Transformer-Based Language Model Surprisal Predicts Human Reading Times Best with About Two Billion Training Tokens

Byung-Doh Oh   William Schuler

Department of Linguistics
The Ohio State University

Findings of the ACL: EMNLP 2023

The Ohio State University

Prediction in human sentence processing

cake

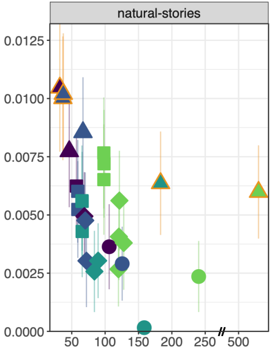
The boy will eat the

ball

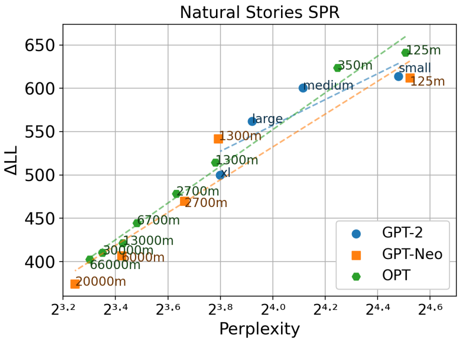
- *cake* is easier to process than *ball* because $\mathrm{P}(\textit{cake} \mid \ldots) > \mathrm{P}(\textit{ball} \mid \ldots)$ (Hale, 2001; Levy, 2008)
- Surprisal has gained strong empirical support from measures of comprehension difficulty (e.g. Demberg & Keller, 2008; Shain et al., 2020; Smith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

This work

- Conflicting results about the relationship between LM perplexity and fit to reading times



Wilcox et al. (2020)



Oh and Schuler (2023)

- Covering the middle ground by evaluating *smaller models* trained on *less data*

Experiment 1: Influence of training data size

- Regression models fit to reading times of Natural Stories and Dundee corpora

  (Futrell et al., 2021; Kennedy et al., 2003)

- Baseline predictors: word length/position, saccade length, previous word fixated

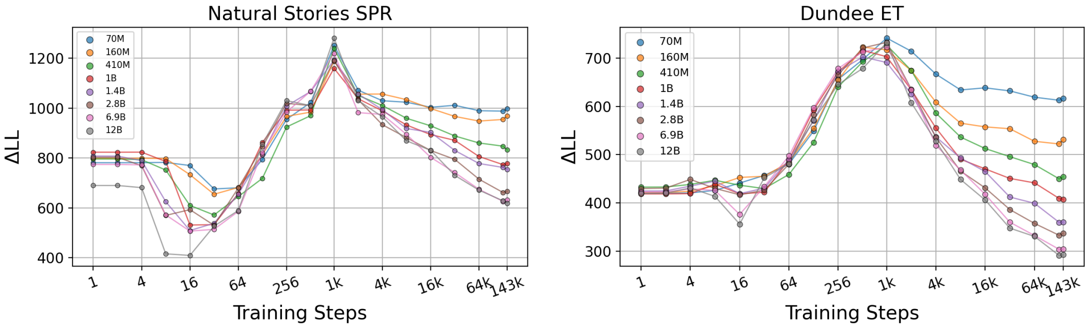
- Predictors of interest: LLM surprisal

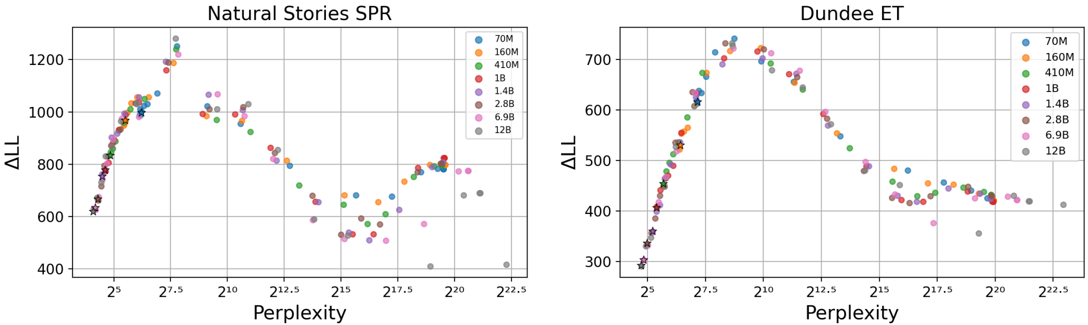
  (Biderman et al., 2023)

- Evaluation metric: Δlog-likelihood (ΔLL)

Model	#L	#H	d_{model}
Pythia 70M	6	8	512
Pythia 160M	12	12	768
Pythia 410M	24	16	1024
Pythia 1B	16	8	2048
Pythia 1.4B	24	16	2048
Pythia 2.8B	32	32	2560
Pythia 6.9B	32	32	4096
Pythia 12B	36	40	5120

- Trained in batches of 1024×2048 tokens
- Checkpoints available after $\{1, 2, 4, \ldots, 512, 1000, 2000, \ldots, 142000, 143000\}$ training steps


Natural Stories SPR
Dundee ET
ΔLL
Training Steps
70M
160M
410M
1B
1.4B
2.8B
6.9B
12B
1
4
16
64
256
1k
4k
16k
64k
143k



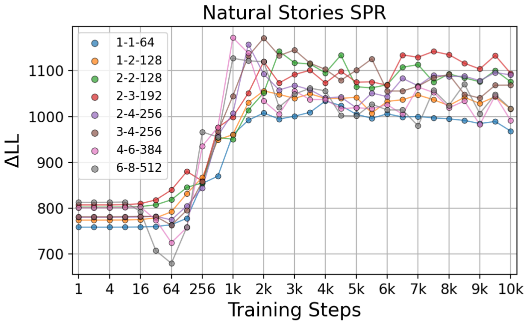
Natural Stories SPR
ΔLL
Perplexity
1200
1000
800
600
400
2^5
$2^{7.5}$
2^{10}
$2^{12.5}$
2^{15}
$2^{17.5}$
2^{20}
$2^{22.5}$
70M
160M
410M
1B
1.4B
2.8B
6.9B
12B
Dundee ET
ΔLL
Perplexity
700
600
500
400
300
2^5
$2^{7.5}$
2^{10}
$2^{12.5}$
2^{15}
$2^{17.5}$
2^{20}
$2^{22.5}$
70M
160M
410M
1B
1.4B
2.8B
6.9B
12B

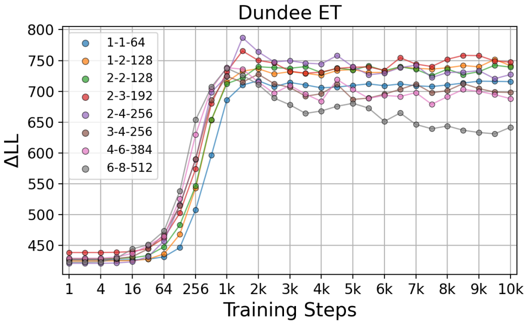
Experiment 2: Influence of model size

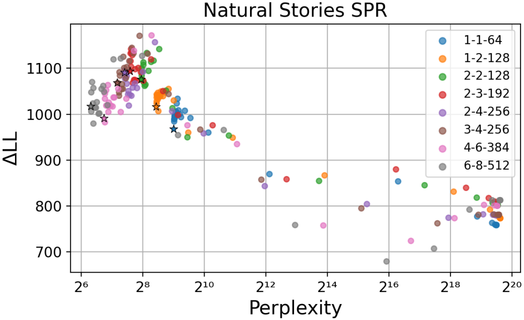
- Smaller LMs trained following the procedures of the Pythia LM

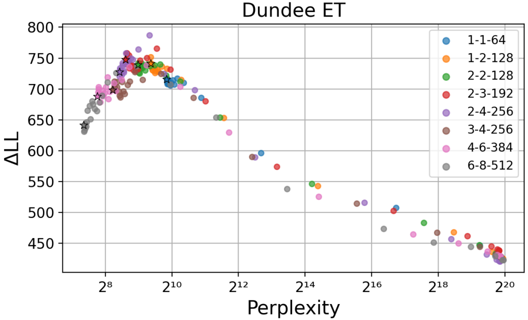
Model	#L	#H	d_{model}	#Parameters
Repro 1-1-64	1	1	64	~6M
Repro 1-2-128	1	2	128	~13M
Repro 2-2-128	2	2	128	~13M
Repro 2-3-192	2	3	192	~20M
Repro 2-4-256	2	4	256	~27M
Repro 3-4-256	3	4	256	~28M
Repro 4-6-384	4	6	384	~46M
Repro 6-8-512	6	8	512	~70M

- LMs evaluated after {1, 2, 4, ..., 512, 1000, 1500, ..., 9500, 10000} training steps


Natural Stories SPR
ΔLL
Training Steps
1-1-64
1-2-128
2-2-128
2-3-192
2-4-256
3-4-256
4-6-384
6-8-512


Dundee ET
ΔLL
Training Steps
1-1-64
1-2-128
2-2-128
2-3-192
2-4-256
3-4-256
4-6-384
6-8-512


Natural Stories SPR
ΔLL
Perplexity
1100
1000
900
800
700
2⁶
2⁸
2¹⁰
2¹²
2¹⁴
2¹⁶
2¹⁸
2²⁰
1-1-64
1-2-128
2-2-128
2-3-192
2-4-256
3-4-256
4-6-384
6-8-512


Dundee ET
ΔLL
Perplexity
800
750
700
650
600
550
500
450
2⁸
2¹⁰
2¹²
2¹⁴
2¹⁶
2¹⁸
2²⁰
1-1-64
1-2-128
2-2-128
2-3-192
2-4-256
3-4-256
4-6-384
6-8-512

Summary: Bigger-is-worse effect of training data

- Fit to reading times starts to degrade after about two billion tokens of training data
- Very strong interaction between model size and amount of training data
- Consolidates conflicting results about LM perplexity and fit to reading times
- This systematic divergence sheds light on what human sentence processing is not

Thank you for listening!

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

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

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

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

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*.

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

References II

Oh, B.-D., & Schuler, W. (2023). 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

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

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. P. (2020). On the predictive power of neural language models for human real-time comprehension behavior. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, 1707–1713. https://cognitivesciencesociety.org/cogsci20/papers/0375