities/surprisals when the illusion occurs.[13] We find that LMs captured some of the distinctions in the *too...to* condition and the *so...as to* condition. However, im-

[11] More statistic model information: Iterating over LMs, metrics, and the subject structure (NP vs. pronoun), we initiated statistical models taking both repeatability (reference = repeatable) and subject number (reference = plural) as the main effects with the random effect including a random intercept for the items.

[12] The full suite of paradigms is shown in Table 4 in the Appendix.

[13] Iterating over sentence pairs, LMs, and metrics, we ran mixed-effects linear regression models on scores over the plausibility contrast (reference = plausible).

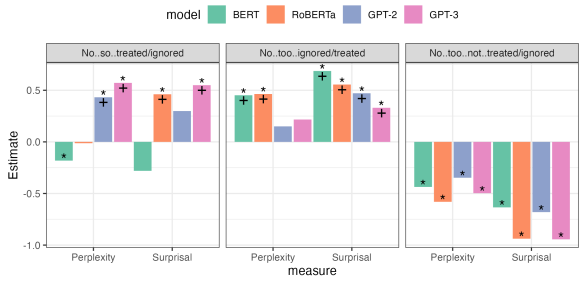

Figure 3: Estimated coefficients for the plausibility contrast (reference = plausible) in **depth-charge illusions**. The y axis shows the increase in perplexity/surprisal when the sentence is implausible vs. plausible. "*" means statistically significant contrasts; "+" means human-like behavior. While we see differences among LMs and metrics in the "no...so...as to" and the "no...too...to" conditions, the condition of "no...too...to not" yielded completely opposite results to humans.

plausible sentences with *too...to not* were rated as more acceptable than their plausible counterparts, which flouts what linguistic rules predict.[14] The fact that "No head injury is too trivial to be treated" and "No head injury is too trivial to not be ignored" generate opposite results while having the same meaning suggests LMs still struggled with negation, antonyms, and meaning composition (Kim and Linzen, 2020; She et al., 2023; Truong et al., 2023).

6 NPI illusion

Negative polarity items and their licensing conditions have been investigated in prior work with language models. For a canonical NPI (e.g. *ever*, *any*) to be acceptable, it has to be in the scope of negation.[15] Existing computational research has shown that the syntactic dependency between the licensor and the NPI is captured by language models (Jumelet and Hupkes, 2018; Jumelet et al., 2021; Shin et al., 2023) but with more difficulty as compared to subject-verb agreement or other syntactic dependencies (Marvin and Linzen, 2018; Warstadt et al., 2019, 2020). In this task, we expanded the suite of LMs and metrics and explored sensitivities to four types of licensors.

[14]The sentence *No head injury is too trivial to not be ignored* should be plausible because compositionally, "too trivial to not be ignored" means "too trivial to be treated" which yields a plausible sentence given the sentential negation.

[15]The licensing conditions of negative polarity items are far more than in the scope of negation. We focus on the classic licensing condition and refer to Giannakidou et al. (2019) for a review.

Our materials came from Orth et al. (2021) with 32 items. The essential triad is (4) where the illusion condition has the NPI *ever* not in the scope of the negation word *no*.

(4) a. (?) The hunter who no villager believed to be trustworthy will ever shoot a bear. (NPI illusion)
b. No hunter who the villager believed to be trustworthy will ever shoot a bear. (Matrix No, acceptable)
c. (*) The hunter who the villager believed to be trustworthy will ever shoot a bear. (Licensor Absent, unacceptable)

6.1 Acceptability differentiation

Table 1 shows that all the four LMs could capture the acceptability difference of control sentences (4-b) and (4-c) (with both metrics).

6.2 Illusion effect

Figure 1 and Table 2 show that only in the case of surprisal did we see an illusion effect where the unacceptable sentences (e.g. (4-c)) received significantly higher surprisals than the illusion sentence (e.g., (4-a)). This finding replicates Shin et al. (2023) in that, for the illusion condition ((4-a)) where *no* linearly precedes *ever* but is in an unlicensing position, *ever* incurs higher surprisal. It is interesting to see the sharp discrepancy between surprisal and perplexity, which we leave to Section 7.4 for discussion.

6.3 Sensitivity to variations

The linguistic manipulations we explored concern the illusion effect in the illusion condition with different NPI licensors. Among the ones we tested, *didn't*, *did not*, and *never*,[16] human research shows that none of these triggers illusion effects (Orth et al., 2021; cf. Vasishth et al., 2008).

Iterating over licensors, LMs, and metrics, we ran statistical models with the same structure in Section 6.2. We plotted the estimated coefficients of the unacceptable main effect in Figure 4 and predicted that a significantly positive coefficient indicates an illusion effect. Contrary to human-like behavior, for all three licensors there were some LM-metric combinations that indicate an illusion

[16]Please refer to Table 5 for the full experimental conditions.

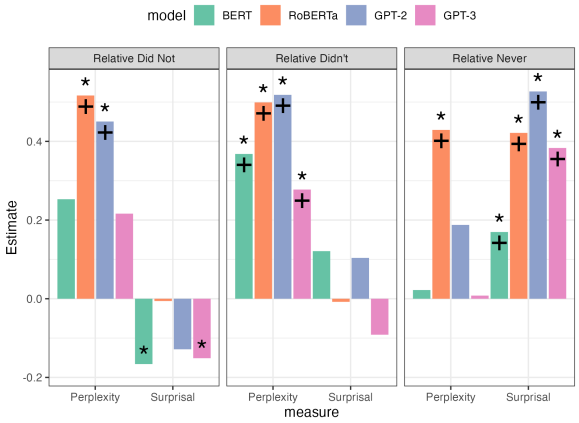

Figure 4: Estimated coefficients for the illusion effect (unacceptable vs. illusion = reference) in **NPI illusions**. The y axis shows the increase in perplexity/surprisal when the sentence is ungrammatical vs. is in the illusion condition. "+" marks an illusion effect while none of the three licensors should trigger an illusion effect according to human behavior; "*" means a significant contrast.

effect: for the licensor *did not*, RoBERTa (perplexity) and GPT-2 (perplexity) show an illusion effect; for *didn't*, all four LMs with perplexity show an illusion effect; for *never*, all four LMs with surprisal, plus RoBERTa with perplexity, show an illusion effect. This pattern shows that with NPI illusions, LMs are more easily tricked than humans.

7 Discussion

7.1 Illusion effect

Successful language processing requires a dynamic integration of lexical knowledge, grammatical knowledge, logical reasoning, and world knowledge, among other cognitive abilities and sources of knowledge. An illusion effect in humans where unacceptable sentences receive unexpectedly high acceptability presents a unique case where the comprehender might prioritize different processing mechanisms or linguistic constraints for meaning inference over those employed for common processing. Studying how language models process language illusions helps us understand (1) from a superficial level, whether LMs appear to be human-like – circumventing some grammatical facts and reaching a good-enough sentence representation, and (2) from a deeper level, whether LMs employ the same set of resources and abilities to process a sentence (i.e. whether they can serve as cognitive models).

In this research, we aim for the first level of understanding. By studying four language models' acceptability judgments of three language illusions, we found that LMs were good at the basic acceptability differentiation task and yet no LMs showed consistent human-like illusion effects among three illusion phenomena by any metric (Figure 5). We conclude from this result that LMs might not be a good cognitive model of human language processing. With this said, we do observe a divergence between the comparative/depth-charge illusion and the NPI illusion – it seems more likely for LMs to be tricked by the NPI illusion compared to the former two. Since the NPI illusion is more relevant to the hierarchical structure of language whereas both the comparative illusion and depth-charge illusion emphasize semantic nuances, we tentatively conclude that LMs are more easily tricked by syntactic illusion rather than semantic illusions.

7.2 Human-like behaviors & Potential processing mechanisms

For both the comparative illusion and depth-charge illusion, the illusion effect test did not show human-like behavior. This could either mean that LMs strictly abide by linguistic rules to compose the language literally or that LMs have trouble understanding this complicated set of sentences overall. For the comparative illusion, the sensitivity task (Section 4.3) suggests that they might have some capacity to process comparative structures. For the depth-charge illusion, that LMs seem to have trouble understanding the literal contrast between plausible/implausible pairs (Section 5.3) suggests sentences involving multiple negations could pose a challenge to LMs. The two cases indicate we still need to develop more robust evaluations to gauge LMs' semantic capabilities in various semantic domains.

For the NPI illusion, the interpretation could be more complicated. On one hand, the illusion test for the licensor *no* yields human-like results (with surprisal) but other licensors also elicit non-human-like illusion effect (cf. Orth et al., 2021). On the other hand, the discrepancy between sentence perplexity and surprisal makes it difficult to conclude to what degree LMs and humans overlap (cf. Shin et al., 2023).

Ultimately, we want to address whether LMs are like humans that utilize not only grammatical rules but also contexts, frequencies, and semantic priors to rationally process language, or LMs are like

		BERT		RoBERTa		GPT-2		GPT-3	
		PPL	Surp	PPL	Surp	PPL	Surp	PPL	Surp
Comparative illusion	Acceptability differentiation	✓		✓		✓	✓	✓	✓
	Illusion effect	✓		✓					
	Number effect: Pronoun	✓	✓			✓			
	Number effect: NP	✓	✓	✓	✓	✓	✓	✓	✓
	Repeatability: Pronoun	✓	✓	✓	✓	✓	✓	✓	✓
	Repeatability: NP			✓				✓	
Depth-charge illusion	Acceptability differentiation	✓	✓	✓	✓		✓	✓	✓
	Illusion effect				✓				✓
	Plausibility contrast (so...as to)				✓	✓		✓	✓
	Plausibility contrast (too...to)	✓	✓	✓	✓		✓		✓
	Plausibility contrast (too...to not)								
NPI illusion	Acceptability differentiation	✓	✓	✓	✓	✓	✓	✓	✓
	Illusion effect for Relative No		✓		✓		✓		✓
	Illusion effect for Relative Did not	✓	✓		✓		✓	✓	✓
	Illusion effect for Relative Didn't		✓		✓		✓		✓
	Illusion effect for Relative Never	✓				✓		✓	

Figure 5: Language models' performance on all three illusions. ✓means LMs show human-like behavior.

grammarians that interpret string inputs in a strict compositional manner. Our investigation does not yield consistent results given the three language illusions but the behavioral inconsistency suggests that language models are far from being a cognitive model of human language.

7.3 Language models' performance in general

All four language models performed on par with each other. If we tallied the number of tests where LMs reported expected results from Figure 5 and averaged between perplexity and surprisal, we have a ranking order from RoBERTa (N=10) and GPT-3 (N=9), to BERT (N=8.5) and GPT-2 (N=8). The successors of both the masked language model and the autoregressive model perform better than their predecessors.

7.4 Perplexity & Surprisal

It is surprising to see that the two widely used probability-based metrics can generate different results for a given hypothesis and a given language model. Future work should (i) investigate both mathematically and practically why the difference could occur and (ii) check if better definitions for the critical regions exist to capture surprisals. Future evaluation work that utilizes one metric should be mindful of the intrinsic limitations of that metric.

7.5 Limitations

Considering the research methodology, acceptability judgment tasks (even with carefully controlled minimal pairs) are indirect measures of language comprehension and it is hard to infer the exact interpretation based on probability-based measures. Further studies should work on direct comprehension measures (e.g. generating paraphrases) that reveal LMs' hidden knowledge.

8 Conclusion

We tested four language models' ability to process three language illusions and asked (1) whether they judge unacceptable illusion sentences to be more acceptable as humans (termed an illusion effect) and (2) whether they are sensitive to linguistic manipulations that modulate human judgments. Our results are based on whole-sentence perplexity and critical word surprisal. We show that none of the LMs demonstrated consistent illusion effects or exhibited overall human-like judgment behaviors. We conclude that given the case of language illusions, language models neither behave like humans with full sets of cognitive abilities and error-prone behavior nor possess the necessary linguistic knowledge for error-free, literal sentence processing. Language models cannot be viewed as cognitive models of language processing, which makes understanding them even more intriguing.

Acknowledgements

We thank the three anonymous reviewers for their helpful feedback. We thank Ankana Saha, Carina Kauf and Hayley Ross for the helpful discussion about the project.

References

Suhas Arehalli, Brian Dillon, and Tal Linzen. 2022. Syntactic Surprisal From Neural Models Predicts, But Underestimates, Human Processing Difficulty From Syntactic Ambiguities. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 301–313, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kathryn Bock and Carol A Miller. 1991. Broken agreement. *Cognitive psychology*, 23(1):45–93.

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

Ken Ramshøj Christensen. 2016. The dead ends of language: The (mis) interpretation of a grammatical illusion. *Let us have articles betwixt us–Papers in Historical and Comparative Linguistics in Honour of Johanna L. Wood. Aarhus: Aarhus University*, pages 129–160.

Forrest Davis. 2022. Incremental Processing of Principle B: Mismatches Between Neural Models and Humans. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 144–156, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Forrest Davis and Marten van Schijndel. 2020. Discourse structure interacts with reference but not syntax in neural language models. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 396–407, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv:1810.04805 [cs].

Allyson Ettinger. 2020. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8:34–48. Place: Cambridge, MA Publisher: MIT Press.

Fernanda Ferreira, Karl G.D. Bailey, and Vittoria Ferraro. 2002. Good-Enough Representations in Language Comprehension. *Current Directions in Psychological Science*, 11(1):11–15.

Richard Futrell, Edward Gibson, and Roger P. Levy. 2020. Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. *Cognitive Science*, 44(3).

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. *arXiv preprint arXiv:1903.03260*.

Anastasia Giannakidou, Klaus von Heusinger, Claudia Maienborn, and Paul Portner. 2019. Negative and positive polarity items. *Semantics—Sentence and information structure*, pages 69–134.

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

Herbert S. Gross. 1983. Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen, and Hand. *The journal of nervous and mental disease*, 171(12):753–.

Michael Hahn, Richard Futrell, Roger Levy, and Edward Gibson. 2022. A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, 119(43):e2122602119. Publisher: Proceedings of the National Academy of Sciences.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Jaap Jumelet, Milica Denic, Jakub Szymanik, Dieuwke Hupkes, and Shane Steinert-Threlkeld. 2021. Language Models Use Monotonicity to Assess NPI Licensing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4958–4969, Online. Association for Computational Linguistics.

Jaap Jumelet and Dieuwke Hupkes. 2018. Do Language Models Understand Anything? On the Ability of LSTMs to Understand Negative Polarity Items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 222–231, Brussels, Belgium. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2019. Negated LAMA: birds cannot fly. *CoRR*, abs/1911.03343.

Carina Kauf and Anna Ivanova. 2023. A Better Way to Do Masked Language Model Scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935, Toronto, Canada. Association for Computational Linguistics.

Carina Kauf, Anna A Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan S She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2022. Event knowledge in large language models: the gap between the impossible and the unlikely. *arXiv preprint arXiv:2212.01488*.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. *CoRR*, abs/2010.05465.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive Science*, 41(5):1202–1241.

Soo-Hwan Lee and Sebastian Schuster. 2022. Can language models capture syntactic associations without surface cues? a case study of reflexive anaphor licensing in English control constructions. In *Proceedings of the Society for Computation in Linguistics 2022*, pages 206–211, online. Association for Computational Linguistics.

Roger Levy. 2008. A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '08*, page 234, Honolulu, Hawaii. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv:1907.11692 [cs].

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective. ArXiv:2301.06627 [cs].

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

Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov, and Ekaterina Artemova. 2022. RuCoLA: Russian corpus of linguistic acceptability. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5207–5227, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mario M. Montalbetti. 1984. *After binding: On the interpretation of pronouns*. PhD Thesis, Massachusetts Institute of Technology.

Sathvik Nair, Mahesh Srinivasan, and Stephan Meylan. 2020. Contextualized Word Embeddings Encode Aspects of Human-Like Word Sense Knowledge. *arXiv:2010.13057 [cs]*. ArXiv: 2010.13057.

Ellen O'Connor. 2015. *Comparative illusions at the syntax-semantics interface*. Ph.D. thesis, University of Southern California.

Wesley Orth, Masaya Yoshida, and Shayne Sloggett. 2021. Negative polarity item (NPI) illusion is a quantification phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47(6):906–947. Publisher: American Psychological Association.

Dario Paape. 2023. When Transformer models are more compositional than humans: The case of the depth charge illusion. *Experiments in Linguistic Meaning*, 2(0):202–218. Number: 0.

Dario Paape, Shravan Vasishth, and Titus von der Malsburg. 2020. Quadruplex negatio invertit? The on-line processing of depth charge sentences. *Journal of Semantics*, 37(4):509–555.

Ludovica Pannitto and Aurélie Herbelot. 2020. Recurrent babbling: Evaluating the acquisition of grammar from limited input data. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 165–176, Online. Association for Computational Linguistics.

Dan Parker and Colin Phillips. 2016. Negative polarity illusions and the format of hierarchical encodings in memory. *Cognition*, 157:321–339.

Colin Phillips, Matthew W. Wagers, and Ellen F. Lau. 2011. Grammatical illusions and selective fallibility in real-time language comprehension. *Experiments at the Interfaces*, 37:147–180.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Jingyuan Selena She, Christopher Potts, Samuel R Bowman, and Atticus Geiger. 2023. Scone: Benchmarking negation reasoning in language models with fine-tuning and in-context learning. *arXiv preprint arXiv:2305.19426*.

Unsub Shin, Eunkyung Yi, and Sanghoun Song. 2023. Investigating a neural language model's replicability of psycholinguistic experiments: A case study of NPI licensing. *Frontiers in Psychology*, 14:937656.

Taiga Someya and Yohei Oseki. 2023. JBLiMP: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia. Association for Computational Linguistics.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. SLING: Sino linguistic evaluation of large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Elias Stengel-Eskin and Benjamin Van Durme. 2022. The curious case of control. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11065–11076, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Judith Tonhauser and Lisa Matthewson. 2015. Empirical evidence in research on meaning.

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. Language models are not naysayers: An analysis of language models on negation benchmarks. *arXiv preprint arXiv:2306.08189*.

Marten van Schijndel and Tal Linzen. 2021. Single-Stage Prediction Models Do Not Explain the Magnitude of Syntactic Disambiguation Difficulty. *Cognitive Science*, 45(6):e12988.

Shravan Vasishth, Sven Brüssow, Richard L. Lewis, and Heiner Drenhaus. 2008. Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, 32(4):685–712. Publisher: Wiley Online Library.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretič, and Samuel R. Bowman. 2019. Investigating BERT's Knowledge of Language: Five Analysis Methods with NPIs. ArXiv:1909.02597 [cs].

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.

Peter C. Wason and Shuli S. Reich. 1979. A Verbal Illusion. *Quarterly Journal of Experimental Psychology*, 31(4):591–597.

Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. 2021. Frequency Effects on Syntactic Rule Learning in Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 932–948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Leonie Weissweiler, Taiqi He, Naoki Otani, David R. Mortensen, Lori Levin, and Hinrich Schütze. 2023. Construction grammar provides unique insight into neural language models. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, pages 85–95, Washington, D.C. Association for Computational Linguistics.

Alexis Wellwood, Roumyana Pancheva, Valentine Hacquard, and Colin Phillips. 2018. The Anatomy of a Comparative Illusion. *Journal of Semantics*, 35(3):543–583.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN Language Models Learn about Filler-Gap Dependencies? In *Proceedings of BlackboxNLP 2018*. ArXiv: 1809.00042.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

M Xiang, B Dillon, and C Phillips. 2009. Illusory licensing effects across dependency types: ERP evidence. *Brain and Language*, 108(1):40–55.

Yuhan Zhang, Wenqi Chen, Ruihan Zhang, and Xiajie Zhang. 2022. Representing affect information in word embeddings.

Yuhan Zhang, Carina Kauf, and Edward Gibson. 2023a. A noisy-channel explanation of the comparative illusion.

Yuhan Zhang, Carina Kauf, and Edward Gibson. 2023b. a noisy-channel explanation of the comparative illusion. *Architectures and Mechanisms for Language Processing*.

Illusion type	BERT		RoBERTa		GPT-2		GPT-3	
	PPL	Surp	PPL	Surp	PPL	Surp	PPL	Surp
Comparative	**0.43**	-0.07	**0.45**	**-0.22**	**-0.33**	**-0.08**	0.15	-0.04
Depth-charge	**-0.61**	-0.01	-0.20	**0.28**	**-0.41**	-0.01	0.12	**0.90**
NPI	**-0.87**	**0.27**	-0.21	**0.54**	**-0.79**	**0.48**	**-0.70**	**0.41**

Illusion sentences are more acceptable than unacceptable sentences.
The unacceptable sentences are more acceptable than illusion sentences.
No significant difference between the two conditions.

Table 2: Estimates of the main effect (unacceptable sentences vs. illusion sentences) for each statistical model. Positive estimates mean larger perplexity or word surprisals for the unacceptable condition which indicates an illusion effect. Negative estimates mean the unacceptable condition is more acceptable than the illusion condition, which is opposite to the prediction. Bolded estimates represent statistical significance ($p < .05$). We mark the cell in green if there is an illusion effect; in orange for no illusion effect.

COMPARATIVE ILLUSION

Number	**VP**	**Examples**
When the *than*-clause subject is **noun phrase**:		
Singular	Repeatable	More students *have been to Russia* than **the teacher** has.
Singular	Non-repeatable	More students *have escaped from Russia* than **the teacher** has.
Plural	Repeatable	More students *have been to Russia* than **the teachers** have.
Plural	Non-repeatable	More students *have escaped from Russia* than **the teachers** have.
Control	Repeatable	More students *have been to Russia* than **teachers** have. (Good)
Control	Non-repeatable	More students *have escaped from Russia* than **teachers** have. (Good)
When the *than*-clause subject is **pronoun**:		
Singular	Repeatable	More students *have been to Russia* than **I** have.
Singular	Non-repeatable	More students *have escaped from Russia* than **I** have.
Plural	Repeatable	More students *have been to Russia* than **we** have.
Plural	Non-repeatable	More students *have escaped from Russia* than **we** have.
Control	Repeatable	Many students *have been to Russia* more than **I** have. (Good)
Control	Non-repeatable	Many students *have escaped from Russia* more than **I** have. (Bad)

Table 3: Full manipulation for the Comparative illusion

DEPTH CHARGE ILLUSION

Conditions		**Examples**
Canonical depth-charge		No head injury is too trivial to be ignored.
Plausible control		Some head injury is too severe to be ignored.
Implausible control		Some head injury is too trivial to be ignored.
too...to	plausible	No head injury is too trivial to be treated.
too...to	implausible	No head injury is too trivial to be ignored.
too...to not	plausible	No head injury is too trivial to not be ignored.
too...to not	implausible	No head injury is too trivial to not be treated.
so...as to	plausible	No head injury is so trivial as to be ignored.
so...as to	implausible	No head injury is so trivial as to be treated.

Table 4: Full manipulation for the Depth-charge illusion

NPI ILLUSION

Conditions	Examples
Matrix No	**No** hunter who the villager believed to be trustworthy will ever shoot a bear.
Licensor Absent	The hunter who the villager believed to be trustworthy will ever shoot a bear.
Relative No	The hunter who **no** villager believed to be trustworthy will ever shoot a bear.
Relative Didn't	The hunter who **didn't** believe the villager to be trustworthy will ever shoot a bear.
Relative Did not	The hunter who **did not** believe the villager to be trustworthy will ever shoot a bear.
Relative Never	The hunter who **never** believed the villager to be trustworthy will ever shoot a bear.

Table 5: Full manipulation for the NPI illusion

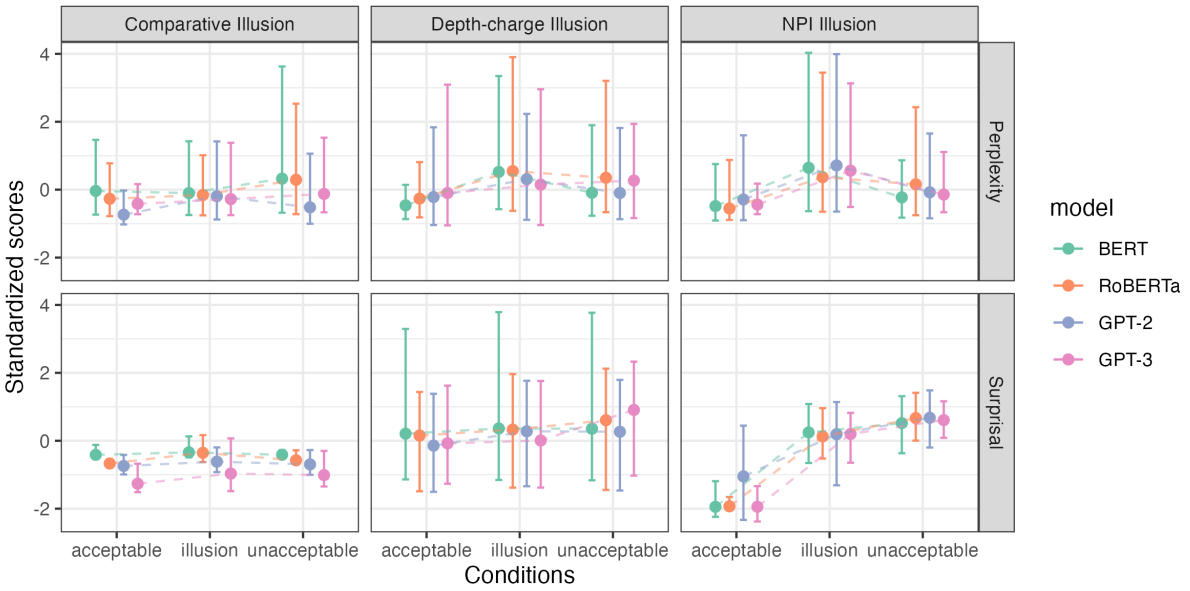

Figure 6: Standardized scores of the Perplexity & Surprisal for sentences in three conditions crossing LMs and language illusion types. If the illusion effect appears, the illusion condition should be rated more acceptable (thus lower in the graph) than the unacceptable condition and therefore has lower perplexity/surprisal. (Error bars are 95% bootstrapped confidence intervals).

View publication stats