Leveraging Eye-Tracking Data to Align Language Models with Human Repeated Reading Behavior

Recent studies have suggested that language models (LMs) and humans may employ similar mechanisms for structuring and recalling information from memory[1,2]. However, these apparent parallels have been challenged, with evidence showing divergence between LMs and humans in next-word prediction task[3,4]. In particular, Vaidya et al. (2023)[3] show that in repeated text presentation, in stark contrast to humans, LMs correctly predict the next word in nearly all cases. They further propose to address this discrepancy by fine-tuning the model's attention on human cloze data.

In this work, we extend Vaidya et al. (2023) from cloze to reading times during natural reading and propose a new approach for aligning multiple LMs with human reading behavior in repeated reading. We use the OneStop dataset, which includes 360 L1 English participants reading Guardian articles, where 20% of the material is presented to participants for a second time. We first use this dataset to show a linear relationship between reading time and surprisal in repeated reading. Building on this result, we directly optimize model probability differences between first and repeated reading using the observed reading time differences, thus aligning the model's behavior with human reading patterns.

To further increase the alignment of language model probabilities to human reading times, we optimize model attention patterns in repeated text presentation. Specifically, we find that induction heads, which are considered crucial for in-context learning capabilities[4,5], directly reduce the model's surprisal when encountering repeated text. To address this, we propose fine-tuning these heads using parameter addition methods, e.g. LoRA, to better match reading times.

We evaluate the fine-tuned models on three main axes. First, we measure the models fit to human reading times in repeated reading. Next, we measure the model's fit to human reading times in first reading. Finally, we assess our method's performance on in-context-learning benchmarks to ensure that modifications to the induction heads do not impair the model's overall performance on these tasks[5]. By leveraging eye-tracking data and multiple alignment techniques, our work aims to bridge the gap between LMs and human reading behavior.

(1) Shain, C.; Meister, C.; Pimentel, T.; Cotterell, R.; Levy, R. P. Large-Scale Evidence for Logarithmic Effects of Word Predictability on Reading Time. **2024**. https://doi.org/10.31234/osf.io/4hyna.

(2) Wilcox, E. G.; Gauthier, J.; Hu, J.; Qian, P.; Levy, R. On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior. arXiv June 2, 2020. http://arxiv.org/abs/2006.01912 (accessed 2024-03-19).

(3) Vaidya, A.; Turek, J.; Huth, A. Humans and Language Models Diverge When Predicting Repeating Text. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*; Jiang, J., Reitter, D., Deng, S., Eds.; Association for Computational Linguistics: Singapore, 2023; pp 58–69. https://doi.org/10.18653/v1/2023.conll-1.5.