Effects of Recency Bias on Transformers' Predictions of Reading Times

Christian Clark (clark.3664@osu.edu),[1] Byung-Doh Oh,[2] William Schuler[1] [1]Ohio State, [2]NYU

A range of recent studies have shown a strong fit between surprisal estimates from Transformer language models (LMs) and human reading times [19, 11]. These findings raise the question of whether Transformers internally process language in a way that mirrors human comprehension. Although Transformers' attention mechanism bears some similarity to psycholinguistic models of cue-based retrieval [15, 12, 18], the fact that Transformers can access lossless representations of hundreds or thousands of previous tokens seems unrealistic for modeling human memory.

We therefore experiment with altering Transformers' attention mechanism to include a *recency bias*, which assigns a higher weight to tokens which are closer to the current token being processed. In a standard Transformer, the attention score of the ith token is calculated as *softmax* $\left(\mathbf{K}^\top \mathbf{q}_i / \sqrt{d}\right)$, where $\mathbf{q}_i \in \mathbb{R}^d$ is the current token's query vector, $\mathbf{K} \in \mathbb{R}^{d \times i}$ contains the first i keys, and $\sqrt{d}$ is a scaling factor. We test two methods to incorporate a recency bias by adding a vector $\mathbf{b}_i \in \mathbb{R}^i$ to the attention scores. The first method, from recent work by de Varda and Marelli [4], sets each $\mathbf{b}_i[j] = e^{-\lambda(i-j)}$ for $j \in \{1, \ldots, i\}$ and defines the ith token's attention score as *softmax* $\left(\alpha \mathbf{b}_i + (1-\alpha)\mathbf{K}^\top \mathbf{q}_i / \sqrt{d}\right)$, where λ and α are hyperparameters. The second method, ALiBi [14], sets each $\mathbf{b}_i[j] = m \cdot (j - i)$ for $j \in \{1, \ldots, i\}$ and defines the ith token's attention score as *softmax* $\left(\mathbf{b}_i + \mathbf{K}^\top \mathbf{q}_i / \sqrt{d}\right)$, where the slope hyperparameter m is specific to each attention head (the part of a Transformer layer that attends to different preceding tokens).

To evaluate the effects of these recency biases on surprisal estimates, a set of Transformer LMs was trained from scratch on the first $\sim$2 billion tokens of a large corpus in English [7]. Separate LMs were trained containing the de Varda and Marelli bias and ALiBi; an LM with no recency bias was also trained. The LMs' architecture followed Pythia LMs [2]; the specific hyperparameters and training data amount were based on the best-performing model in a previous reading time study [13]. Surprisal estimates from each LM were calculated across six English self-paced reading (SPR) and eye-tracking (ET) corpora [17, 6, 5, 3, 8, 10]. Following [4], surprisal estimates were also calculated from LMs that included a recency bias at inference time but not during training.

Subsequently, linear mixed-effects (LME; [1]) models were fit to reading times from each corpus to find the increase in regression model log-likelihood (ΔLogLik) due to including each surprisal predictor over a model with the following baseline predictors: word length, word position, unigram surprisal, and whether the previous word was fixated (ET corpora only). LME models for SPR corpora included by-subject random slopes for word position, word length, and surprisal of current and previous word, and a by-subject random intercept. LME models for ET corpora included random slopes for word position and surprisal of current word, and a by-subject random intercept.

As Table 1 shows, surprisal estimates from the LM including ALiBi during both training and inference increased ΔLogLik by a margin of $\sim$352 compared to surprisal from an LM with no recency bias. A permutation test showed that this improvement was significant ($p < 0.001$). However, no improvement over the LM without bias was observed from models including the de Varda and Marelli bias, nor from a model with ALiBi at inference time only. A second experiment (Figure 1) revealed that a simplified version of ALiBi with uniform attention head slopes generally failed to perform better than the LM without bias, suggesting that mixed slopes are important for ALiBi's success. An additional analysis (Figure 2) provided evidence that mixed slopes in ALiBi may enable different attention heads to track different semantic dependencies. These results suggest that incorporating varying rates of memory decay, rather than a single decay parameter (e.g., [9]), may be helpful for developing humanlike models of language processing; more generally, this work opens up possibilities for modeling memory effects using broad-coverage corpora.

Recency Bias	Training	Inference	ΔLogLik (↑)
None	—	—	3003
dVM [4]	✗	✓	2988
dVM [4]	✓	✓	2948
ALiBi [14]	✗	✓	2926
ALiBi [14]	✓	✓	**3355**

Table 1: Improvements in log likelihood (ΔLogLik) from regression models that include surprisal estimates from LMs with no recency bias (None), the de Varda and Marelli recency bias (dVM), or ALiBi. The middle two columns mark whether the indicated recency bias was included during both training and inference or only during inference. ΔLogLik scores were taken on held-out partitions of the six SPR and ET corpora and are aggregated over all corpora.

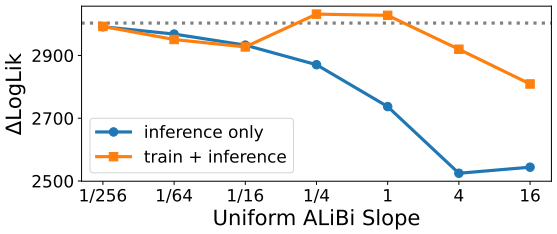

Figure 1: Aggregated improvements in log likelihood from a simplification of ALiBi in which all attention heads have the same slope (marked on the x-axis). Blue circles are LMs that included the bias at inference time only; orange squares are LMs that included the bias during both training and inference. The gray dashed line shows ΔLogLik from an LM with no recency bias.

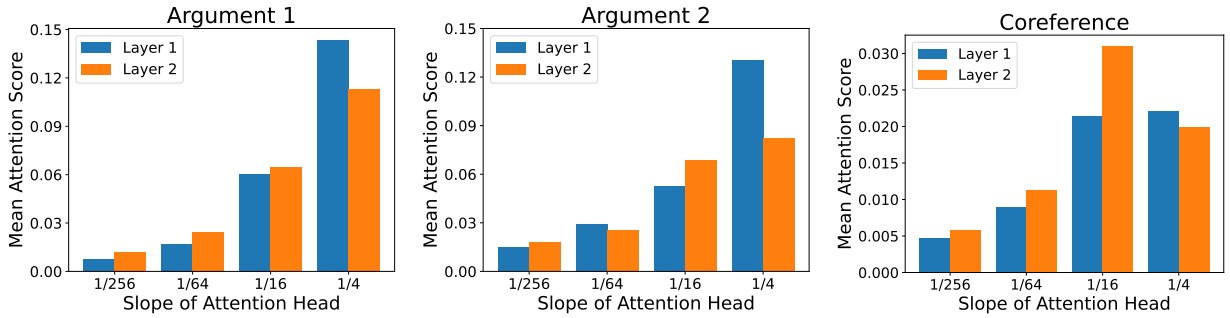

Figure 2: Mean attention scores for three types of semantic dependencies, across the four attention heads and two layers of an LM with mixed ALiBi slopes. These scores were computed on the Natural Stories corpus [6], with argument and coreference information extracted from existing syntactic annotations [16]. Argument 1 and argument 2 dependencies typically correspond to the relationship between a verb and its subject and its direct object, respectively. Mean attention score is the average attention weight assigned by the head word of a dependency to a preceding dependent word. These results suggest that the LM tends to make greater use of attention heads with smaller slopes (i.e., less decay) for coreference dependencies and larger slopes for argument attachment.

References

[1] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[2] S. Biderman et al. Pythia: A suite for analyzing large language models across training and scaling. In *Proc. ICML*, 2023.

[3] U. Cop et al. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 2017.

[4] A. G. de Varda and M. Marelli. Locally biased transformers better align with human reading times. In *Proc. CMCL*, 2024.

[5] S. L. Frank et al. Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, 2013.

[6] R. Futrell et al. The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *LREC*, 2021.

[7] L. Gao et al. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv*, 2020.

[8] A. Kennedy et al. The Dundee Corpus. In *Proc. ECEM*, 2003.

[9] R. L. Lewis and S. Vasishth. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 2005.

[10] S. G. Luke and K. Christianson. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 2018.

[11] D. Merkx and S. L. Frank. Human sentence processing: Recurrence or attention? In *Proc. CMCL*, June 2021.

[12] B.-D. Oh and W. Schuler. Entropy- and distance-based predictors from GPT-2 attention patterns predict reading times over and above GPT-2 surprisal. In *Proc. EMNLP*, 2022.

[13] B.-D. Oh and W. Schuler. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *EMNLP Findings*, 2023.

[14] O. Press et al. Train short, test long: Attention with linear biases enables input length extrapolation. In *Proc. ICLR*, 2022.

[15] S. H. Ryu and R. L. Lewis. Accounting for agreement phenomena in sentence comprehension with Transformer language models: Effects of similarity-based interference on surprisal and attention. In *Proc. CMCL*, 2021.

[16] C. Shain et al. Deep syntactic annotations for broad-coverage psycholinguistic modeling. In *Workshop on Linguistic and Neuro-Cognitive Resources*, 2018.

[17] N. J. Smith and R. Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, 2013.

[18] W. Timkey and T. Linzen. A language model with limited memory capacity captures interference in human sentence processing. In *EMNLP Findings*, 2023.

[19] E. G. Wilcox et al. On the predictive power of neural language models for human real-time comprehension behavior. In *Proc. Cognitive Science Society*, 2020.