# Analyzing the Effects of Temperature-Scaled Surprisal for Subword Reading Times

**Sneha Chetani**[1], **Iza Škrjanec**[1], **Vera Demberg**[1]
[1]Saarland University, Germany
snehachetani45@gmail.com, {skrjanec, vera}@coli.uni-saarland.de

Surprisal [1, 2] measures predictability in context and has been accepted as a metric of per-word human processing effort [3-5] with surprisal estimated using large neural language models (LMs). Recent work indicates LMs with a higher quality (and a lower perplexity) do not necessarily correspond to a better fit to human reading times (RT) [6] and that they likely underestimate the surprisal of low-frequency words [7]. A way of adjusting for this is to use temperature scaling of LM outputs to make the probability distribution less certain. Liu et al. [8] show that surprisal estimates are closer to human reading times when surprisal is temperature-scaled. Additionally, they show this benefit is driven by words that are split into multiple subwords. This poses the question whether LMs are overestimating the processing difficulty due to subword tokenization and how this overestimation explains why temperature scaling differentially impacts RT of standalone versus split subwords.

In our study, we model reading times during naturalistic reading and calculate surprisal with the small GPT2 LM. We use the Dundee eye-tracking corpus [9], but instead of using word-level gaze duration measures, we re-calculate the measures for each subword based on the GPT2 tokenizer. We consider the log-transformed total reading time of each subword and fit a baseline mixed-effects regression with subword surprisal and length, word and subword frequency, position as predictors, including a binary split-indicator of whether a subword stands alone or is rather a part of a word. The experimental model included surprisal from GPT2, which has been temperature-scaled. We explore the range between 1 and 10 for scalar values, where a larger value results in a less certain probability distribution over the vocabulary and a higher surprisal for most words. To establish the effect of temperature scaling, we compare the log-likelihoods of the base and experimental model in the delta log-likelihood metrics, where a larger value indicates a better fit above the baseline.

Our results show that the split-indicator has a main effect above and beyond the length and frequency of the subword: words that are split are read more slowly. Our results also reveal that the effect of temperature scaling is not equally beneficial for all subwords. As shown in Figure 1, delta log-likelihood values indicate that single-subword words and the first subword of a split word profit, as their predicted reading times are closer to observed times after their surprisal is increased via temperature scaling. This contrasts the result for subwords that are in the middle or at the end of a word (e.g. *can**'t***, *four-year**ly***). We suspect that these are high-predictability continuations of the first subword. Increasing their surprisal does not correspond to human processing effort.

We plan to extend the analysis to synthetic languages (such as German or Finnish) with longer compound words and richer inflection. The analyses will add to the discussion of cognitive plausibility of subword tokenizers [10, 11].

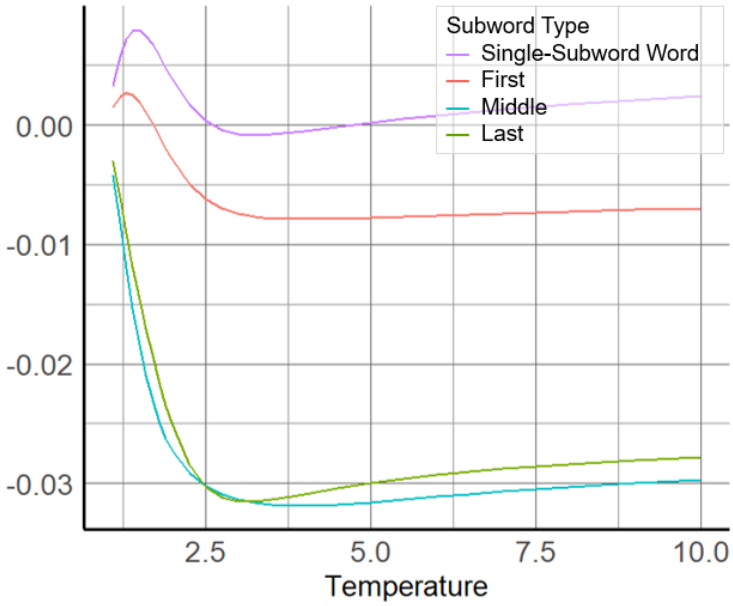


Figure 1: Effects of temperature scaling on delta log-likelihood values for different subword types across temperatures $T \in [1, 10]$.

**References** [1] Hale. Probabilistic Earley parser as a psycholinguistic model. NACL 2001. • [2] Levy. Expectation-based syntactic comprehension. Cognition 2008. • [3] Demberg, Keller. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. Cognition 2008. • [4] Smith, Levy. The effect of word predictability on reading time is logarithmic. Cognition 2013. • [5] Wilcox, Gauthier, Hu, Qian, Levy. On the predictive power of neural language models for human real-time comprehension behavior. CogSci 2020. • [6] Oh, Schuler. Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times?. ACL 2023. • [7] Oh, Yue, Schuler. Frequency Explains the Inverse Correlation of Large Language Models' Size, Training Data Amount, and Surprisal's Fit to Reading Times. EACL 2024. • [8] Liu, Škrjanec, Demberg. 2024. Temperature-scaling surprisal estimates improve fit to human reading times – but does it do so for the "right reasons"?. ACL 2024. • [9] Kennedy, Hill, Pynte. The Dundee corpus. European conference on eye movement 2003. • [10] Beinborn, Pinter. Analyzing Cognitive Plausibility of Subword Tokenization. EMNLP 2023. • [11] Nair, Resnik. Words, Subwords, and Morphemes: What Really Matters in the Surprisal-Reading Time Relationship?. EMNLP 2023.