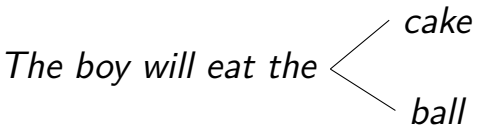
The Bigger-is-Worse Effects of Model Size and Training Data of Large Language Model Surprisal on Human Reading Times

Byung-Doh Oh

Department of Linguistics
The Ohio State University

November 9, 2023
New York University

Prediction in human sentence processing



- *cake* is easier to process than *ball* because $P(\textit{cake} \mid ...) > P(\textit{ball} \mid ...)$ (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)

Prediction in human sentence processing

- Open questions about the probability distribution of the human comprehender
- Computational modeling helps us understand what this distribution is (or is not)
- This talk will highlight the *systematic divergence* of large language models (LLMs)

Roadmap

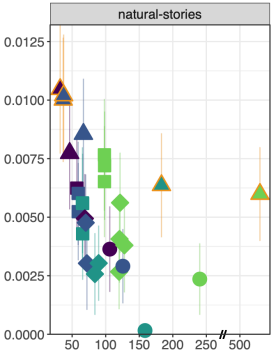
1. Phenomenon #1: The bigger-is-worse effect of model size (Oh & Schuler, 2023a)
2. Phenomenon #2: The bigger-is-worse effect of training data (Oh & Schuler, 2023b)
3. Preliminary analyses towards a unified explanation
4. Conclusion

Phenomenon #1: The bigger-is-worse effect of model size

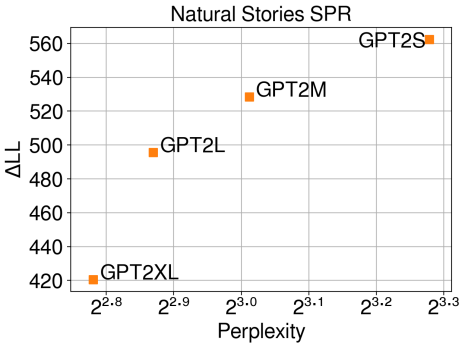
Oh and Schuler (2023a). Why does surprisal from larger Transformer-based language models provide a poorer fit to human reading times? *TACL*.

Introduction

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



Wilcox et al. (2020)



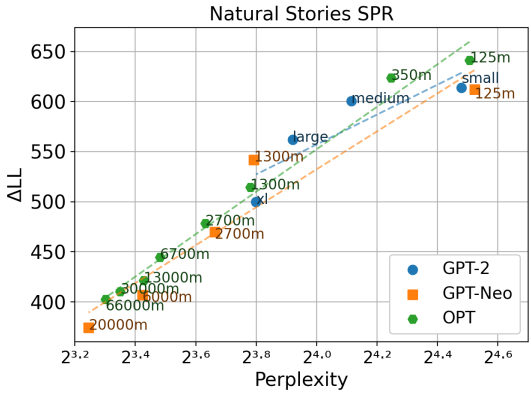
Oh et al. (2022)

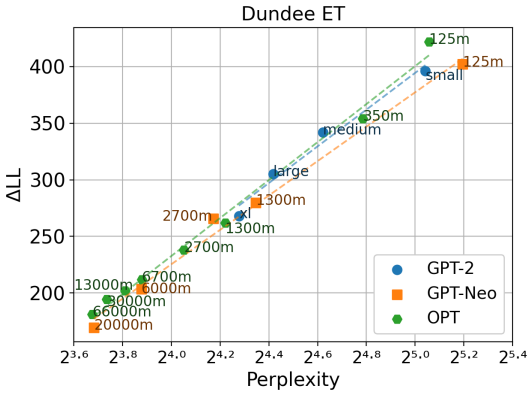
- Qualitative analysis into the trend observed for LLM surprisal

Main experiment: Evaluation of LLM surprisal on human reading times

- 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
  (Black et al., 2022; Black et al., 2021; Radford et al., 2019; Wang & Komatsuzaki, 2021; Zhang et al., 2022)
- Evaluation metric: Δlog-likelihood (ΔLL)

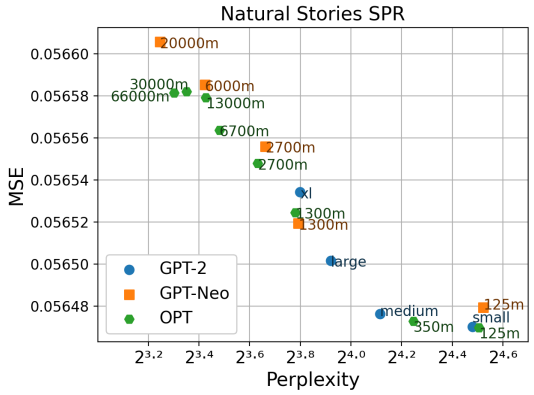
Model	#L	#H	d_{model}
GPT-2 Small	12	12	768
GPT-2 Medium	24	16	1024
GPT-2 Large	36	20	1280
GPT-2 XL	48	25	1600
GPT-Neo 125M	12	12	768
GPT-Neo 1.3B	24	16	2048
GPT-Neo 2.7B	32	20	2560
GPT-J 6B	28	16	4096
GPT-NeoX 20B	44	64	6144
OPT 125M	12	12	768
OPT 350M	24	16	1024
OPT 1.3B	24	32	2048
OPT 2.7B	32	32	2560
OPT 6.7B	32	32	4096
OPT 13B	40	40	5120
OPT 30B	48	56	7168
OPT 66B	64	72	9216

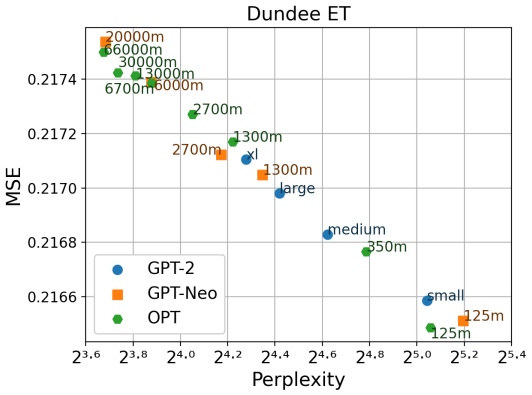

Natural Stories SPR
ΔLL
Perplexity
GPT-2
GPT-Neo
OPT
125m
small
350m
medium
125m
large
1300m
1300m
xl
2700m
2700m
6700m
13000m
30000m
30000m
66000m
20000m

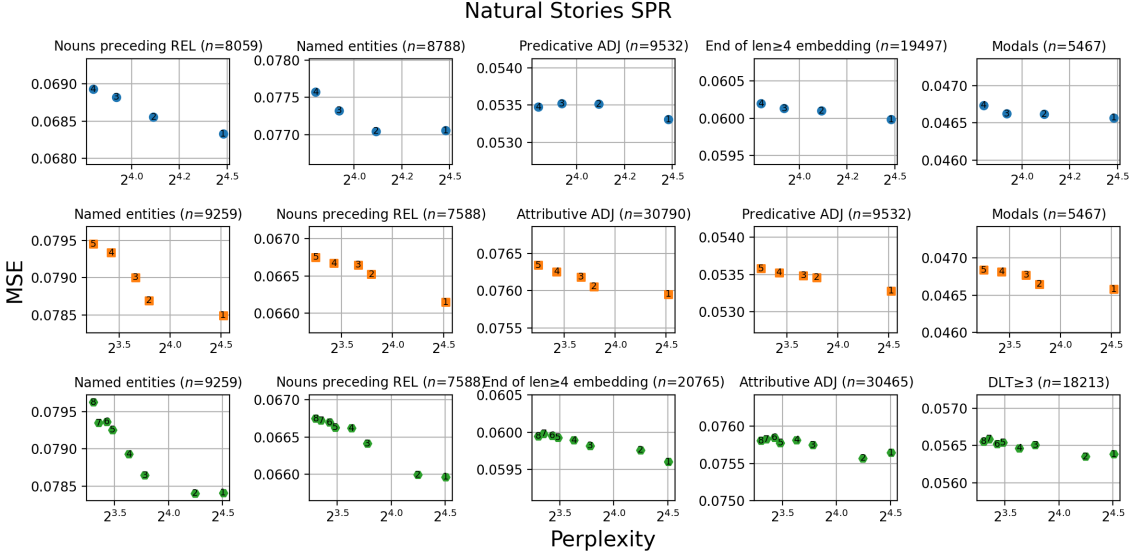

Dundee ET
ΔLL
Perplexity
GPT-2
GPT-Neo
OPT
125m
125m
small
350m
medium
large
2700m
1300m
xl
1300m
2700m
13000m
6700m
30000m
66000m
20000m

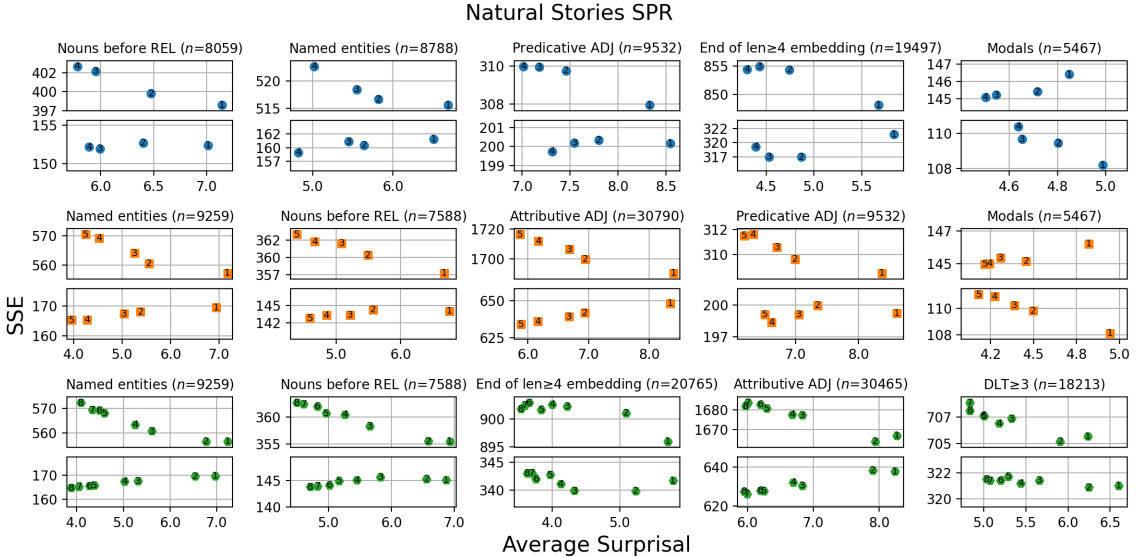
Analysis: Linguistic phenomena underlying the trend

- Data points associated with word-level and syntactic properties (Shain et al., 2018)
- Subsets with the largest differences in MSE between models identified






Natural Stories SPR
Nouns preceding REL (n=8059)
Named entities (n=8788)
Predicative ADJ (n=9532)
End of len≥4 embedding (n=19497)
Modals (n=5467)
Named entities (n=9259)
Nouns preceding REL (n=7588)
Attributive ADJ (n=30790)
Predicative ADJ (n=9532)
Modals (n=5467)
Named entities (n=9259)
Nouns preceding REL (n=7588)
End of len≥4 embedding (n=20765)
Attributive ADJ (n=30465)
DLT≥3 (n=18213)
MSE
Perplexity


Natural Stories SPR
Nouns before REL (n=8059)
Named entities (n=8788)
Predicative ADJ (n=9532)
End of len≥4 embedding (n=19497)
Modals (n=5467)
Named entities (n=9259)
Nouns before REL (n=7588)
Attributive ADJ (n=30790)
Predicative ADJ (n=9532)
Modals (n=5467)
Named entities (n=9259)
Nouns before REL (n=7588)
End of len≥4 embedding (n=20765)
Attributive ADJ (n=30465)
DLT≥3 (n=18213)
SSE
Average Surprisal

Summary: Bigger-is-worse effect of model size

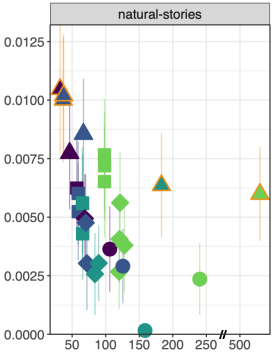
- Strictly *monotonic, positive* relationship between LM perplexity and fit to reading times
- Effect mostly driven by underprediction of reading times by LLM surprisal
  (see e.g. Arehalli et al., 2022; Hahn et al., 2022; van Schijndel & Linzen, 2021)
- Likely due to extensive domain knowledge from massive amounts of training examples

Phenomenon #2: The bigger-is-worse effect of training data

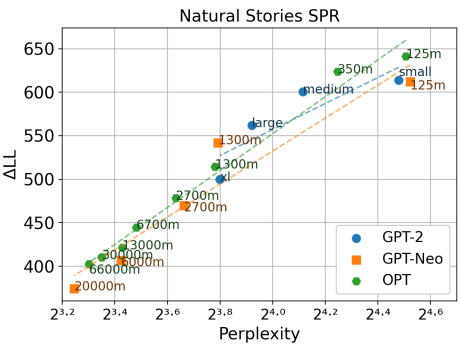
Oh and Schuler (2023b). Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. *Findings of the ACL: EMNLP 2023*.

Introduction

- (Still) conflicting results about LM perplexity and fit to reading times



Wilcox et al. (2020)



Oh and Schuler (2023a)

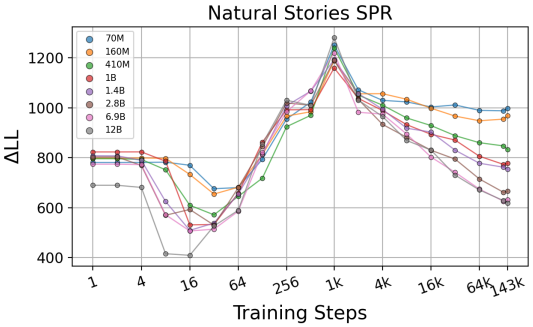
- Covering the middle ground by evaluating *smaller models* trained on *less data*

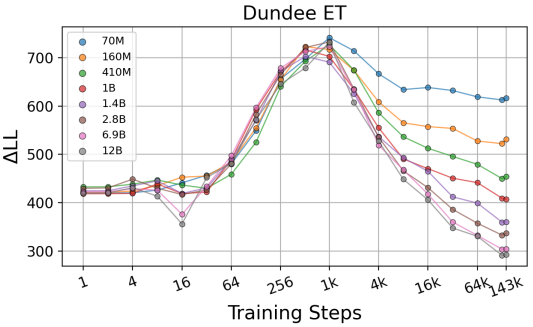
Experiment 1: Influence of training data size

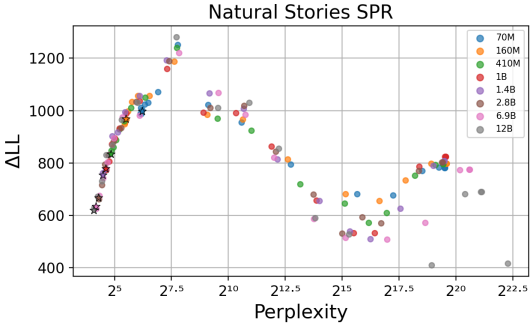
- Regression models fit to reading times of Natural Stories and Dundee corpora
  (Futrell et al., 2021; Kennedy et al., 2003)
- Baseline predictors: word length/position, saccade length, previous word fixated
- Predictors of interest: LLM surprisal
  (Biderman et al., 2023)
- Evaluation metric: Δlog-likelihood (ΔLL)

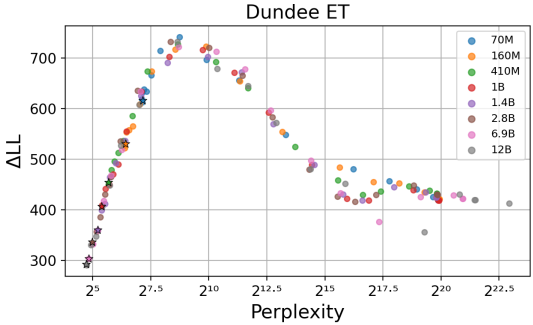
Model	#L	#H	d_{model}
Pythia 70M	6	8	512
Pythia 160M	12	12	768
Pythia 410M	24	16	1024
Pythia 1B	16	8	2048
Pythia 1.4B	24	16	2048
Pythia 2.8B	32	32	2560
Pythia 6.9B	32	32	4096
Pythia 12B	36	40	5120

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


Natural Stories SPR
ΔLL
Training Steps
70M
160M
410M
1B
1.4B
2.8B
6.9B
12B
1
4
16
64
256
1k
4k
16k
64k
143k
400
600
800
1000
1200


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
300
400
500
600
700


Natural Stories SPR
ΔLL
Perplexity
70M
160M
410M
1B
1.4B
2.8B
6.9B
12B

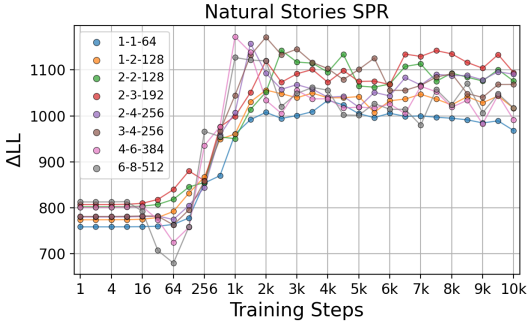

Dundee ET
ΔLL
Perplexity
70M
160M
410M
1B
1.4B
2.8B
6.9B
12B

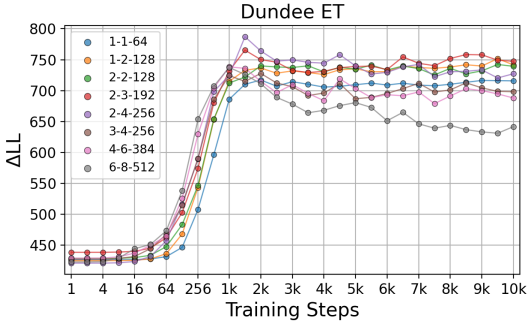
Experiment 2: Influence of model size

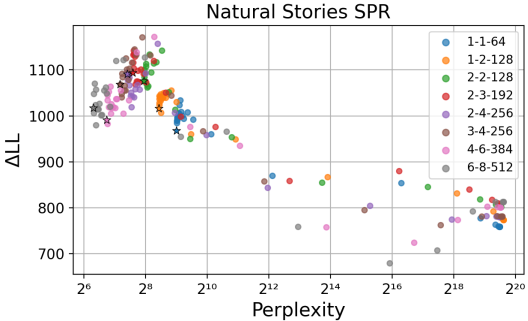
- Smaller LMs trained following the procedures of the Pythia LM

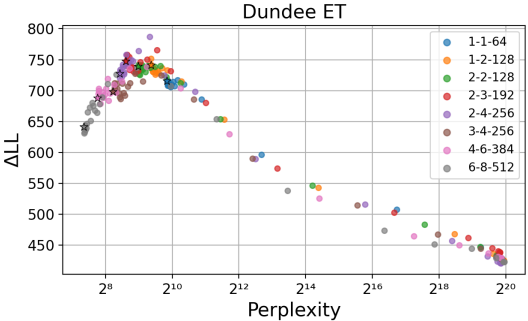
Model	#L	#H	d_{model}	#Parameters
Repro 1-1-64	1	1	64	∼6M
Repro 1-2-128	1	2	128	∼13M
Repro 2-2-128	2	2	128	∼13M
Repro 2-3-192	2	3	192	∼20M
Repro 2-4-256	2	4	256	∼27M
Repro 3-4-256	3	4	256	∼28M
Repro 4-6-384	4	6	384	∼46M
Repro 6-8-512	6	8	512	∼70M

- 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
700
800
900
1000
1100
1
4
16
64
256
1k
2k
3k
4k
5k
6k
7k
8k
9k
10k


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
450
500
550
600
650
700
750
800
1
4
16
64
256
1k
2k
3k
4k
5k
6k
7k
8k
9k
10k


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

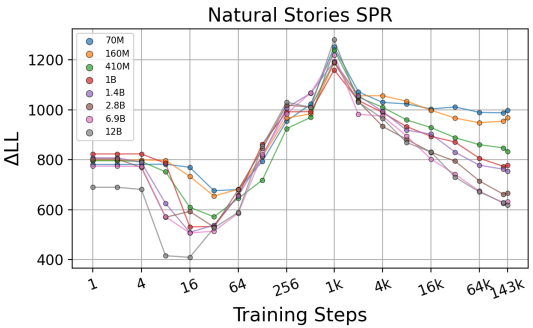

Dundee ET
ΔLL
Perplexity
1-1-64
1-2-128
2-2-128
2-3-192
2-4-256
3-4-256
4-6-384
6-8-512

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

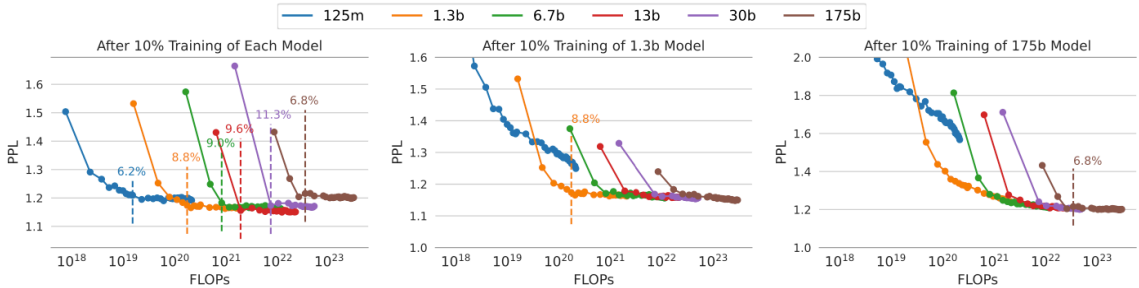
Preliminary analyses towards a unified explanation

Two open questions



- What do all LMs learn similarly during early training?
- What do larger LMs learn differently after early training?

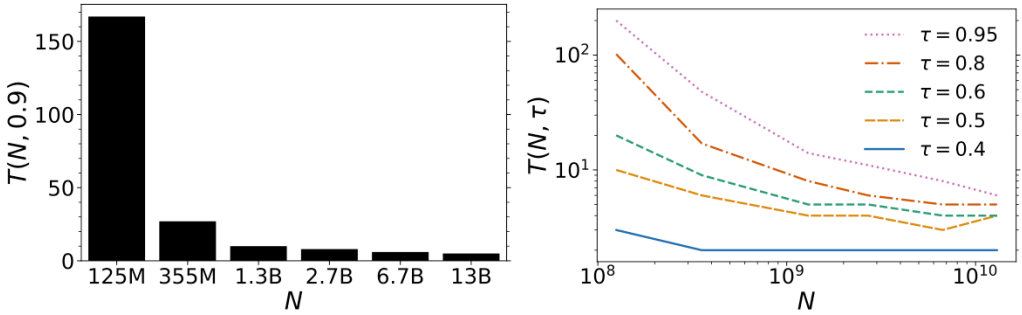
Function words seem to be learned early



"On manual inspection, stagnated tokens are *primarily non-content words such as prepositions, determiners, and punctuations.*"

Xia et al. (2023). Training trajectories of language models across scales. In *Proc. ACL*.

Larger LMs learn quicker with limited examples



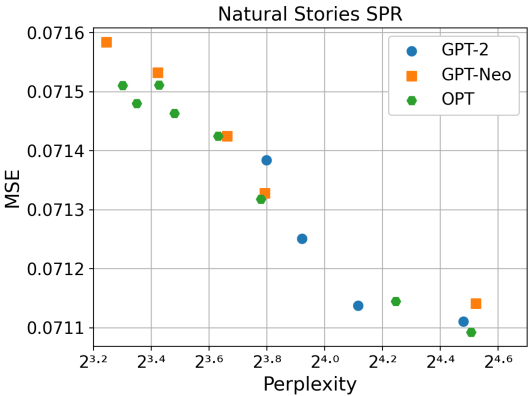
$T(N, \tau)$: Number of gradient updates for a model with N parameters to reach probability of τ

Tirumala et al. (2022). Memorization without overfitting: Analyzing the training dynamics of large language models. In *Proc. NeurIPS*.

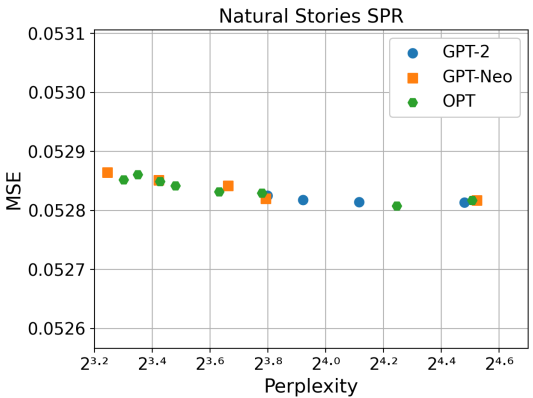
Working hypothesis

LMs begin to learn to predict rare words after a certain point in training, with larger LMs doing so more efficiently. This leads to the bigger-is-worse effects of model size and training data.

Bigger-is-worse effect of model size, revisited

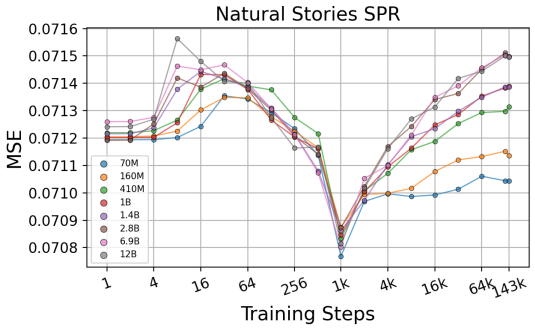


Lowest 20% by unigram frequency

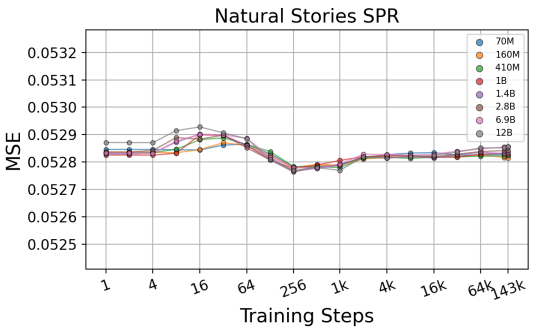


Other 80%

Bigger-is-worse effect of training data, revisited



Lowest 20% by unigram frequency



Other 80%

Conclusion

Conclusion

- Bigger-is-worse effects of model size and training data (Oh & Schuler, 2023a, 2023b)
- Preliminary results suggest that frequency may explain these two effects
- This systematic divergence sheds light on what human sentence processing is not

Thank you for listening!

oh.531@osu.edu byungdoh.github.io
byungdoh/{llm_surprisal,slm_surprisal}

References I

Arehalli, S., Dillon, B., & Linzen, T. (2022). Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. *Proceedings of the 26th Conference on Computational Natural Language Learning*, 301–313. https://aclanthology.org/2022.conll-1.20

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

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

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

Hahn, M., Futrell, R., Gibson, E., & Levy, R. P. (2022). A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, *119*(43). https://doi.org/10.1073/pnas.2122602119

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*.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations*. https://arxiv.org/abs/1412.6980

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

References III

Oh, B.-D., Clark, C., & Schuler, W. (2022). Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, *5*, 777963. https://doi.org/10.3389/frai.2022.777963

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*. https://arxiv.org/abs/2304.11389

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., 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

Shain, C., van Schijndel, M., & Schuler, W. (2018). Deep syntactic annotations for broad-coverage psycholinguistic modeling. *Workshop on Linguistic and Neuro-Cognitive Resources*. http://lrec-conf.org/workshops/lrec2018/W9/pdf/9_W9.pdf

References IV

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

van Schijndel, M., & Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, *45*(6). https://doi.org/10.1111/cogs.12988

Wang, B., & Komatsuzaki, A. (2021). GPT-J-6B: A 6 billion parameter autoregressive language model.

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

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

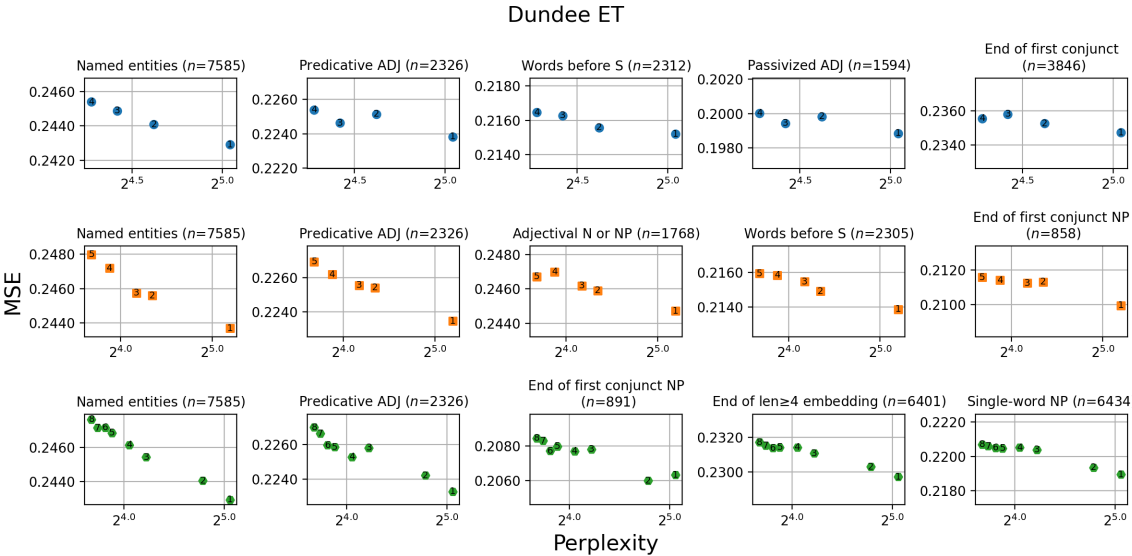
References V

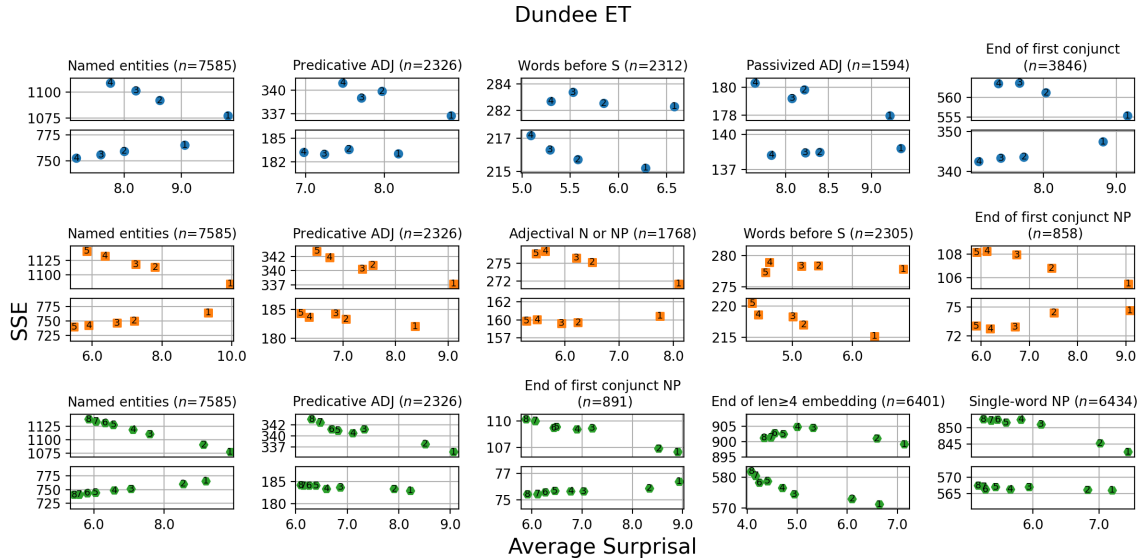
Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., & Zettlemoyer, L. (2022). OPT: Open pre-trained Transformer language models. *arXiv preprint, arXiv:2205.01068v4.*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

Supplementary slides

Regression modeling procedures

- Log-transformed reading times (log ms) in the exploratory partition ($\sim$50%)
- Filtering criteria
  - Natural Stories: initial/final words, $<$100 ms, $>$3000 ms, $<$4 correct answers
  - Dundee: initial/final words, unfixated words, after $>$4 word saccades
- By-subject random slopes for all main effects
  - (1 | subject:sentence) for Natural Stories
  - (1 | sentence) for Dundee

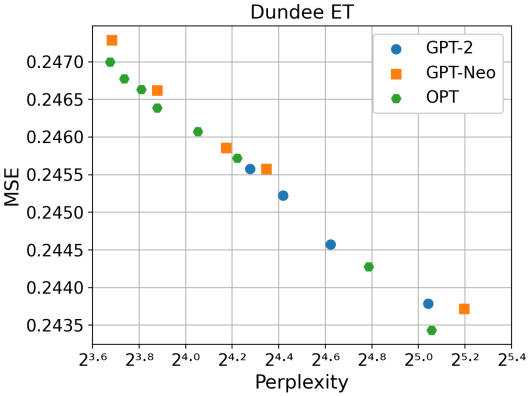

Dundee ET
Named entities (n=7585)
Predicative ADJ (n=2326)
Words before S (n=2312)
Passivized ADJ (n=1594)
End of first conjunct (n=3846)
Named entities (n=7585)
Predicative ADJ (n=2326)
Adjectival N or NP (n=1768)
Words before S (n=2305)
End of first conjunct NP (n=858)
Named entities (n=7585)
Predicative ADJ (n=2326)
End of first conjunct NP (n=891)
End of len≥4 embedding (n=6401)
Single-word NP (n=6434)
MSE
Perplexity


Dundee ET
Named entities (n=7585)
Predicative ADJ (n=2326)
Words before S (n=2312)
Passivized ADJ (n=1594)
End of first conjunct (n=3846)
Named entities (n=7585)
Predicative ADJ (n=2326)
Adjectival N or NP (n=1768)
Words before S (n=2305)
End of first conjunct NP (n=858)
SSE
Named entities (n=7585)
Predicative ADJ (n=2326)
End of first conjunct NP (n=891)
End of len≥4 embedding (n=6401)
Single-word NP (n=6434)
Average Surprisal

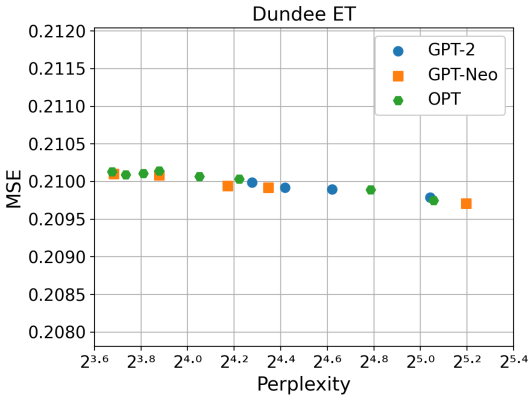
SLM training procedures

- 10k batches from the Pile (Gao et al., 2020), in identical order as the Pythia LMs
- Adam optimizer (Kingma & Ba, 2015), LR warmed up linearly to 0.001 over the first 1% of training steps, then lowered to 0.0001 following a cosine annealing schedule
- Assumes 143000 training steps for comparability with Pythia LMs

Bigger-is-worse effect of model size, revisited (Dundee)

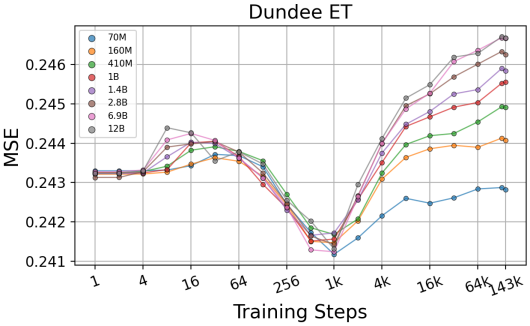


Lowest 20% by unigram frequency

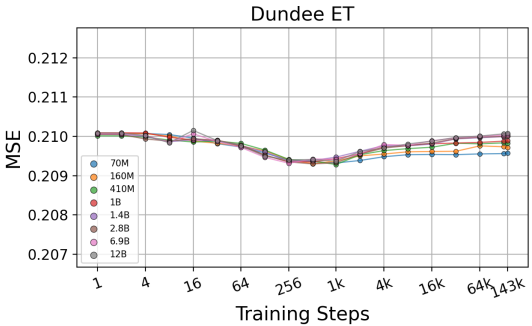


Other 80%

Bigger-is-worse effect of training data, revisited (Dundee)



Lowest 20% by unigram frequency



Other 80%