The impact of token granularity on the predictive power of language model surprisal

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

[1]New York University
[2]The Ohio State University

ACL 2025

Language model surprisal in cognitive modeling

If you were to journey ...

Language model surprisal in cognitive modeling

If you were to journey ...

$$\text{Processing difficulty of } \textit{journey} \propto \underbrace{-\log_2 \mathrm{P}(\textit{journey} \mid \textit{If you were to})}_{\text{surprisal}}$$

Hale (2001), Levy (2008)

Language model surprisal in cognitive modeling

If you were to journey ...

$$\text{Processing difficulty of } \textit{journey} \propto \underbrace{-\log_2 \mathrm{P}(\textit{journey} \mid \textit{If you were to})}_{\text{surprisal}}$$

Hale (2001), Levy (2008)

Word	*If*	*you*	*were*	*to*	*journey*
Reading Time	571 ms	354 ms	386 ms	383 ms	457 ms
LM1 Surprisal	7.76	0.81	5.42	2.09	14.62
LM2 Surprisal	6.71	0.78	5.22	2.30	13.93
LM3 Surprisal	7.10	0.56	5.15	2.39	15.02

Wilcox et al. (2020), Oh and Schuler (2023), *i.a.*

The field has overlooked token granularity

The field has overlooked token granularity

Finer granularity, more character-like ($|V| = 256$)

␣ I f ␣ y o u ␣w er e ␣to ␣ j o ur n e y

The field has overlooked token granularity

Finer granularity, more character-like ($|V| = 256$)

␣ I f ␣ y o u ␣w er e ␣to ␣ j o ur n e y

Coarser granularity, more word-like ($|V| = 128000$)

␣If ␣you ␣were ␣to ␣journey

The field has overlooked token granularity

Finer granularity, more character-like ($|V| = 256$)

␣ I f ␣ y o u ␣w er e ␣to ␣ j o ur n e y

Coarser granularity, more word-like ($|V| = 128000$)

␣If ␣you ␣were ␣to ␣journey

1. Encodes word length and frequency information

The field has overlooked token granularity

Finer granularity, more character-like ($|V| = 256$)

␣ I f ␣ y o u ␣w er e ␣to ␣ j o ur n e y

Coarser granularity, more word-like ($|V| = 128000$)

␣If ␣you ␣were ␣to ␣journey

1. Encodes word length and frequency information
2. Changes co-occurrence statistics, sequence lengths, ...

The field has overlooked token granularity

Finer granularity, more character-like ($|V| = 256$)

␣ I f ␣ y o u ␣w er e ␣to ␣ j o ur n e y

Coarser granularity, more word-like ($|V| = 128000$)

␣If ␣you ␣were ␣to ␣journey

1. Encodes word length and frequency information
2. Changes co-occurrence statistics, sequence lengths, ...

→ *We evaluate surprisal with different token granularities against reading time data*

Methods 1: Tokenizer training

Tokenizer: Unigram language model (Kudo, 2018) tokenizer

Methods 1: Tokenizer training

Tokenizer: Unigram language model (Kudo, 2018) tokenizer

Vocabulary sizes: {256, 512, 1k, 2k, 4k, 8k, 16k, 32k, 48k, 64k, 128k}

Methods 1: Tokenizer training

Tokenizer: Unigram language model (Kudo, 2018) tokenizer

Vocabulary sizes: {256, 512, 1k, 2k, 4k, 8k, 16k, 32k, 48k, 64k, 128k}

Data: 1M articles from English Wiki-40B train (Guo et al., 2020)

Methods 2: Language model training

Neural network architecture: Mamba-2 (Dao & Gu, 2024)

Methods 2: Language model training

Neural network architecture: Mamba-2 (Dao & Gu, 2024)

Models: 11 tokenizers $\times$ 3 sizes

Model	#L	#H	d_{model}	#Parameters
Small	6	8	256	$\sim$2.6M
Medium	12	16	512	$\sim$19.8M
Large	24	24	768	$\sim$88.0M

Methods 2: Language model training

Neural network architecture: Mamba-2 (Dao & Gu, 2024)

Models: 11 tokenizers $\times$ 3 sizes

Model	#L	#H	d_{model}	#Parameters
Small	6	8	256	$\sim$2.6M
Medium	12	16	512	$\sim$19.8M
Large	24	24	768	$\sim$88.0M

Data: $\sim$5.2M articles ($\sim$1.5B words) from English Wiki-40B train (Guo et al., 2020)

Evaluation 1: Impact on fit to naturalistic reading times

Reading times from Natural Stories, Brown, GECO, Dundee, Provo

(Cop et al., 2017; Futrell et al., 2021; Kennedy et al., 2003; Luke & Christianson, 2018; Smith & Levy, 2013)

Evaluation 1: Impact on fit to naturalistic reading times

Reading times from Natural Stories, Brown, GECO, Dundee, Provo

(Cop et al., 2017; Futrell et al., 2021; Kennedy et al., 2003; Luke & Christianson, 2018; Smith & Levy, 2013)

Surprisal calculated from the LMs, both at the start and end of training

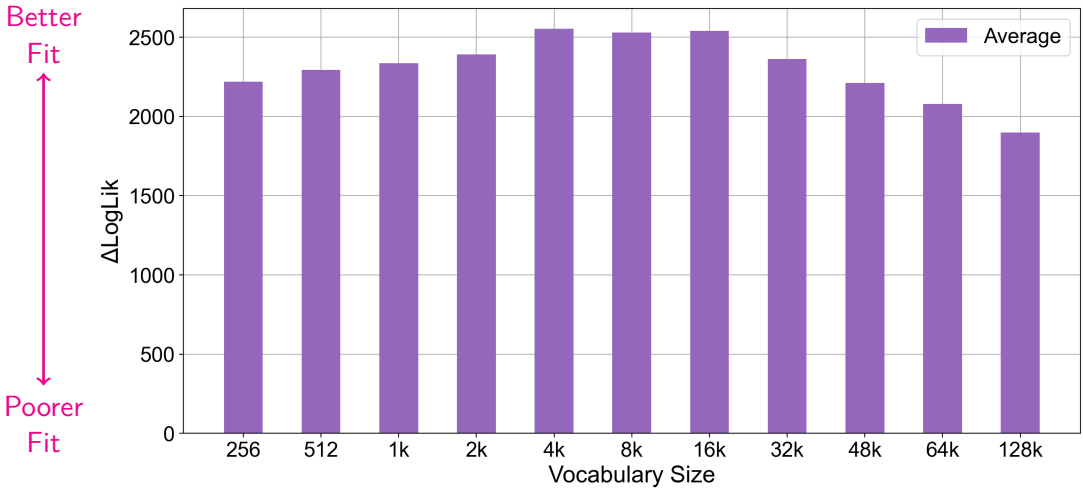
Evaluation 1: Impact on fit to naturalistic reading times

Reading times from Natural Stories, Brown, GECO, Dundee, Provo
(Cop et al., 2017; Futrell et al., 2021; Kennedy et al., 2003; Luke & Christianson, 2018; Smith & Levy, 2013)

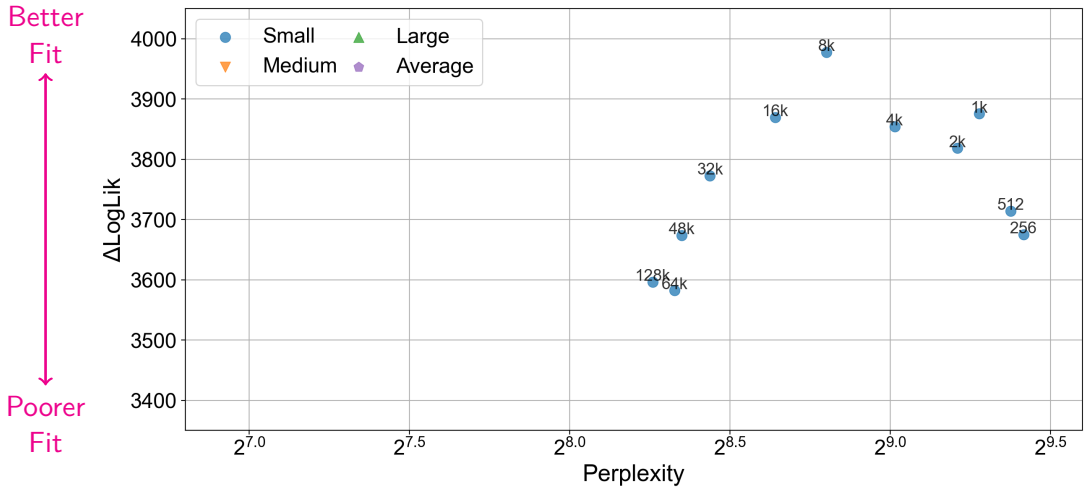
Surprisal calculated from the LMs, both at the start and end of training

Surprisal's contribution to held-out regression log-likelihood (ΔLogLik) measured
($\sim$641k data points)

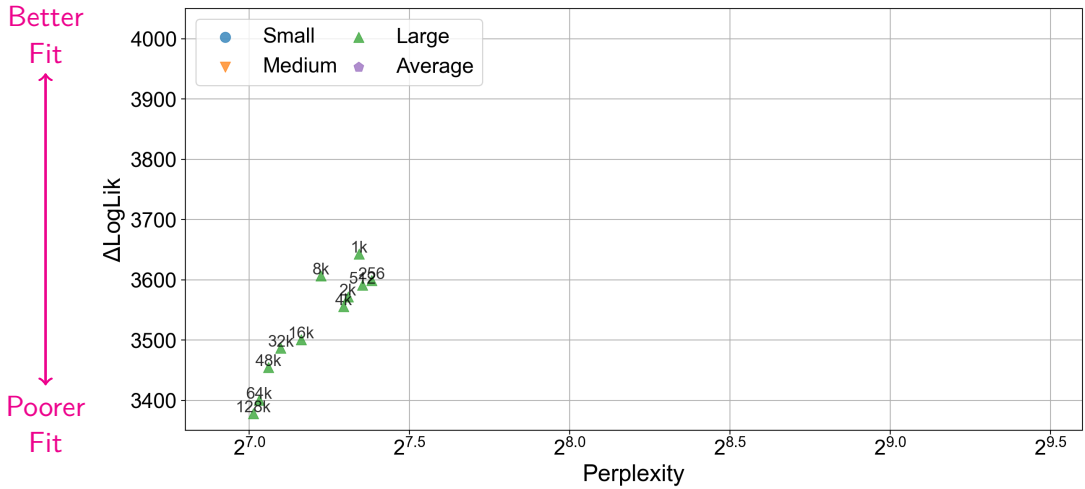
Before LM training: Strong influence of token granularity



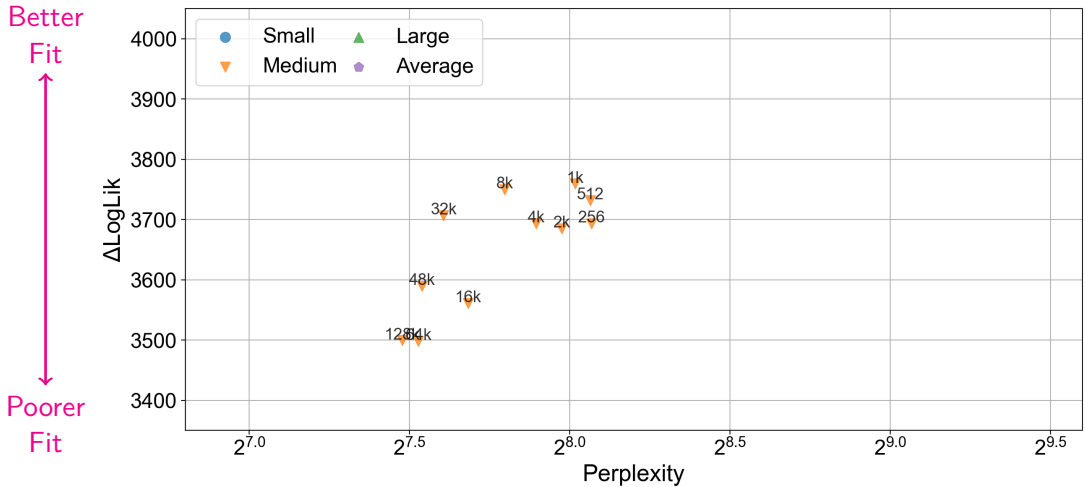
After LM training: Strong interaction between model size and token granularity



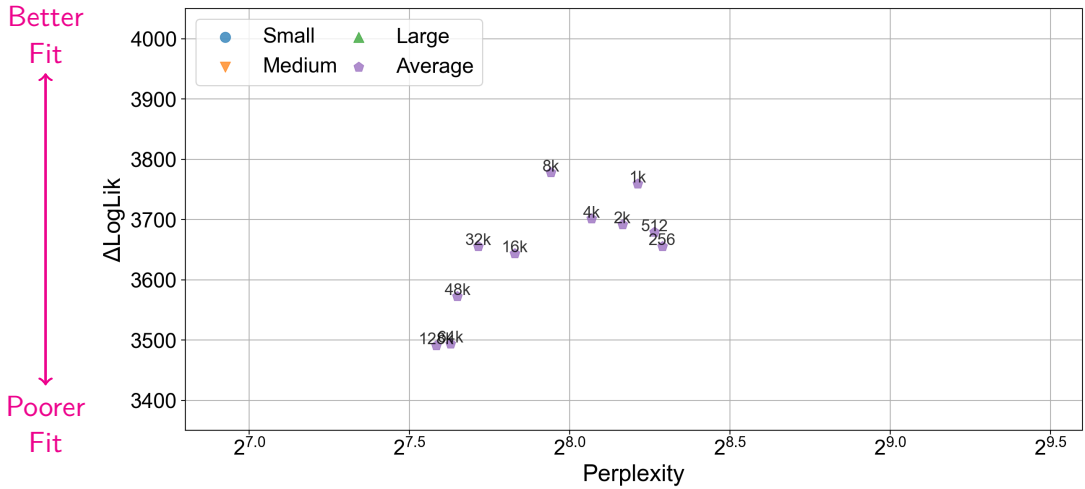
After LM training: Strong interaction between model size and token granularity



After LM training: Strong interaction between model size and token granularity



After LM training: Strong interaction between model size and token granularity



Experiment 2: Impact on surprisal-based garden-path effects

The suspect sent the file deserved further investigation ...
The suspect who was sent the file deserved further investigation ...

→ People read deserved much more slowly than deserved

Experiment 2: Impact on surprisal-based garden-path effects

The suspect sent the file deserved further investigation ...
The suspect who was sent the file deserved further investigation ...

→ People read deserved much more slowly than deserved

Stimuli pairs and self-paced reading data from the SAP Benchmark
(3 garden-path constructions, 24 pairs each; Huang et al., 2024)

Experiment 2: Impact on surprisal-based garden-path effects

The suspect sent the file deserved further investigation ...
The suspect who was sent the file deserved further investigation ...

→ People read deserved much more slowly than deserved

Stimuli pairs and self-paced reading data from the SAP Benchmark
(3 garden-path constructions, 24 pairs each; Huang et al., 2024)

Surprisal-to-RT conversion model fit to reading times of non-garden-path sentences
(∼996k data points)

Experiment 2: Impact on surprisal-based garden-path effects

The suspect sent the file deserved further investigation ...
The suspect who was sent the file deserved further investigation ...

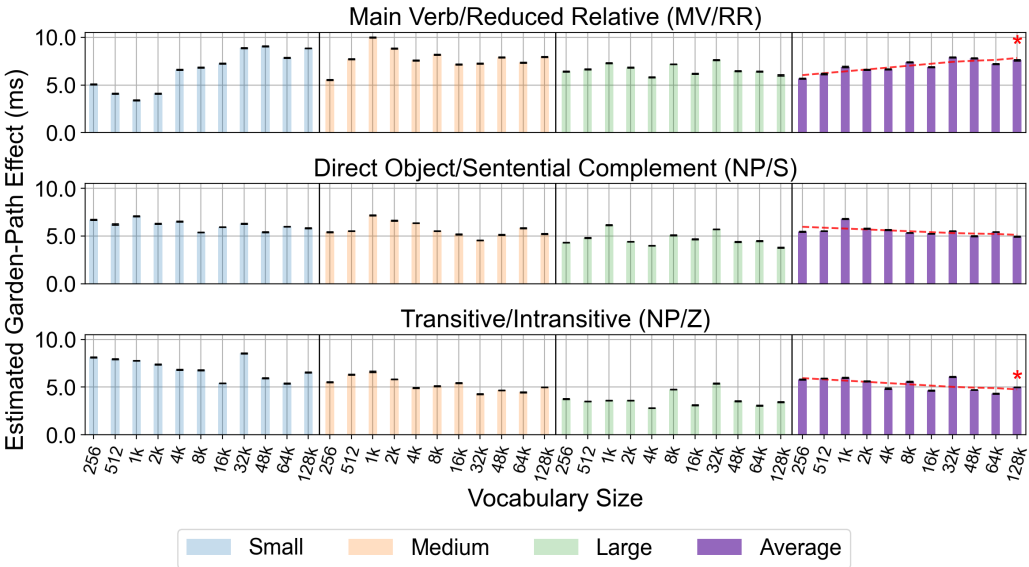
→ People read deserved much more slowly than deserved

Stimuli pairs and self-paced reading data from the SAP Benchmark
(3 garden-path constructions, 24 pairs each; Huang et al., 2024)

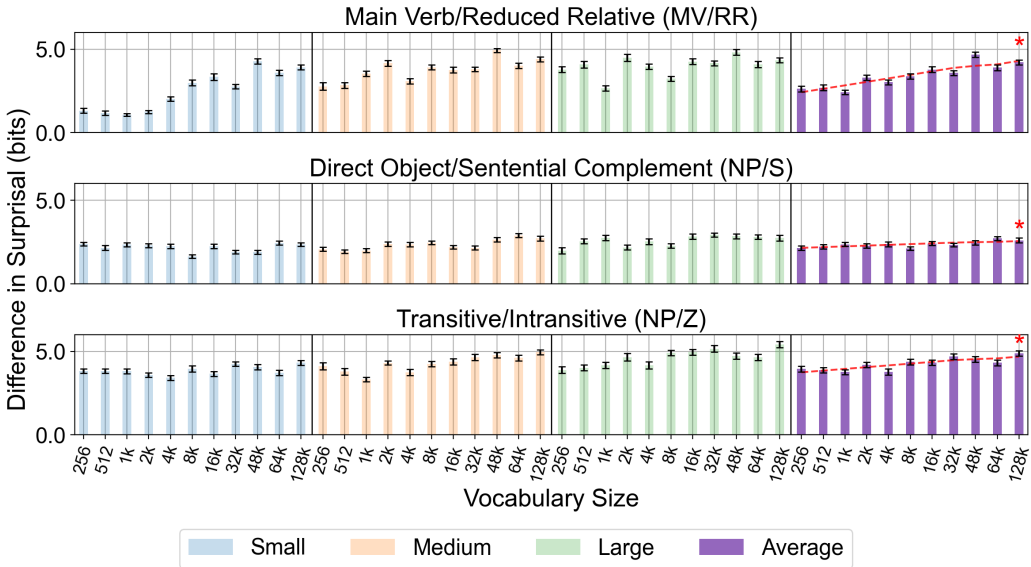
Surprisal-to-RT conversion model fit to reading times of non-garden-path sentences
(∼996k data points)

Conversion model used to predict difference in reading times at deserved and deserved
(∼48k data points)

No clear trend in estimated garden-path effects



Coarser-grained tokens lead to larger differences in raw surprisal



Conclusion

Strong influence of token granularity, especially for smaller models

Conclusion

Strong influence of token granularity, especially for smaller models

Improved fit to reading times probably due to ‘sneaky’ word length and frequency

Conclusion

Strong influence of token granularity, especially for smaller models

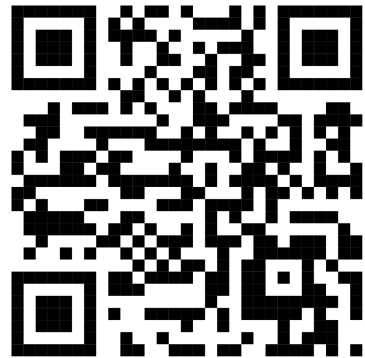
Improved fit to reading times probably due to ‘sneaky’ word length and frequency

Hot take: Let’s use coarser-grained tokens – less prone to this issue, easier to interpret

Thank you for listening!

oh.b@nyu.edu
byungdoh/ssm-surprisal
byungdoh/ssm-token-granularity

This work was supported by NSF grant #1816891 and NYU IT High Performance Computing resources, services, and staff expertise.



References I

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

Dao, T., & Gu, A. (2024). Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. *Proceedings of the 41st International Conference on Machine Learning*, *235*, 10041–10071. https://proceedings.mlr.press/v235/dao24a.html

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

Guo, M., Dai, Z., Vrandečić, D., & Al-Rfou, R. (2020). Wiki-40B: Multilingual language model dataset. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2440–2452. https://aclanthology.org/2020.lrec-1.297

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/

Huang, K.-J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2024). Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, *137*, 104510. https://doi.org/10.1016/j.jml.2024.104510

References II

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

Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 66–75. https://aclanthology.org/P18-1007

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. (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

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