Error Sensitivity in the Next-Word Predictions of Humans and Language Models

Siyuan Song[1], Thomas Hikaru Clark[2]

Background. According to the noisy-channel processing account, comprehenders rationally infer an intended message from perceived input by weighing both the prior over intended meanings and the likelihood of errors [1, 2]. However, previous research has typically focused on processing at the location of the error (e.g. longer RTs for typos). An open question remains: do comprehenders show sensitivity to possible errors in their expectations about upcoming words? Concretely, following a correctable error (The ferryman *rows* $\rightarrow$ *sows* . . .), are readers less surprised by seeing a word compatible with the *corrected* version of the error (*boat*), relative to when the error is uninformative (*rows* $\rightarrow$ *plants*)? This study considers sentences which vary in how detectable an error is, and measures processing for words downstream of the error. We analyze both humans and large language models (LMs), which are known to have some error-correcting abilities but also have a strongly incremental bias [3, 4]; it is unknown whether LMs will behave similarly to humans, who may employ non-incremental reading to resolve possible errors. Under a **strong account** of error sensitivity, predictions about next words will reflect plausible continuations of not only the veridical context, but error-sensitive alternatives of the veridical context. ***Methods.*** 35 items are constructed with a "[context][subject][verb][object]" structure in a $5 \times 3 \times 3$ design (Table 1) manipulating (i) the verb condition, which included one clean form and four error conditions crossing lexical status (real-word vs. non-word) with error type (orthographic neighbor vs. non-neighbor), and (ii) two bias factors, the preceding context and the subject, each varying across three levels (neutral, biased toward the clean verb, or biased toward the neighbor real-word variant). We conducted an *incremental processing* experiment on 9 LMs by comparing the log probabilities of object log $P($[object]$|$[context][subject][verb]), and a *comprehension question* experiment on 6 instruction-tuned LMs by comparing the relative probability of answering "yes" $\frac{P(\text{"yes"}|Q)}{P(\text{"yes"}|Q)+P(\text{"no"}|Q)}$ after a simple comprehension question Q: "[aux verb] [subject] [verb-clean] [object]?" In the LM experiments, higher object-probability and higher "yes"-probability indicate a more inferential interpretation. We also conducted a mouse-tracking experiment to collect reading measures and responses to the same comprehension questions from $N = 100$ monolingual English speakers on Prolific. We hypothesized that, if predictions (of either a model or human) are error-sensitive, *neighbor* errors should yield more inferential behavior, compared to *non-neighbor* errors with same lexical status. To test this, we fit separate linear regressions predicting each behavioral measure from verb condition, running one model with the non-neighbor real-word as the reference and another with the non-neighbor non-word as the reference. ***Results.*** In the *incremental processing* experiment, focusing on items with neutral contexts and subjects, we do not observe any significant effect on log probability of object of the *neighbor real-word* condition relative to the *non-neighbor real-word* condition, either across LMs or within individual LMs ($p > 0.1$). In contrast, we find a significant positive overall effect when comparing *neighbor non-word* to *non-neighbor non-word* ($b = 0.657, SE = 0.157, p < 0.001$). Similarly, for same items in the *comprehension question* experiment, *neighbor real-word* does not yield a higher relative "yes" probability than *non-neighbor real-word* ($p > 0.1$), whereas *neighbor non-word* is significantly higher than *non-neighbor non-word* overall ($b = 0.252, SE = 0.025, p < 0.001$). In the mouse tracking experiment, we did not find any significant differences in the gaze duration of the object across verb conditions. However, participants showed a significantly higher proportion of "yes" responses to the comprehension questions in the neighbor conditions than in the non-neighbor conditions, for both real-word and non-word. ***Conclusion.*** Our results indicate that LM predictions are only sensitive to non-word errors, without parallel alternatives being considered for temporarily plausible sentence prefixes. Humans, on the other hand, are not sensitive to possible errors at the level of next-word prediction, but show high sensitivity to errors when answering comprehension questions.

[1]Department of Linguistics, University of Texas at Austin
[2]Department of Brain and Cognitive Sciences, MIT
Correspondence to: siyuansong@utexas.edu

[context],	[subject]	[verb]	[object]
/ (no context)	he (neutral)	rows (clean)	the boat
on the river (context 1)	the ferryman (subject 1)	sows (neighbor real-word)	
in the field (context 2)	the farmer (subject 2)	plants (non-neighbor real-word)	
		aows (neighbor non-word)	
		hamnts (non-neighbor non-word)	

Table 1: Example item illustrating different verb, subject, and context conditions. We compared the probability of the object given different prefixes and the relative probability of "yes" given a full sentence and a question (e.g., "Does he row the boat?").

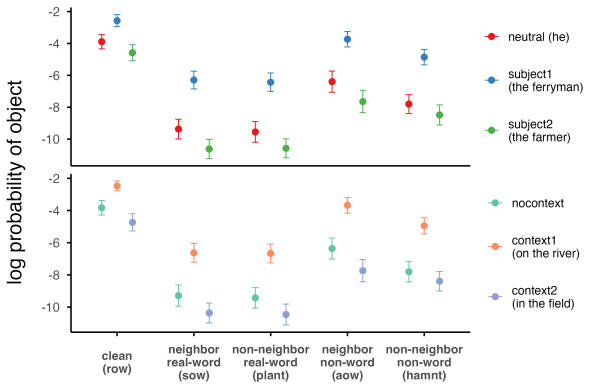

Figure 1: Means and 95% CIs of Gemma 12B log probabilities across subject (top) and context (bottom) conditions. Subject and context manipulations show clear, directionally expected effects on probability. Similar patterns are observed in other LMs.

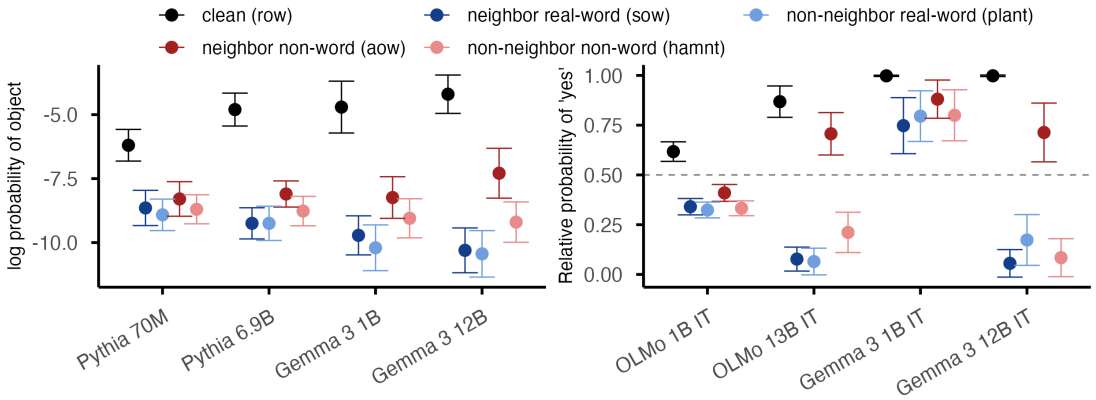

Figure 2: Left: Base models' log probabilities of objects; Right: Instruction-tuned models relative probabilities of answering "yes" to a question "[aux verb] [subject] [verb-clean] [object]?" (i.e. interpretation consistent with clean verb); dashed baseline at 0.5. Points are means; error bars 95% CI.

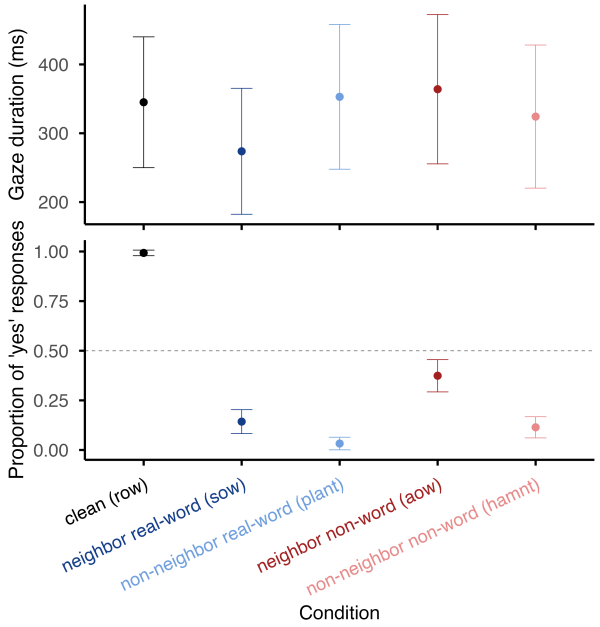

Figure 3: Means and 95% CIs of gaze durations on the object (top) and yes-response proportions (bottom) across verb conditions for neutral-subject, no-context items.

[1] Roger Levy. A noisy-channel model of human sentence comprehension under uncertain input. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 234–243, 2008.

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

[3] Yunxiang Zhang, Liangming Pan, Samson Tan, and Min-Yen Kan. Interpreting the robustness of neural NLP models to textual perturbations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3993–4007, 2022.

[4] Esther Gan, Yiran Zhao, Liying Cheng, Mao Yancan, Anirudh Goyal, Kenji Kawaguchi, Min-Yen Kan, and Michael Shieh. Reasoning robustness of LLMs to adversarial typographical errors. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10449–10459, 2024.