On the Bigger-is-Worse Nature of Pre-Trained Language Model Surprisal

Byung-Doh Oh (oh.531@osu.edu) and William Schuler (The Ohio State University)

In cognitive modeling, predictability operationalized by information-theoretic surprisal [9] has been shown to be a strong predictor of behavioral and neural measures of incremental processing difficulty [10, 4]. As language models (LMs) directly define a conditional probability distribution of a word given its context required for surprisal calculation, they are frequently evaluated as surprisal-based cognitive models of sentence processing. Recently, it was observed that surprisal from larger variants of the pre-trained GPT-2 LM [7] that have more parameters and achieve lower perplexity is *less* predictive of self-paced reading times and eye-gaze durations collected during naturalistic reading of English text [6]. As the different variants of the pre-trained GPT-2 model share the primary architecture and training data, this offers an especially strong counterexample to earlier work that showed a *negative* relationship between LM perplexity and predictive power of surprisal estimates [3, 11], leaving open the question of why larger LMs perform worse.

To examine whether this trend generalizes to other pre-trained LMs, linear mixed-effects regression analyses were conducted to evaluate the quality of surprisal predictors in terms of improvement in regression model log-likelihood (ΔLL). To this end, surprisal predictors for the Natural Stories self-paced reading corpus [2] and the Dundee eye-tracking corpus [5] were calculated from variants of the GPT-2 [7], GPT-Neo [1], and the OPT LM [12] that vary mostly in terms of model size. The baseline predictors include word length measured in characters and index of word position within each sentence (both Natural Stories and Dundee), as well as saccade length and whether or not the previous word was fixated (Dundee only). All predictors were z-transformed prior to fitting, and all regression models included by-subject random slopes for all fixed effects. The results in Figure 1 show that surprisal from the smallest variants made the biggest contributions to regression model fit on both self-paced reading times and eye-gaze durations for all three LM families. More notably, surprisal estimates from larger LM variants within each family yielded strictly progressively poorer fits to reading times, robustly replicating the trend observed in [6]. Interestingly, the three LM families also showed a strong *positive* log-linear relationship between perplexity and ΔLL, which was significant by a one-tailed t-test on the slope of the regression lines.

To provide an explanation for this trend, the residual errors from the regression models were analyzed to identify data points that surprisal from larger LM variants accounted for less accurately compared to their smaller counterparts. To this end, each data point in both corpora was associated with various word- and sentence-level properties that are thought to influence real-time processing, which were derived from manually annotated syntactic tree structures [8]. Subsequently, for every corpus-LM combination, subsets that showed the largest differences in MSE between regression models were identified. The results in Figure 2 show that on each corpus, similar subsets were identified as driving the trend of MSEs across different LM families (note that MSE, which can be attributed to each point, is lower when ΔLL is higher). On Natural Stories, these subsets were primarily determined by named entities and syntactic categories such as named entities, nouns before relativizers, attributive and predicative adjectives, and modal auxiliaries. The top subsets of Dundee were similarly determined by named entities and syntactic categories such as predicative and passivized adjectives, and single-word noun phrases. A further breakdown of residuals errors shows that surprisal estimates from larger LM variants underpredict reading times of open-class words more severely and make compensatory overpredictions for reading times of function words.

These results suggest that the propensity of larger LMs to glean extensive domain knowledge from vast quantities of text during training makes their surprisal estimates diverge from human-like expectations, which warrants caution in using them to study human language processing. As pre-trained LMs get larger, they may be more problematic for cognitive modeling as they are increasingly better-read than humans and therefore better able to predict descriptions from context.

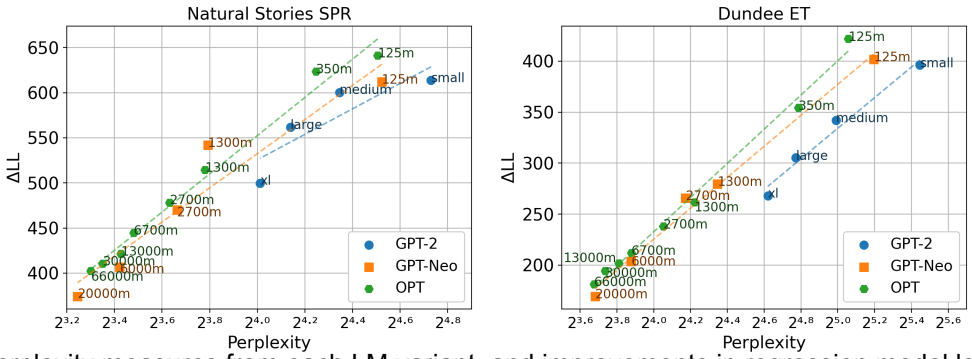

Figure 1: Perplexity measures from each LM variant, and improvements in regression model log-likelihood from including each surprisal estimate on the exploratory set of Natural Stories (left) and Dundee data (right).

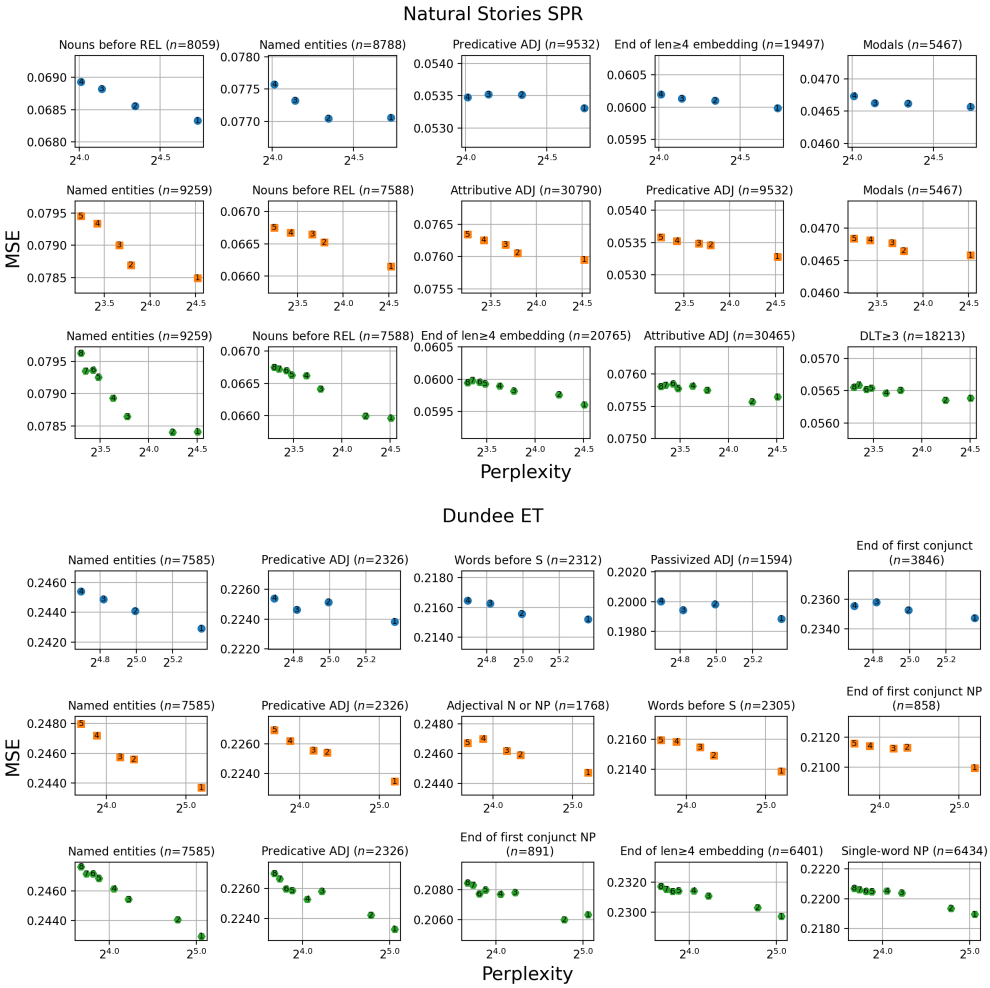

Figure 2: Top five subsets with largest differences in MSE of Natural Stories (top) and Dundee data (bottom).

[1] S. Black et al. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. *Zenodo*, 2021.
[2] R. Futrell et al. The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *LREC*, 55:63–77, 2021.
[3] A. Goodkind and K. Bicknell. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proc. CMCL*, pages 10–18, 2018.
[4] J. Hale et al. Finding syntax in human encephalography with beam search. In *Proc. ACL*, pages 2727–2736, 2018.
[5] A. Kennedy et al. The Dundee Corpus. In *Proceedings of the 12th European Conference on Eye Movement*, 2003.
[6] B.-D. Oh et al. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5(777963), 2022.
[7] A. Radford et al. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019.
[8] C. Shain et al. Deep syntactic annotations for broad-coverage psycholinguistic modeling. In *Workshop on Linguistic and Neuro-Cognitive Resources*, 2018.
[9] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.
[10] N. J. Smith and R. Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–319, 2013.
[11] E. G. Wilcox et al. On the predictive power of neural language models for human real-time comprehension behavior. In *Proc. CogSci*, pages 1707–1713, 2020.
[12] S. Zhang et al. OPT: Open pre-trained Transformer language models. *arXiv preprint*, 2022.