Frequency Explains the Inverse Correlation of Large Language Models' Size, Training Data Amount, and Surprisal's Fit to Reading Times

Byung-Doh Oh[1] Shisen Yue[2] William Schuler[1]

[1]The Ohio State University
[2]Shanghai Jiao Tong University

EACL 2024

The Ohio State University

上海交通大學
Shanghai Jiao Tong University

Prediction in human sentence processing

cake

The boy will eat the

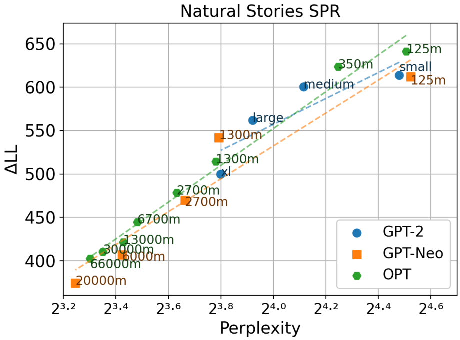
ball

- *cake* is easier to process than *ball* because $\mathrm{P}(\textit{cake} \mid \ldots) > \mathrm{P}(\textit{ball} \mid \ldots)$ (Hale, 2001; Levy, 2008)
- Surprisal has gained strong empirical support from measures of comprehension difficulty (e.g. Demberg & Keller, 2008; Shain et al., 2020; Smith & Levy, 2013)
- Research goal of characterizing the probability distribution of the human comprehender

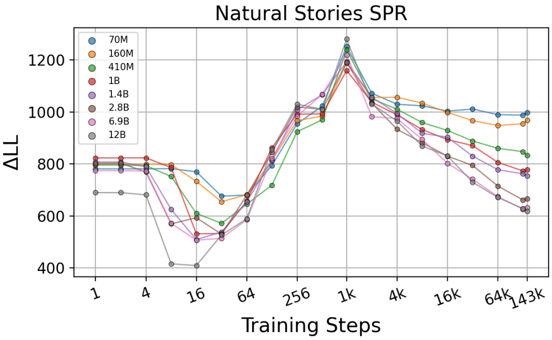
Systematic divergence of Transformer-based LM surprisal

Better Fit

Poorer Fit



Oh and Schuler (2023a)



Oh and Schuler (2023b)

- How does model size and training data interact to result in such divergence?

Insights from the scaling behavior of LLMs

- Larger models ‘learn faster’ given the same amount of exposure (Tirumala et al., 2022)
- Early in training, all models similarly learn to predict frequent function words (Xia et al., 2023)

Word frequency modulates the difference in surprisal estimates as a function of model size and training data amount, which drives their adverse effects on fit to human reading times.

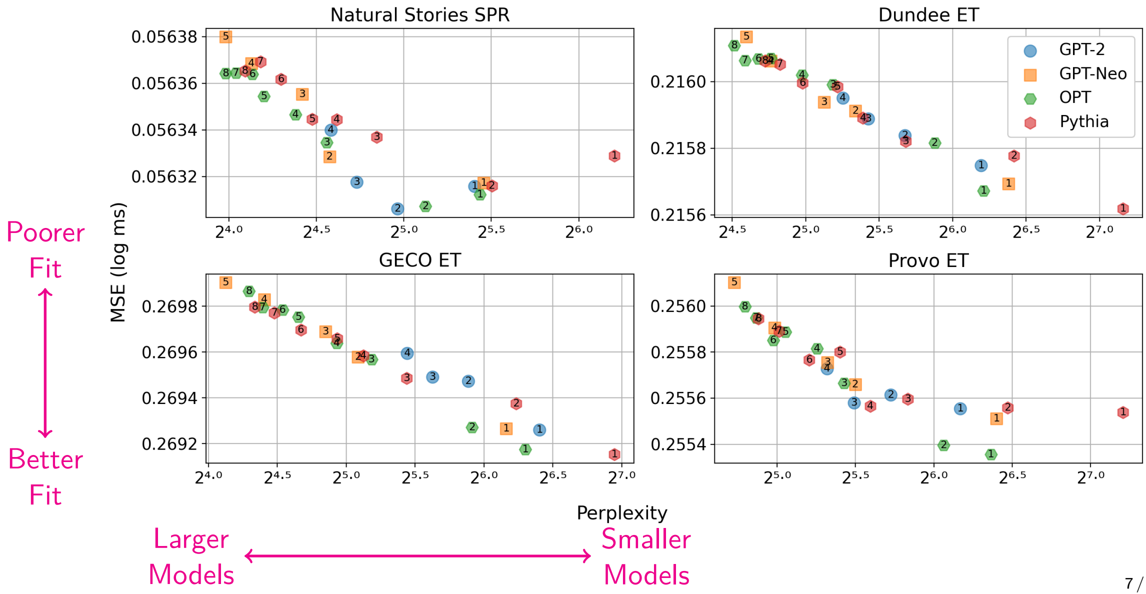
This work

- Experiment 1: Word frequency and adverse effect of model size
- Experiment 2: Word frequency and adverse effect of training data amount
- Follow-up analysis: What enables larger models to predict rare words?
- Discussion and conclusion

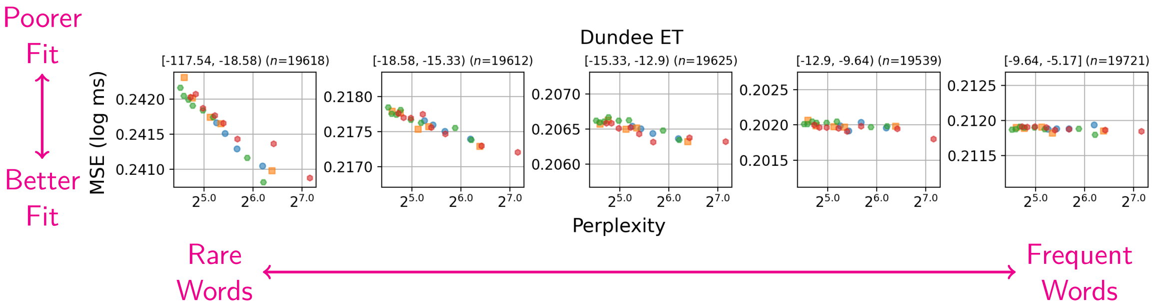
Experiment 1: Word frequency and adverse effect of model size

- LME models fit to reading times of Natural Stories, Dundee, Ghent, and Provo corpora (Cop et al., 2017; Futrell et al., 2021; Kennedy et al., 2003; Luke & Christianson, 2018)
- Baseline predictors: Word length/position, unigram surprisal (tokens from Gao et al., 2020), saccade length, previous word fixated
- Predictors of interest: GPT-2, GPT-Neo, OPT, Pythia surprisal (Biderman et al., 2023; Black et al., 2022; Black et al., 2021; Radford et al., 2019; Wang & Komatsuzaki, 2021; Zhang et al., 2022)
- Mean squared errors calculated on each quintile defined by unigram log-probability

Larger models yield poorer fits to reading times



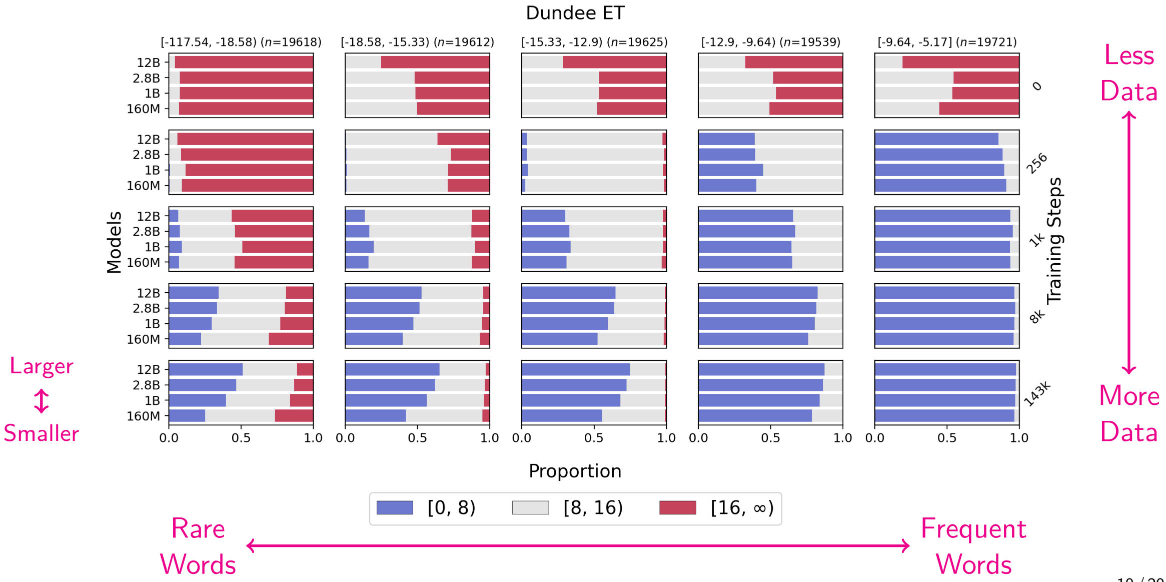
These effects are the largest on rare words



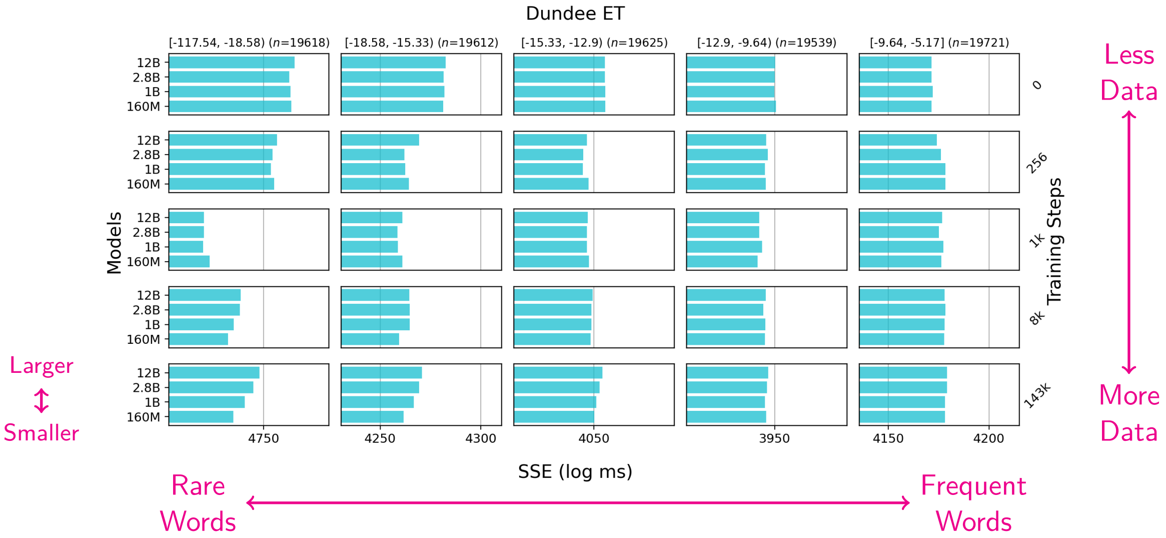
Experiment 2: Word frequency and adverse effect of training data amount

- Similar regression modeling procedures as Experiment 1
- Predictors of interest: Pythia surprisal after {0, 128, 256, 512, 1k, 2k, 4k, 8k, 143k} training steps (Biderman et al., 2023)
- Surprisal values and MSEs analyzed by quintile defined by unigram log-probability

Rare words are learned more accurately by larger models with more data



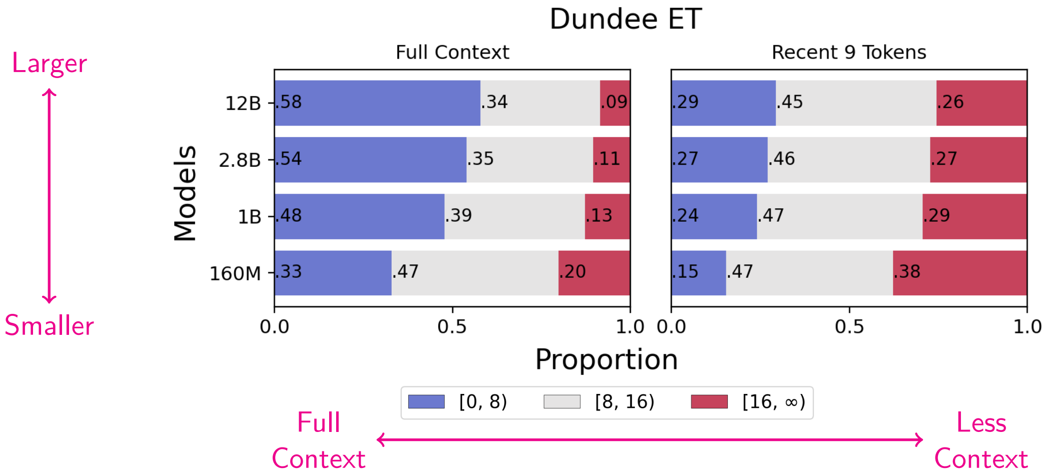
Rare words are learned more accurately by larger models with more data



Follow-up analysis: What enables larger models to predict rare words?

- One possibility is that larger models have a longer 'effective' context window
- We examine this possibility through a feature attribution analysis
- Method: Limiting the context to the most recent $\{49, 24, 9\}$ tokens (Kuribayashi et al., 2022)
- Change in Pythia surprisal values analyzed on the quintile of least frequent words

Larger models have widespread associations for predicting rare words



Discussion and conclusion

- Word frequency explains the adverse effects of model size and training data amount
- Larger model and training data sizes contribute to accurate predictions of rare words
- This has implications for studying the dissociability of frequency vs. predictability effects (Goodkind & Bicknell, 2021; Shain, 2019, 2023)
- Possible extension to data collected in other languages (de Varda & Marelli, 2023; Kuribayashi et al., 2021; Wilcox et al., 2023)

Thank you for listening!

oh.531@osu.edu byungdoh.github.io

byungdoh/llm_surprisal

References I

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., & van der Wal, O. (2023). Pythia: A suite for analyzing large language models across training and scaling. *Proceedings of the 40th International Conference on Machine Learning*, *202*, 2397–2430. https://proceedings.mlr.press/v202/biderman23a.html

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., Pieler, M., Prashanth, U. S., Purohit, S., Reynolds, L., Tow, J., Wang, B., & Weinbach, S. (2022). GPT-NeoX-20B: An open-source autoregressive language model. *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, 95–136. https://aclanthology.org/2022.bigscience-1.9

Black, S., Gao, L., Wang, P., Leahy, C., & Biderman, S. (2021). GPT-Neo: Large scale autoregressive language modeling with Mesh-Tensorflow. *Zenodo*. https://doi.org/10.5281/zenodo.5297715

Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, *49*(2), 602–615. https://doi.org/10.3758/s13428-016-0734-0

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210. https://doi.org/10.1016/j.cognition.2008.07.008

References II

de Varda, A., & Marelli, M. (2023). Scaling in cognitive modelling: A multilingual approach to human reading times. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 139–149. https://aclanthology.org/2023.acl-short.14

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S., & Fedorenko, E. (2021). The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, *55*, 63–77. https://doi.org/10.1007/s10579-020-09503-7

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., & Leahy, C. (2020). The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint, arXiv:2101.00027*. https://arxiv.org/abs/2101.00027

Goodkind, A., & Bicknell, K. (2021). Local word statistics affect reading times independently of surprisal. *arXiv preprint, arXiv:2103.04469v2*. https://arxiv.org/abs/2103.04469

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, 1–8. https://www.aclweb.org/anthology/N01-1021/

Kennedy, A., Hill, R., & Pynte, J. (2003). The Dundee Corpus. *Proceedings of the 12th European Conference on Eye Movement*.

References III

Kuribayashi, T., Oseki, Y., Brassard, A., & Inui, K. (2022). Context limitations make neural language models more human-like. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 10421–10436. https://aclanthology.org/2022.emnlp-main.712

Kuribayashi, T., Oseki, Y., Ito, T., Yoshida, R., Asahara, M., & Inui, K. (2021). Lower perplexity is not always human-like. *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 5203–5217. https://aclanthology.org/2021.acl-long.405

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Luke, S. G., & Christianson, K. (2018). The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, *50*(2), 826–833. https://doi.org/10.3758/s13428-017-0908-4

Oh, B.-D., & Schuler, W. (2023a). Why does surprisal from larger Transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, *11*, 336–350. https://doi.org/10.1162/tacl_a_00548

Oh, B.-D., & Schuler, W. (2023b). Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1915–1921. https://aclanthology.org/2023.findings-emnlp.128/

References IV

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Technical Report*. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Shain, C. (2019). A large-scale study of the effects of word frequency and predictability in naturalistic reading. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4086–4094. https://aclanthology.org/N19-1413/

Shain, C. (2023). Word frequency and predictability dissociate in naturalistic reading. *PsyArXiv preprint*. https://osf.io/preprints/psyarxiv/9zdfw/

Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, *138*, 107307. https://doi.org/https://doi.org/10.1016/j.neuropsychologia.2019.107307

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*, 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Tirumala, K., Markosyan, A., Zettlemoyer, L., & Aghajanyan, A. (2022). Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, *35*, 38274–38290. https://proceedings.neurips.cc/paper_files/paper/2022/file/fa0509f4dab6807e2cb465715bf2d249-Paper-Conference.pdf

References V

Wang, B., & Komatsuzaki, A. (2021). GPT-J-6B: A 6 billion parameter autoregressive language model. https://github.com/kingoflolz/mesh-transformer-jax

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, *11*, 1451–1470. https://doi.org/10.1162/tacl_a_00612

Xia, M., Artetxe, M., Zhou, C., Lin, X. V., Pasunuru, R., Chen, D., Zettlemoyer, L., & Stoyanov, V. (2023). Training trajectories of language models across scales. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 13711–13738. https://aclanthology.org/2023.acl-long.767

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., & Zettlemoyer, L. (2022). OPT: Open pre-trained Transformer language models. *arXiv preprint*, *arXiv:2205.01068v4*. https://arxiv.org/abs/2205.01068