The Inverse Scaling Effect of Pre-Trained Language Model Surprisal Is Not Due to Data Leakage

Byung-Doh Oh[1] Hongao Zhu[2] William Schuler[3]

oh.b@nyu.edu
github.com/byungdoh/rt-leakage

[1]New York University [2]Shanghai Jiao Tong University [3]The Ohio State University

PAPER

Key Takeaways

The inverse scaling effect of pre-trained LM surprisal is unlikely to be due to data leakage. We provide two reasons why:

1. Widely used reading time corpora suffer little from data leakage measured in terms of token n-gram overlap
2. The inverse scaling effect of surprisal is replicated with LMs trained on 'leakage-free' data

Language processing and LM surprisal

People experience processing difficulty at unpredictable words,

The coast guard had seen a shark
The coast guard had seen a zebra

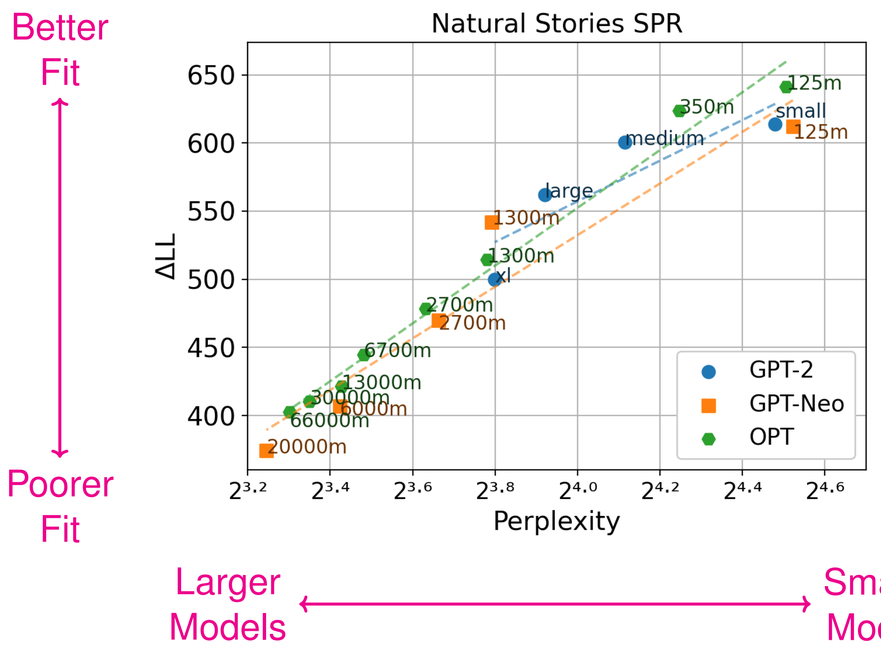
e.g. Ehrlich and Rayner [2]

which has led to research that evaluates LM surprisal (negative log probability) as a predictor of human reading times

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

Background: Inverse scaling effect of LM surprisal

Surprisal from larger pre-trained LMs yields a poorer fit to reading times [figure from 9]



However, this may be due to data leakage, as the text material used to collect reading times is often available online [11]

Method: n-gram overlap detection using CDAWGs

Compacted Directed Acyclic Word Graphs [CDAWGs; 8] efficiently return the longest attested suffix of the query sequence

Reference: `h e l l o w o r l d`
Query: `l l o y d`

Output: `1 2 3 0 1`

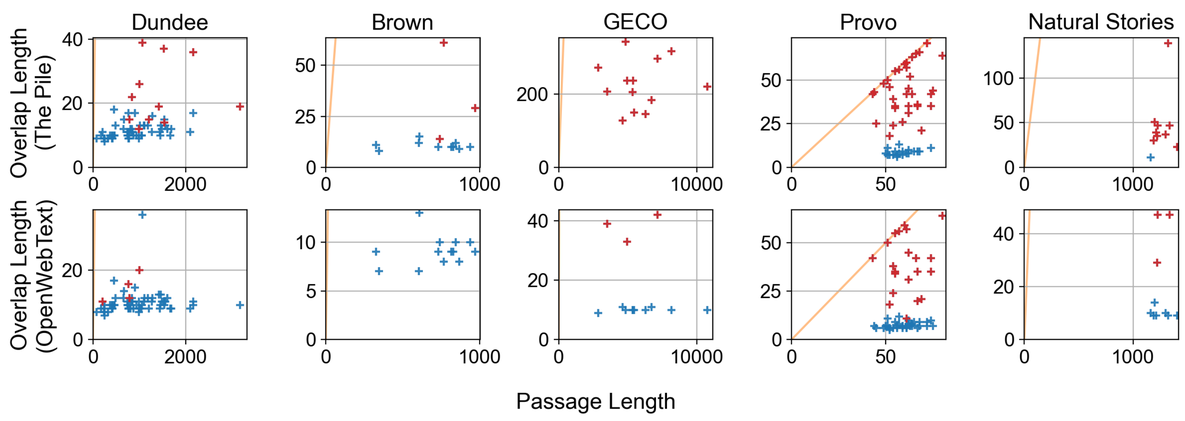
The longest n-gram overlap between reference and query is of length 3

Study 1: Assessment of data leakage

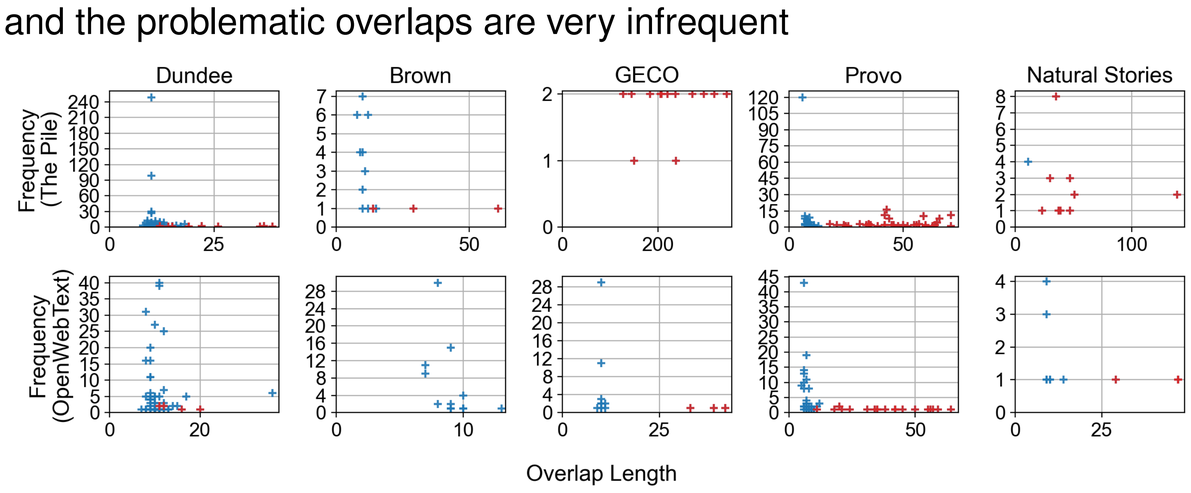
Reference (pre-training data): The Pile, OpenWebText [4, 5]

Query (reading time data): Dundee, Brown, GECO, Provo, Natural Stories [6, 10, 1, 7, 3]

Except for Provo, overlaps are relatively short in light of passage length,



and the problematic overlaps are very infrequent

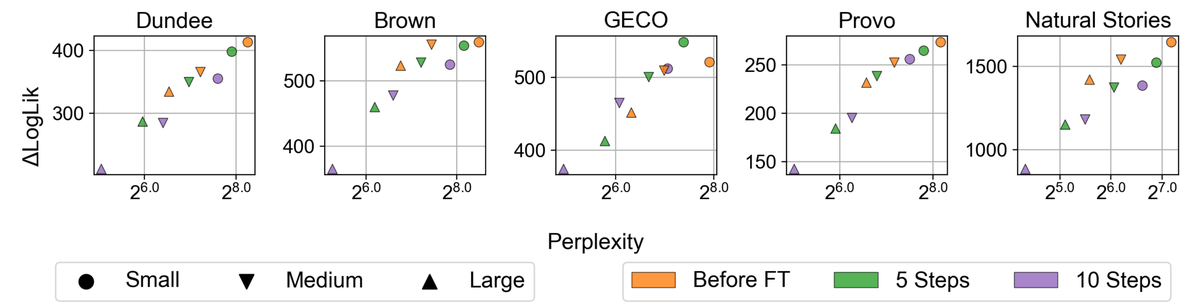


Study 2: Evaluation of LMs trained 'leakage-free'

Clean data: ~21B tokens with at most 11 continuous tokens of overlap

Artificial leakage: Fine-tuning for 5 and 10 steps on reading time corpora

Inverse scaling is replicated with clean data, but leakage strengthens effect



[1] Cop, U., Dirix, N., Drieghe, D., et al. 2017. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading.

[2] Ehrlich, S. F., & Rayner, K. 1981. Contextual effects on word perception and eye movements during reading.

[3] Futrell, R., Gibson, E., Tily, H. J., et al. 2021. The Natural Stories Corpus: A reading-time corpus of English texts containing rare syntactic constructions.

[4] Gao, L., Biderman, S., Black, S., et al. 2020. The Pile: An 800GB dataset of diverse text for language modeling.

[5] Gokaslan, A., & Cohen, V. 2019. OpenWebText Corpus. `http://Skylion007.github.io/OpenWebTextCorpus`

[6] Kennedy, A., Hill, R., & Pynte, J. 2003. The Dundee Corpus.

[7] Luke, S. G., & Christianson, K. 2018. The Provo Corpus: A large eye-tracking corpus with predictability norms.

[8] Merrill, W., Smith, N. A., & Elazar, Y. 2024. Evaluating n-gram novelty of language models using Rusty-DAWG.

[9] Oh, B.-D., & Schuler, W. 2023. Why does surprisal from larger Transformer-based language models provide a poorer fit to human reading times?.

[10] Smith, N. J., & Levy, R. 2013. The effect of word predictability on reading time is logarithmic.

[11] Wilcox, E. G., Meister, C., Cotterell, R., et al. 2023. Language model quality correlates with psychometric predictive power in multiple languages.