

ACL 2025

The Impact of Token Granularity on the Predictive Power of Language Model Surprisal

Byung-Doh Oh

Center for Data Science
New York University
oh.b@nyu.edu

William Schuler

Department of Linguistics
The Ohio State University
schuler.77@osu.edu

紹介者: 笹野遼平 (名大)

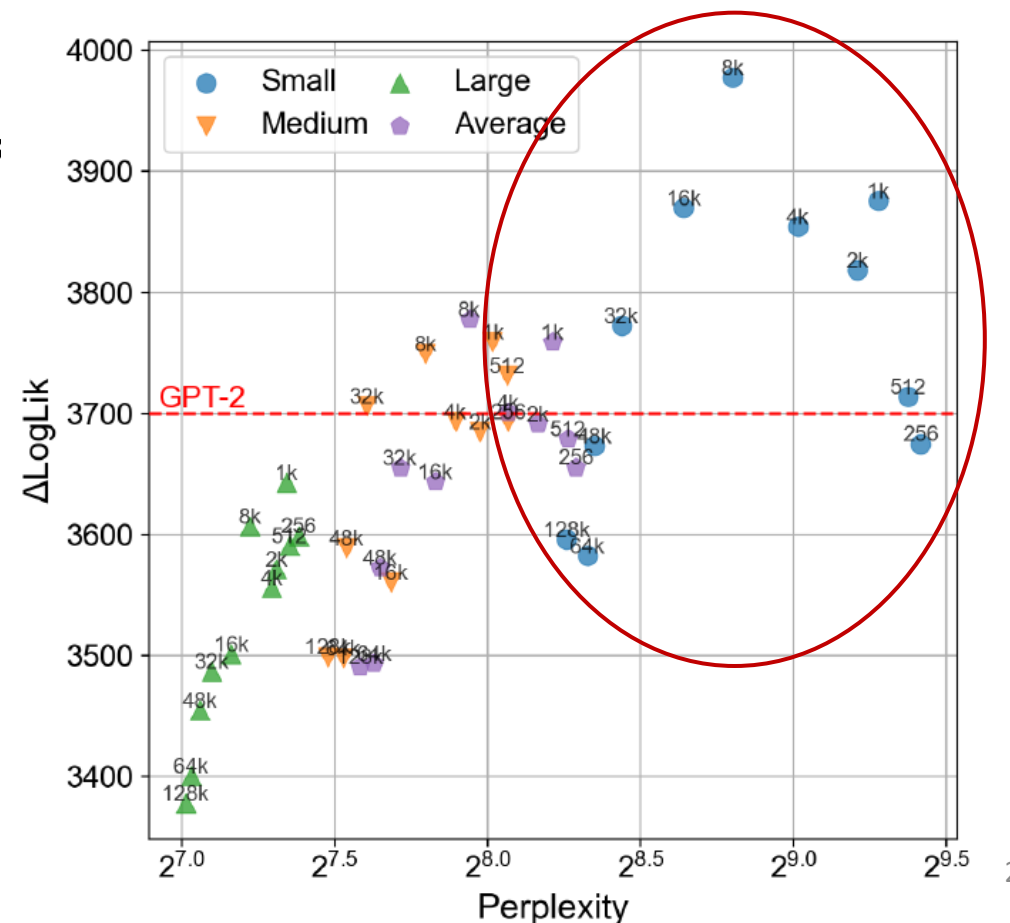
一部の図は、論文または
著者の発表スライドより引用

論文の概要

- 「言語処理時の認知負荷をサプライザルがどの程度説明できるか」を分析する際の「言語モデルのトークン粒度の影響」を調査
 - 読み時間の実験では、Smallモデルに対しては、中間的な粒度 ($|V| \approx 8,000$) が最も予測精度が高くなるという結果
 - ガーデンパス文を用いた実験では

“LMs trained on coarser-grained tokens generally assigned higher surprisal to critical regions, suggesting a greater sensitivity to garden-path effects than previously reported”

と主張



著者について

- サプライザルに関連する論文を多く発表 (psycholinguistics)
- 2023年のTACLの論文
“Why does surprisal from larger Transformer-based language models provide a poorer fit to human reading times?”
は一昨年のSNLPで栗林さんが紹介



Byung-Doh Oh

[New York University](#)

Verified email at nyu.edu - [Homepage](#)

[computational linguistics](#) [psycholinguistics](#) [natural language processing](#)



TITLE	CITED BY	YEAR
Why does surprisal from larger Transformer-based language models provide a poorer fit to human reading times? BD Oh, W Schuler Transactions of the Association for Computational Linguistics 11, 336-350	164	2023
Comparison of structural parsers and neural language models as surprisal estimators BD Oh, C Clark, W Schuler Frontiers in Artificial Intelligence 5, 777963	70	2022
Transformer-based language model surprisal predicts human reading times best with about two billion training tokens BD Oh, W Schuler Findings of the Association for Computational Linguistics: EMNLP 2023, 1915-1921	52 *	2023
Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal 's fit to reading times BD Oh, S Yue, W Schuler Proceedings of the 18th Conference of the European Chapter of the ...	31	2024
Entropy- and distance-based predictors from GPT-2 attention patterns predict reading times over and above GPT-2 surprisal BD Oh, W Schuler Proceedings of the 2022 Conference on Empirical Methods in Natural Language ...	31	2022
Surprisal estimators for human reading times need character models BD Oh, C Clark, W Schuler Proceedings of the 59th Annual Meeting of the Association for Computational ...	22	2021

前提知識 1 : サプライザル理論

- 人間が文を処理するときに経験する「処理の難しさ」を情報理論の枠組みを用いて説明しようとする理論 (Hale 2001, Levy 2008)

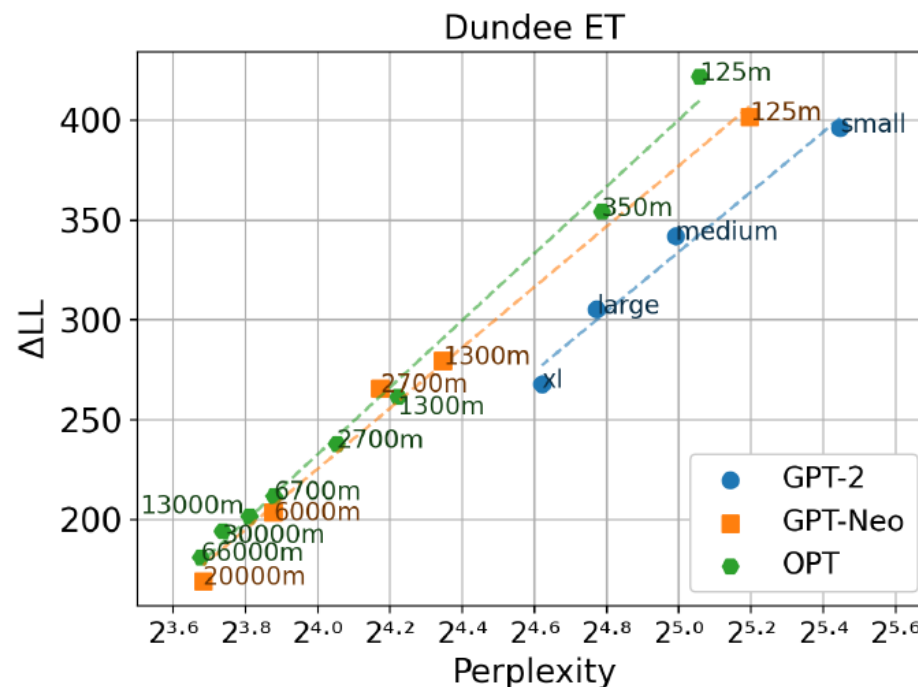
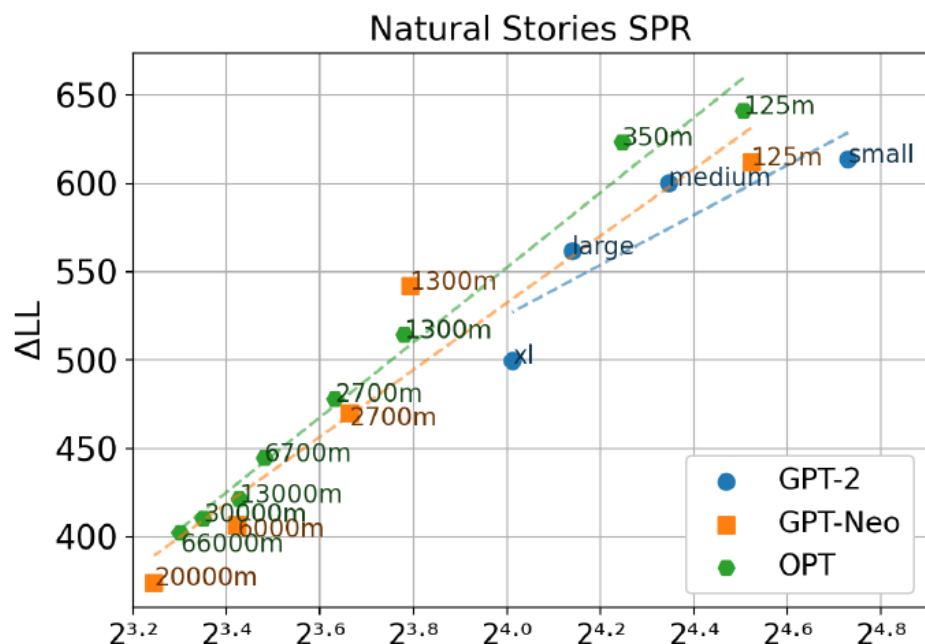
$$\text{Processing difficulty of } w_t \propto -\log P(w_t | w_{1...t-1})$$

サプライザル

Word	<i>If</i>	<i>you</i>	<i>were</i>	<i>to</i>	<i>journey</i>
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

前提知識 2 : LMの性能とサプライザル

- 小さなperplexityが得られるような大規模なTransformerベースのモデルが読み時間推定において高い性能とならない (Kuribayashi et al. 2021, Oh&Schuler 2023b)



縦軸は読み時間推定
への効果 ΔLL (↑)、
横軸はPerplexity (↓)
(Oh&Schuler 2023b)

サブライザル計算におけるトークン粒度の影響

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

▯ I f ▯ y o u ▯ w e r e ▯ t o ▯ j o u r n e y

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

▯ If ▯ you ▯ were ▯ to ▯ journey

- サブライザルの値はトークン粒度による影響を大きく受ける可能性
- $|V| = 256$ の場合、1トークンはほぼ1文字 $\Rightarrow$ to と journey の生成確率はおおよそ6乗分異なる ($p^1 \Leftrightarrow p'^7$)
- $|V| = 128k$ の場合、to と journey の生成確率は同等になる

実験設定

- トークナイザー
 - トークナイザー: Unigram言語モデル (Kudo 2018)
 - 語彙サイズ: {256, 512, 1k, 2k, 4k, 8k, 16k, 32k, 48k, 64k, 128k}
 - 訓練データ: English Wiki-40B trainから100万記事
- 言語モデル
 - アーキテクチャ: Mamba-2 (Dao&Gu 2024) (長い系列でも比較的高速)
 - 訓練データ: English Wiki-40B trainから520万記事(～1.5B words)
 - モデルサイズ:
 - Small, Medium, Large
- 語彙サイズ: 11 ×
モデルサイズ: 3 = 33種類

Model	#L	#H	d_{model}	#Parameters
<i>Small</i>	6	8	256	～2.6M
<i>Medium</i>	12	16	512	～19.8M
<i>Large</i>	24	24	768	～88.0M

実験 1 : 読み時間データを使った実験

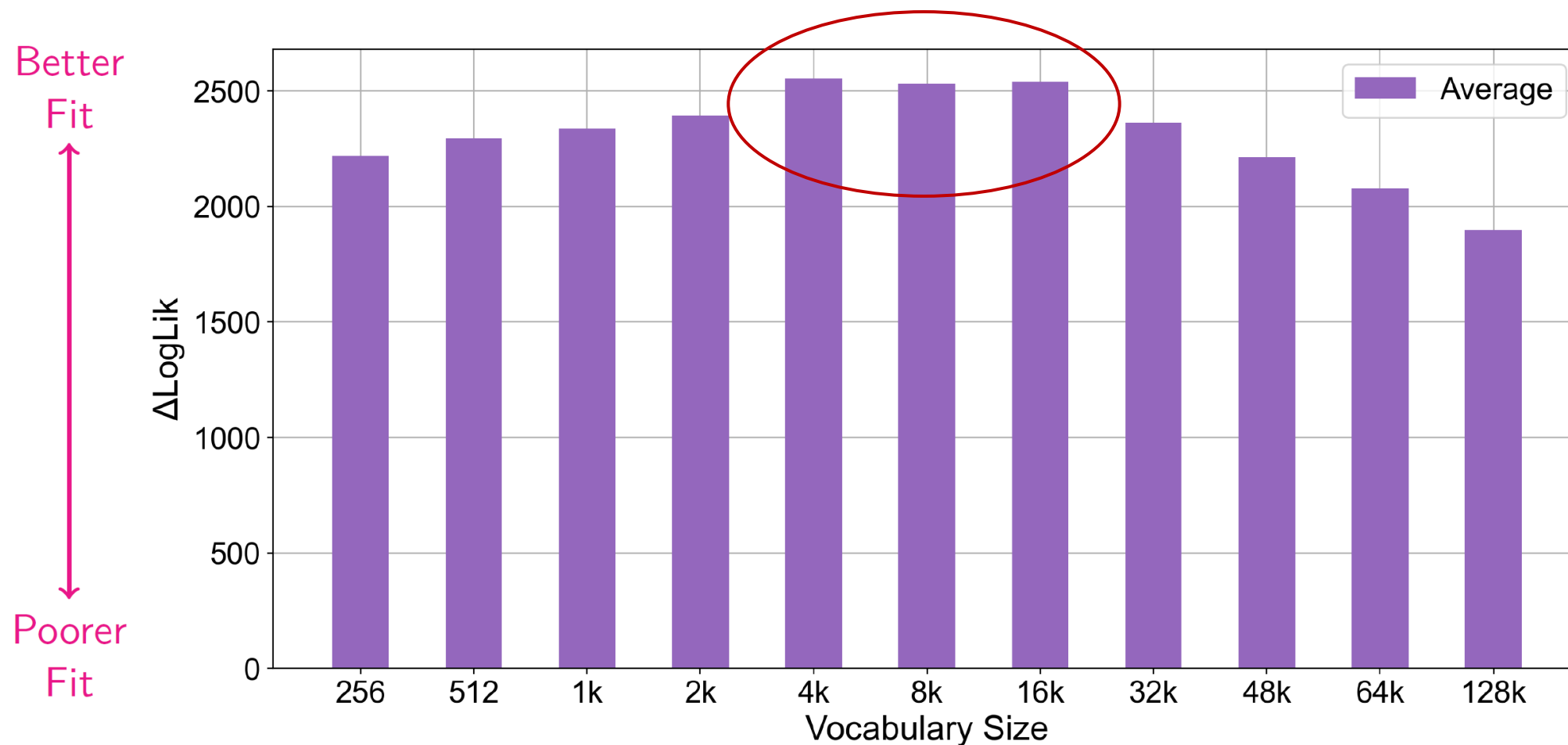
- 手法

- コーパス 5 種類: Natural Stories, Brown, GECO, Dundee, Provo
- 訓練前と訓練後の各言語モデルからサプライズを算出
- 読み時間を線形混合効果(LME)回帰モデルによりモデリング

- 評価

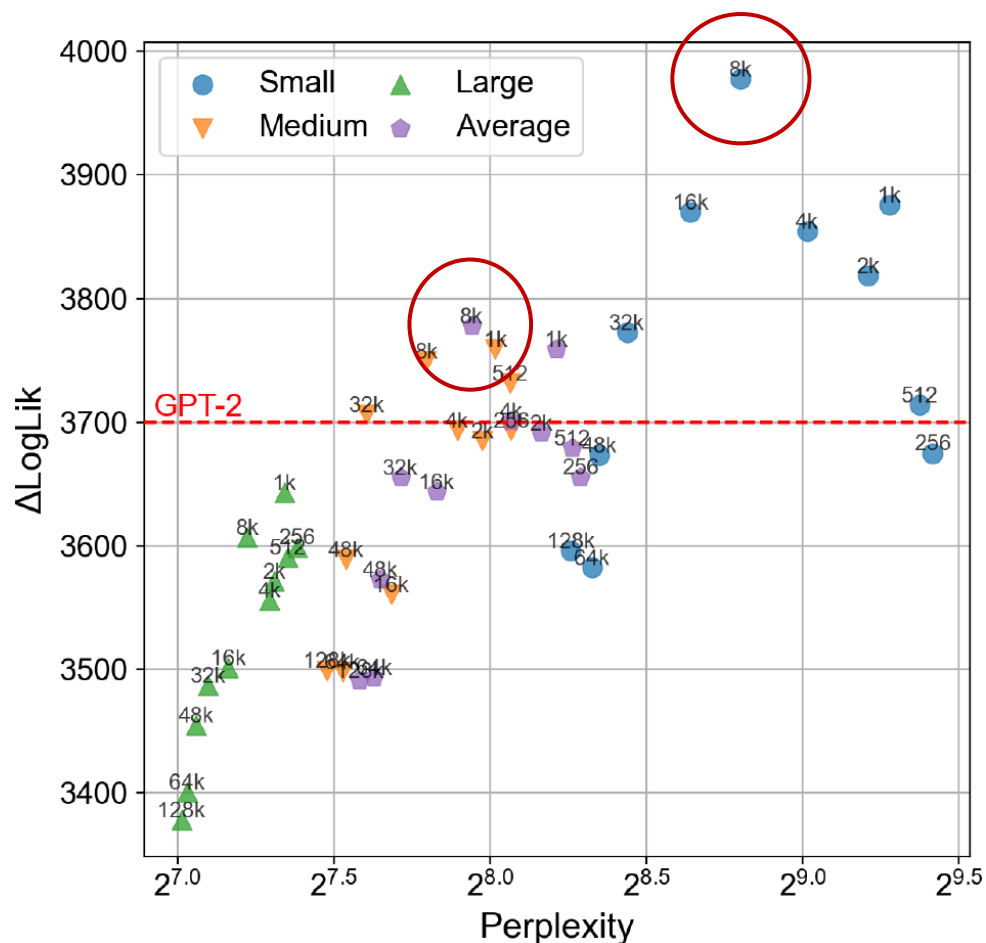
- ベースとする回帰モデルのサプライズ値を、各LMを用いて算出したサプライズ値に置換したときの対数尤度の増加量 ΔLogLik で評価
 - ベースモデルでは単語の文字数、その単語の文中の位置などの情報に加えて、KenLMを使って算出したunigramサプライズを使用
 - LMEモデルはガウス分布に基づくので与えられた読み時間に対する尤度が算出可能

結果 1 : 訓練前のモデルを使った実験



※ 言語モデルのサイズによる違い限定的なのでAverageのみを掲載

結果 2 : 訓練後のモデルを使った実験



- Smallモデルにより算出されたサブライザルは訓練前のピークを再現
- 一方、Medium, Largeモデルについてはピークは観測できず
 - Perplexityが小さいほど ΔLogLik が小さくなる傾向を再現
- 著者らは3モデルの平均について語彙サイズを8kとした場合にGPT-2 Smallを超えることを強調しているが平均をとる妥当性は不明

前提知識 3 : ガーデンパス文

Garden path effect (GPE)

“who was”がないとdeservedを読む時間が大幅に遅くなる

The suspect sent the file deserved further investigation given the new evidence.

who was

ChatGPT 5

以下の文を日本語に翻訳してください。
The suspect sent the file deserved further investigation given the new evidence.

以下の文を日本語に翻訳しました。

「容疑者が送ったファイルは、新しい証拠を踏まえると、さらなる調査に値した。」

Google 翻訳

テキスト 画像 ウェブサイト

英語 日本語

The suspect sent the file deserved further investigation given the new evidence.

81 / 5,000

容疑者は、新たな証拠を考慮すると、さらなる捜査に値するファイルを送った。

DeepL 翻訳

The suspect sent the file deserved further investigation given the new evidence.

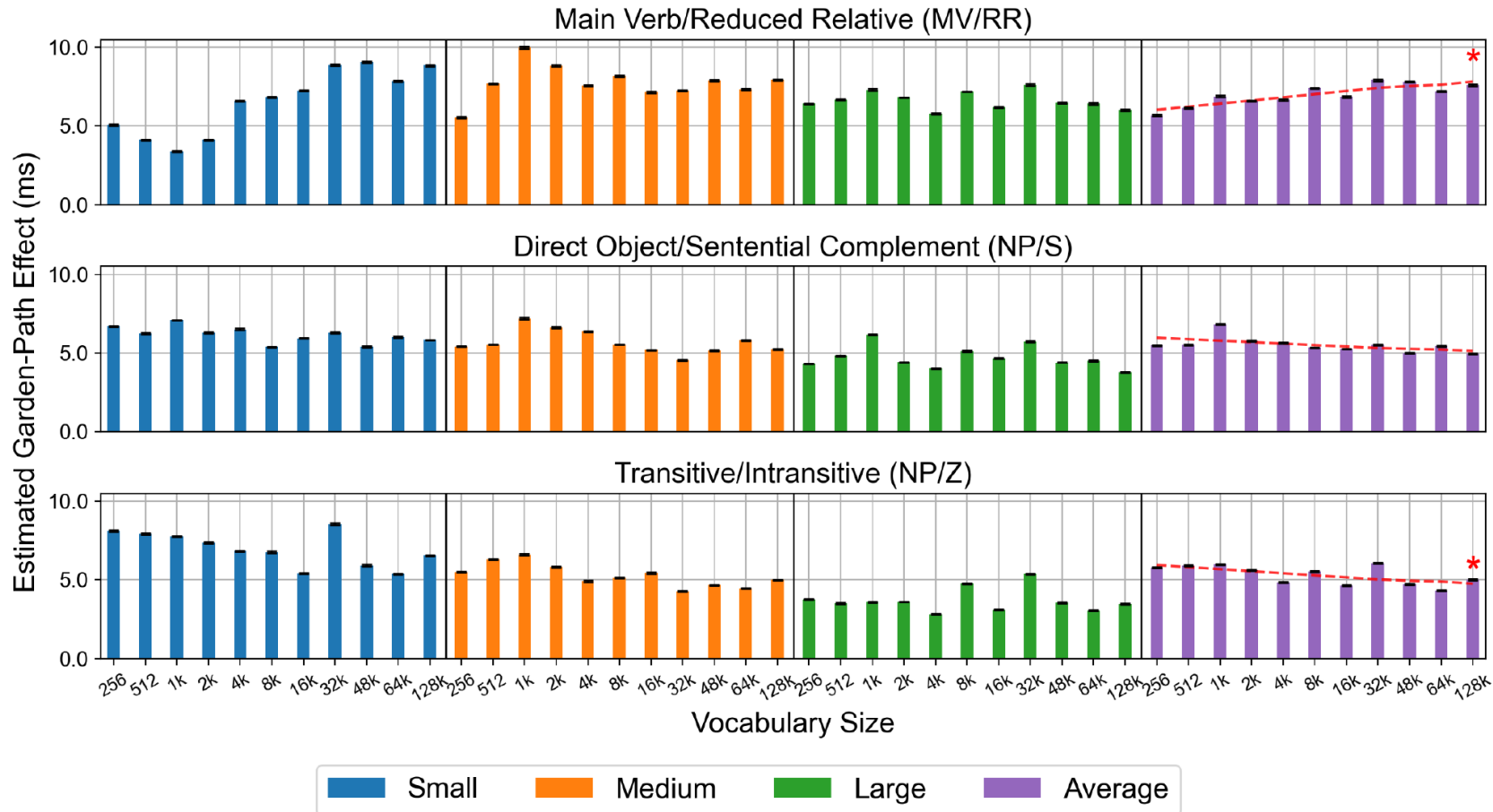
新たな証拠を踏まえると、容疑者が送ったファイルはさらなる調査が必要であった。

実験 2 : ガーデンパス文を使った実験

Construction/Condition	Example
MV/RR Ambiguous	The suspect sent the file <i>deserved</i> further investigation given the new evidence.
MV/RR Unambiguous	The suspect who was sent the file <i>deserved</i> further investigation given the new evidence.
NP/S Ambiguous	The suspect showed the file <i>deserved</i> further investigation during the murder trial.
NP/S Unambiguous	The suspect showed that the file <i>deserved</i> further investigation during the murder trial.
NP/Z Ambiguous	Because the suspect changed the file <i>deserved</i> further investigation during the jury discussions.
NP/Z Unambiguous	Because the suspect changed, the file <i>deserved</i> further investigation during the jury discussions.

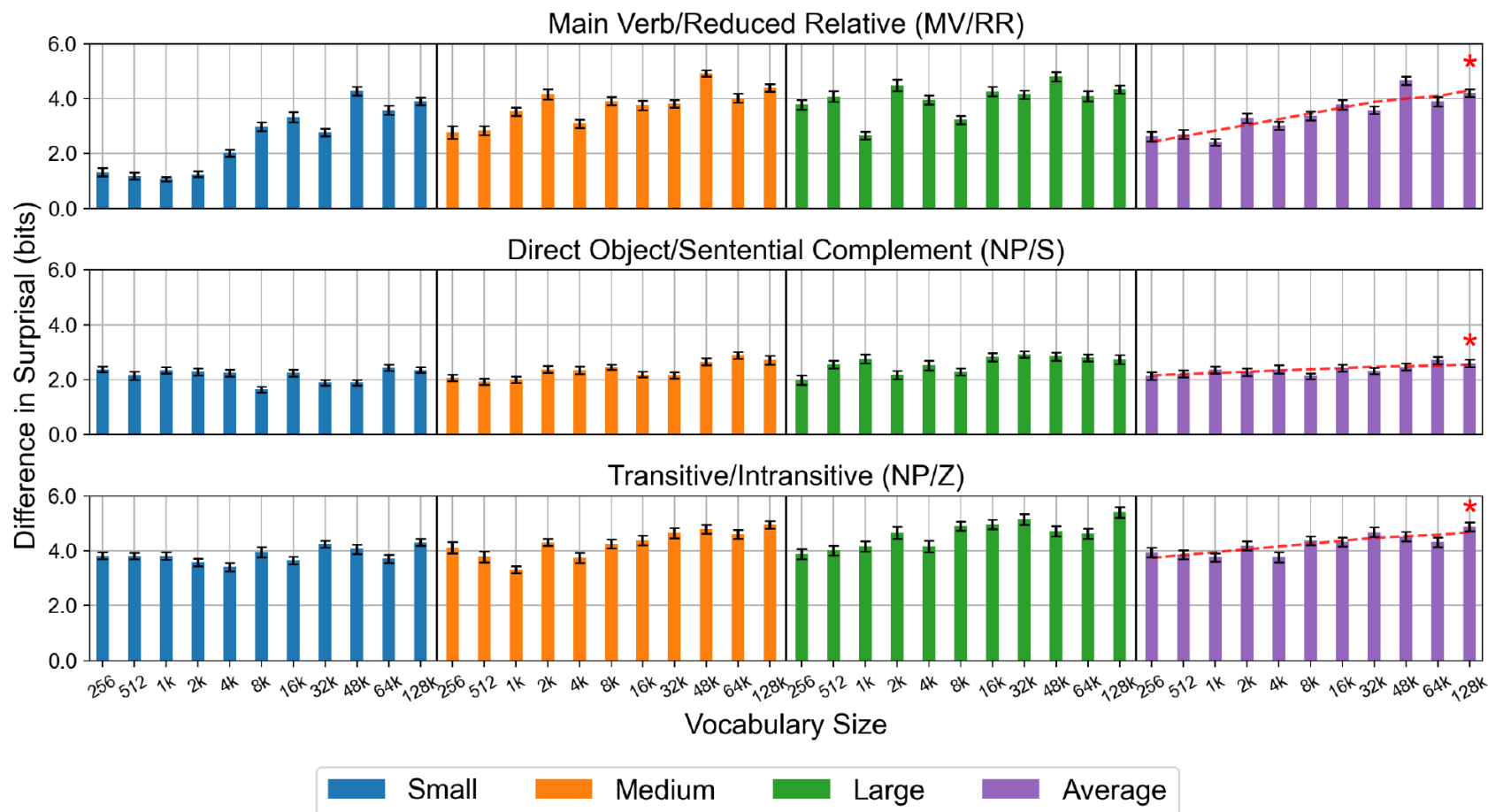
- 3 タイプ、計 24 ペアのデータを使用 (Huang et al. 2024)
- 構文的曖昧性による処理困難さを持たない文を対象にself-paced reading (SPR)にfitするようにLMEモデルを調整
- “*critical word*”と後続する2語(spillover words)におけるサプライザルの増加による読み時間の予測値の増加量を算出

1つ目のspillover wordのGPE予測値



- “no clear trend in estimated garden-path effects” (著者スライドより)
- 人におけるGPEと比べ1～2桁小さい効果しか観測できない問題は改善されず

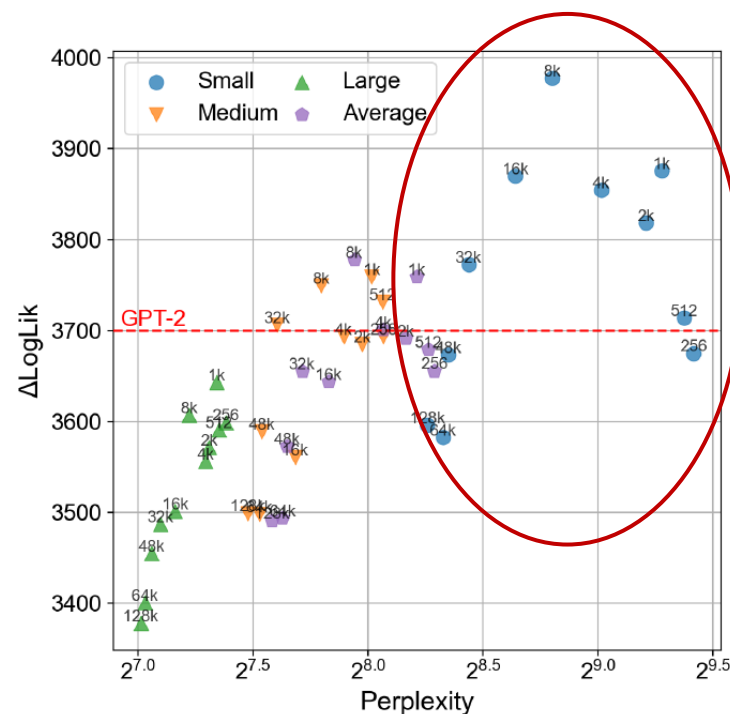
critical wordにおけるサプライザルの増加



- “Coarser-grained tokens lead to larger differences in raw surprisal” (著者スライドより)
- この結果から何らかの結論を得るのは危険な印象

雑感

- サプライザル算出におけるトークン粒度は重要な問題
- 読み時間の実験において、もともと性能が高いSmallモデルに対し中間的なトークン粒度で ΔLogLik が最大となることは有用な知見
 - 今後はSmall $\times$ 8k で実験すると良い？ 一方で意外性は限定的？
- ガーデンパス(GP)文の実験から何らかの知見を主張するのは厳しそう
 - 著者の主張も一貫していない
 - そもそもLLMはGP文を解析できていない？
- なぜmain (oral)に採択されたのか？
 - ドメイン知識を前提とし説明がやや不親切な印象
 - Mamba2を採用するなど技術的な面が評価？
 - 心理言語学分野の査読者の興味に合致？



参考文献

- John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. NAACL
- Levy. 2008. Expectation-based syntactic comprehension. Cognition
- Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower perplexity is not always human-like. ACL-IJCNLP
- Byung-Doh Oh and William Schuler. 2023b. Why does surprisal from larger Transformer-based language models provide a poorer fit to human reading times? TACL
- Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. ACL
- Tri Dao and Albert Gu. 2024. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. ICML
- Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and
- Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. Journal of Memory and Language