(Un)likely words in context: A divergence between humans and large language models

Eun-Kyoung Rosa Lee, Sathvik Nair, Naomi Feldman
University of Maryland, College Park

Human comprehenders sometimes fail to immediately notice (un)likely words given a sentence context. A representative example involves argument roles (*who* did it to *whom*), where people show the same initial responses (e.g., N400, reading times) to verbs that appear in role-appropriate and role-inappropriate contexts. Here, we examined large language models' role-sensitivity in verb predictions, in order to test whether the initial role-blindness observed with humans arises from a prediction process that is mainly driven by distributional information (e.g., co-occurrences of words). Through three experiments, we find that language models show role-sensitive verb predictions, i.e., they can distinguish words that appear in likely and unlikely contexts, but they do not show the same systematic patterns that are observed with humans.

We tested the models' role-sensitivity across three different constructions that elicit specific patterns in human initial predictions. People show the same initial role-blindness when the context contains swapped arguments (*which waitress the customer served*) [1] or when the verb form is changed (*the hearty meal was devouring*) [2]. In contrast, they show immediate sensitivity when an argument is replaced (*which illustrator/readers the author had hired*) [1]. These patterns indicate that the illusion is specific to cases that involve reversing argument role(s), in which changes in the context or the target word yields the same behavior. We tested whether the models exhibited the same patterns across these three conditions (Table 1).

We tested GPT-2 (small, medium, large), BERT (uncased base and large), and RoBERTa (base and large), using English stimulus materials from previous work [1, 2]. In contrast to previous approaches which probe language models' sentence representations [3, 4], we examined representations at the verb, mirroring how humans' role-sensitivity is measured.

Experiment 1 tested models' role-sensitivity by computing the difference of the *surprisal* (negative log probability) of the verb in plausible and implausible contexts. Surprisal is strongly correlated with the N400 response [5]; we expected a difference in verb surprisal between contexts if models were sensitive to argument roles. Results showed positive surprisal effects across all conditions and models, except GPT2-small (Figure 1). The largest effects were found in the *change-verb* condition, not *replace-argument*, i.e., a divergence from human behavior.

Experiment 2 asked whether role-based plausibility information is encoded in the models' vector representations, by training linear classifiers on the models' representations of target verbs in plausible and implausible contexts. We examined representations in each layer to pinpoint the stage of processing when this role-sensitivity arises. Classification accuracy was above chance for all models and increased the most in middle to late layers (Figure 2). However, the *change-verb* condition reached ceiling accuracy based on early layers.

Experiment 3 examined the extent to which the models correctly distinguished subject and object positions. We identified a subject *attention head* [6], which assigned the most attention to the subject when predicting the verb relative to other words, and examined its relative attention to the subject and object. Results showed that the subject heads accurately attended to the subject and not the object in both the *swap-arguments* and *change-verb* conditions (Table 2), suggesting that the difference in role-sensitivity does *not* arise from a difference in the ability to distinguish subjects and objects in the context preceding the verb.

Taken together, our results show that large language models distinguish words that appear in contextually appropriate and inappropriate contexts, but they do not exhibit the same systematic patterns that humans show during real-time prediction. While human comprehenders treat changes in contexts and target words equally, language models show greater sensitivity to changes in target words. This suggests that the human patterns do not naturally arise from prediction based on distributional information alone, and that the models' capacity to detect plausibility does not arise from human-like sentence processing mechanisms.

Table 1. Experiment stimuli. *Target words are in bold font.*

Condition	Plausible	Implausible	N400 effect
swap-arguments	The restaurant owner forgot which *customer* the *waitress* had **served**	The restaurant owner forgot which *waitress* the *customer* had **served**	No
change-verb	The hearty *meal* was **devoured**	The hearty *meal* was **devouring**	No
replace-argument	The secretary confirmed which *illustrator* the *author* had **hired**	The secretary confirmed which *readers* the *author* had **hired**	Yes

Figure 1. Experiment 1 verb surprisal effects (implausible minus plausible).

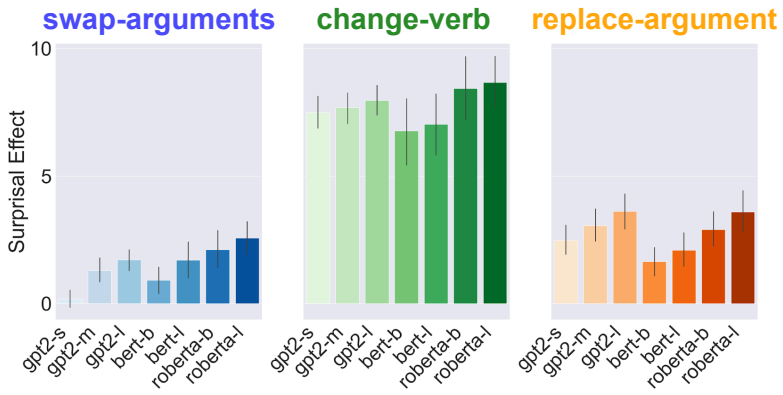

Figure 2. Experiment 2 mean classification accuracies.

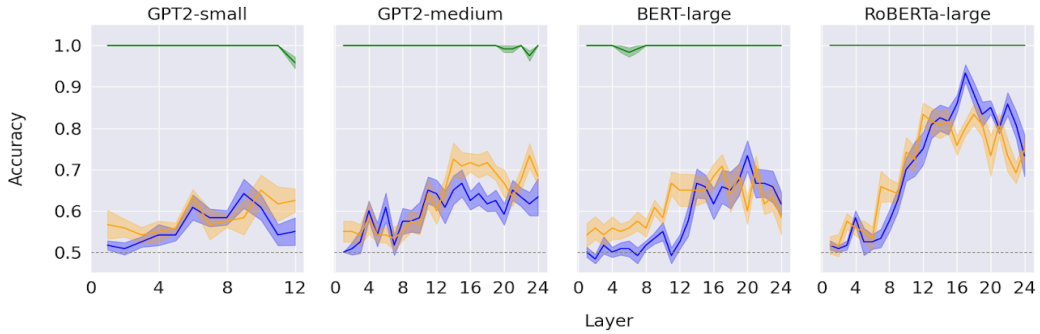

Table 2. Experiment 3 mean attention weights of the subject attention head.

Model	Condition	Attention to Subject		Attention to Object	
		Plausible	Implausible	Plausible	Implausible
GPT2-small	swap-arguments	.53 (.17)	.53 (.17)	.18 (.10)	.19 (.06)
	replace-argument	.51 (.12)	.50 (.13)	.19 (.09)	.21 (.08)
RoBERTa-large	swap-arguments	.68 (.18)	.70 (.20)	.06 (.10)	.05 (.09)
	replace-argument	.65 (.16)	.68 (.16)	.06 (.08)	.04 (.02)

References: [1] Chow et al. (2016). *LCN* [2] Kim & Osterhout (2005). *JML* [3] Papadimitriou et al. (2022). *ACL* [4] Kauf et al. (2023). *Cog. Sci.* [5] Michaelov et al. (2024). *Neurobiology of Language* [6] Ryu & Lewis (2021). *CMCL.*