Alignment between human reading and large language models varies across word class

Jonathan R. Brennan*, James Baybaş, Sid Bhushan, Noa Segal, Simón Campos, Xueyang Huang, Sylvan Jesien, Rachel McCullough, Martin Mössmer, George Stain, Elliot Stork, Hsin-Ju Wu, Yueke Zeng & Lisa Levinson

University of Michigan

*jobrenn@umich.edu

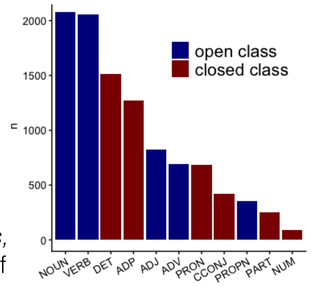
SCAN ME

Introduction

- Familiarly, there are **processing differences between closed- and open-class words** (e.g. Bradley 1978), such as empirical asymmetries in how expectations affect reading times and ERPs across word category (e.g. Roark et al. 2009, Brennan & Hale 2019)
- Oh and Schuler (2023) test how **predictability from large-language models (LLMs) of different sizes** fit with human reading times, and find that **alignment is modulated by grammatical categories** (e.g. nouns and adjectives may be under-predicted by larger models.)
- Building on this, we
  - **(1) test the correlation between LLM surprisal and reading times across part-of-speech (POS)** and
  - **(2) evaluate reading time alignment across LLMs of different sizes** to test whether larger models might be better able to capture certain word classes than others
- These comparisons offer a benchmark for evaluating theories of prediction in reading that incorporate grammatical features.

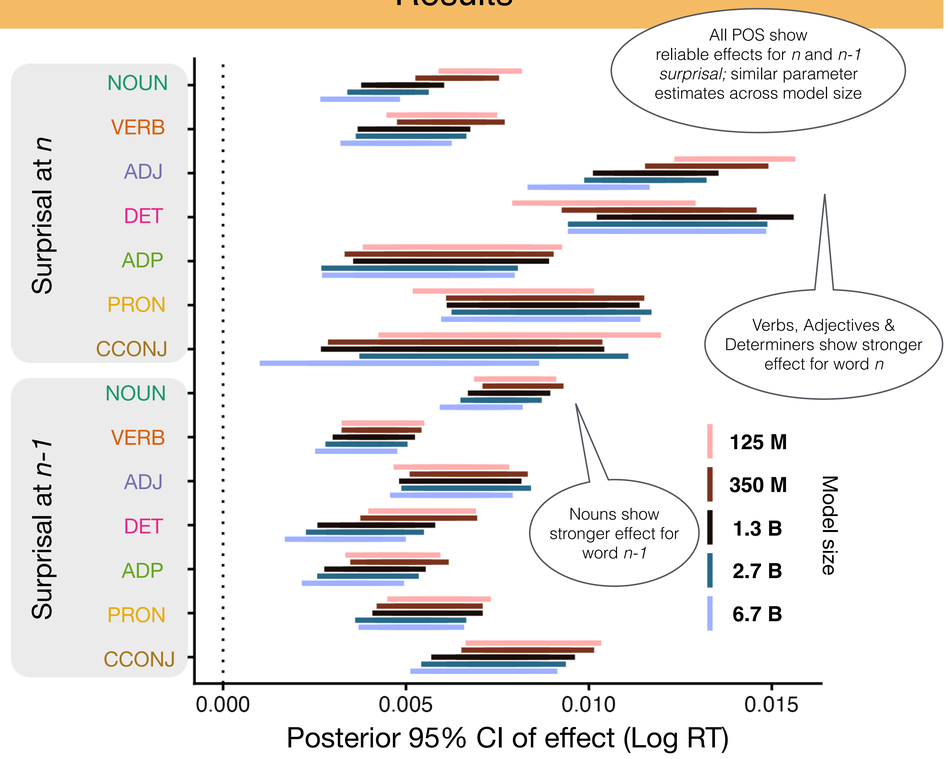
Methods

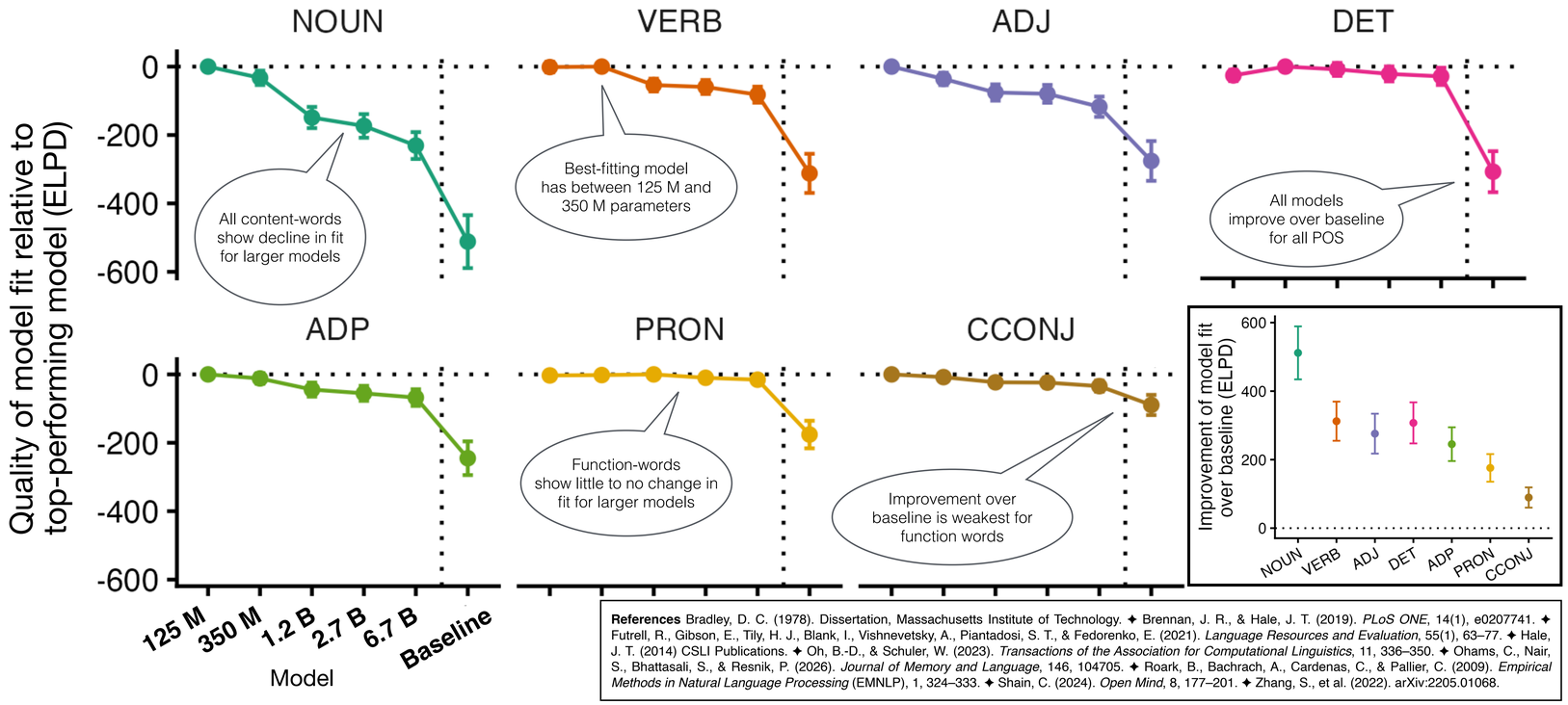
- N=80 datasets from *Natural Stories* self-paced reading-time corpus (Futrell et al. 2021; 1k to 10k words each of English short stories, online data collection.)
- Reading times (RTs) for **open class** (*Nouns, Verbs, Adjectives)* and **closed class** (*Determiners, pronouns, adpositions, conjunctions*). Number of tokens per POS shown on right.



- Predictability via **surprisal** from the pre-trained OPT LLM spanning 125 million to 6.7 billion parameters (Zhang et al. 2022)
- Fit between surprisal and RT evaluated using Bayesian hierarchical linear regression; maximal random effects and fixed effects of *word length*, *POS*, and *surprisal* for current word (*n*) and previous word (*n-1*), as well as all interactions.
  - Including ***n*-1 surprisal** captures well-documented spill-over effects (e.g. Shain 2024) that were not included in prior work testing LLMs of different sizes.
  - Target models compared against **baseline** with control covariates but no surprisal predictors.
- Statistical goodness-of-fit quantified via *Expected Log Pointwise Predictive Density* (ELPD; Vehtari 2017) across LLMs separately for each word category.

Results





Conclusions

- Declining performance for models with >350 M parameters (as found by Oh & Schuler 2023), strongest for open-class words.
- LLM surprisal captures predictability differently across parts-of-speech.
- Consistent with theoretical accounts where lexical predictions reflect an inference process that incorporates syntax (e.g. Hale 2014, Eddine et al. 2024, Ohams et al. 2026).
- **Methodological take-away**: If using LLMs to estimate predictability, consider the word classes in target regions carefully.
- **Theoretical take-away:** Theories of prediction in comprehension must consider empirical asymmetries between word classes.

References Bradley, D. C. (1978). Dissertation, Massachusetts Institute of Technology. ✦ Brennan, J. R., & Hale, J. T. (2019). *PLoS ONE*, 14(1), e0207741. ✦ Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S. T., & Fedorenko, E. (2021). *Language Resources and Evaluation*, 55(1), 63–77. ✦ Hale, J. T. (2014) CSLI Publications. ✦ Oh, B.-D., & Schuler, W. (2023). *Transactions of the Association for Computational Linguistics*, 11, 336–350. ✦ Ohams, C., Nair, S., Bhattasali, S., & Resnik, P. (2026). *Journal of Memory and Language*, 146, 104705. ✦ Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009). *Empirical Methods in Natural Language Processing* (EMNLP), 1, 324–333. ✦ Shain, C. (2024). *Open Mind*, 8, 177–201. ✦ Zhang, S., et al. (2022). arXiv:2205.01068.