What Transformer Attention Mechanism Provides the Best Fit for Human Reading Times?

Lanni Bu (lb1437@georgetown.edu),[1] Xiulin Yang,[1], Christian Clark[2], Alex Warstadt[3], Ethan Wilcox[1]
[1]Georgetown, [2]OSU, [3]UCSD

Summary Transformer-based language models (LMs) are used widely as models of human language processing. But their memory mechanisms differ markedly from those in humans, as their *attention heads* allow lossless access to a full preceding context. We hypothesize that adding biases into LMs limiting their attention can improve their ability to model human language processing data. While previous work has explored individual biases [14, 1] in this context, we conduct a systematic assessment of seven variants to the vanilla attention mechanism. We categorize variants as either corresponding to distance-based or interference-based constraints on memory, and find that the latter leads to improved ability for modeling human reading times.

Methods All of our proposed interventions limit Transformers by changing their attention mechanism. The attention mechanism is learned as part of training and determines how much each previous word contributes to predicting the next word. It has been argued to resemble proposed memory architectures in humans [11]. We implement seven variants of attention (Table 1), which we group into two broad buckets: *Distance-based mechanisms* constrain attention based on the distance between words, restricting access to distant context. *Interference-based mechanisms* constrain attention between two words based on lexical features of the words that occur between them. Our distance-based attention variants include (1) ALiBi [10], which adds a linear distance penalty to attention scores; (2) Dynamic ALiBi [9], which gradually relaxes this penalty during training; and (3-5) N-gram Attention, which restrict attention to fixed windows of 2, 3, or 5 tokens. Our interference-based mechanisms include (6) Forgetting Gate Attention [7], similar to ALiBi but data-dependent: each token computes a learned forget probability based on its content, and these accumulate to down-weight attention to earlier positions; and (7) Stick-Breaking Attention [13], which starts with one budget of attention and distributes it to past words according to their content similarity so that giving more to an important earlier word leaves less for the later ones. We also include, as a baseline, (8) Vanilla Transformer [15], in which attention scores between words are learned during training without any bias being imposed. **Training Procedure:** We train models from scratch on three corpora: (1) BabyLM-10M and (2) BabyLM-100M [16], both human-scale corpora of developmentally plausible input; and (3) a 2-billion-word subset of The Pile [5], a large-scale corpus of web text. For each corpus, we train 2-layer (4 heads) and 4-layer (6 heads) models. **Evaluation:** We evaluate on six English naturalistic reading datasets, including two self-paced reading (Brown [12], Natural Stories [4]); three eye-tracking (GECO [2], Dundee [6], and Provo [8]); and one with both (UCL [3]). We measure a model's psychological predictive power by assessing how well a linear regression fit with its surprisal values predicts word-by-word reading times over a baseline model. This difference is reported as Delta Log-Likelihood (Δ_{llh}), where higher values are taken to indicate better psychological fit of the proposed attention mechanism. For regressions, baseline features include word length, sentence position, unigram surprisal, and previous word's unigram surprisal; eye-tracking models additionally control for whether the previous word was fixated. All regressions include by-subject and by-item random effects.

Results Figure 1 reports the summed Δ_{llh} for each model across all reading time datasets. We test whether Δ_{llh} is significantly above baseline with a paired permutation test. We find that interference-based attention mechanisms consistently achieve higher alignment with human reading times. Although previous work [1] has found that ALiBi produces higher Δ_{llh} than the vanilla model, we find that this is sensitive to hyperparameters, and only replicate this result for its original settings. Figure 2 shows the relationship between LMs' ability to predict text (measured in *perplexity*) and Δ_{llh}. Lower perplexity is correlated with higher Δ_{llh} ($r = -.69$, $p < .001$), however, Forgetting Gate and Stick-breaking consistently perform above the trend line, suggesting they capture aspects of human processing beyond what perplexity alone predicts.

Table 1: Attention mechanisms, their formulations, and types of bias. A_{ij}: attention from position i to j; $\mathbf{q}_i$, $\mathbf{k}_j$: query and key vectors; d: hidden dimension; m_h, m_t: head/epoch-specific slopes with decay rate r and initial slope m_0; $f_k = \sigma(\mathbf{w}_f^\top \mathbf{x}_k + b_f)$: content-dependent forget gate; g_{ij}: sigmoid-transformed attention scores.

Mechanism	**Formula**	**Type of bias***
Forgetting Gate	$A_{ij} = \text{softmax}_j\left(\frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d}} + \sum_{k=j+1}^{i} \log f_k\right), \quad B_{ij} = \sum_{k=j+1}^{i} \log f_k$	Interference
Stick-breaking	$A_{ij} = g_{ij} \prod_{j<k<i}(1 - g_{ik}), \quad g_{ij} = \sigma\left(\frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d}}\right)$	Interference
N(2,3,5)-gram	$A_{ij} = \mathbf{1}[i - j < N]$	Strict capacity limit
Dynamic ALiBi	$A_{ij} = \text{softmax}_j\left(\frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d}} + m_t(i-j)\right), \quad B_{ij} = m_t(i-j), \quad m_t = m_0 \cdot r^t$	Less-is-more
ALiBi	$A_{ij} = \text{softmax}_j\left(\frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d}} + m_h(i-j)\right), \quad B_{ij} = m_h(i-j)$	Recency Effect
Vanilla	$A_{ij} = \text{softmax}_j\left(\frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d}}\right), \quad B_{ij} = 0$	Unlimited capacity

Figure 1: Total Δ_{llh} and rank across training settings. Outline boxes indicate mechanism type: solid = baseline, dashed = distance-based, dotted = interference-based. The *s indicate the significance of a permutation test ($p < .05$) against vanilla Δ_{llh} scores within each setting.

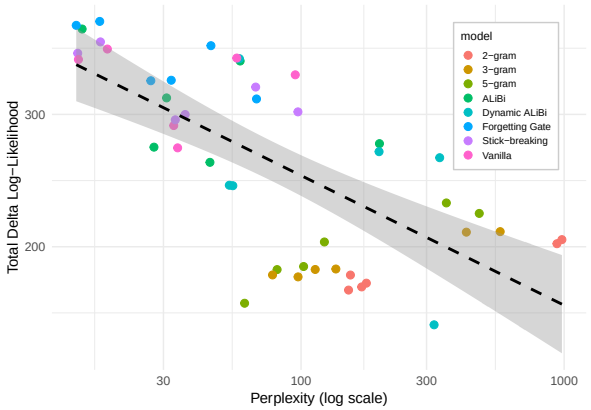

Figure 2: Perplexity vs. Δ_{llh} ($r = -.69$, $p < .001$). Lower perplexity (better language modeling) is associated with higher ΔLL (better psychometric fit), but the correlation is imperfect, suggesting attention mechanisms affect both differently.

References

[1] C. Clark, B.-D. Oh, and W. Schuler. Linear recency bias during training improves transformers' fit to reading times. In *Proc. COLING*, 2025.
[2] U. Cop et al. Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 2017.
[3] S. L. Frank et al. Reading time data for evaluating broad-coverage models of english sentence processing. *Behavior Research Methods*, 2013.
[4] R. Futrell et al. The natural stories corpus: A reading-time corpus of english texts containing rare syntactic constructions. *LREC*, 2021.
[5] L. Gao et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv*, 2020.
[6] A. Kennedy et al. The dundee corpus. In *Proc. ECEM*, 2003.
[7] Z. Lin et al. Forgetting transformer: Softmax attention with a forget gate. *arXiv*, 2025.
[8] S. G. Luke and K. Christianson. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 2018.
[9] M. Mita, R. Yoshida, and Y. Oseki. Developmentally-plausible working memory shapes a critical period for language acquisition. *arXiv*, 2025.
[10] O. Press et al. Train short, test long: Attention with linear biases enables input length extrapolation. In *Proc. ICLR*, 2022.
[11] 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.
[12] N. J. Smith and R. Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, 2013.
[13] S. Tan et al. Scaling stick-breaking attention: An efficient implementation and in-depth study. In *Proc. ICLR*, 2025.
[14] W. Timkey and T. Linzen. A language model with limited memory capacity captures interference in human sentence processing. In *EMNLP Findings*, 2023.
[15] A. Vaswani et al. Attention is all you need. In *Proc. NeurIPS*, 2017.
[16] A. Warstadt et al. Findings of the babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proc. BabyLM Challenge, CoNLL*, 2023.