Linear Recency Bias During Training Improves Transformers' Fit to Reading Times

Christian Clark
The Ohio State University
clark.3664@osu.edu

Byung-Doh Oh
New York University
oh.b@nyu.edu

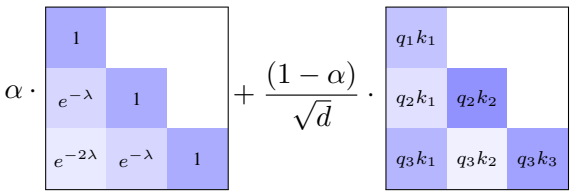
William Schuler
The Ohio State University
schuler.77@osu.edu

Abstract

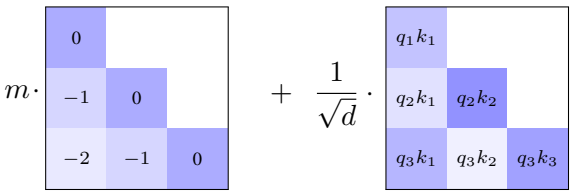
Recent psycholinguistic research has compared human reading times to surprisal estimates from language models to study the factors shaping human sentence processing difficulty. Previous studies have shown a strong fit between surprisal values from Transformers and reading times. However, standard Transformers work with a lossless representation of the entire previous linguistic context, unlike models of human language processing that include memory decay. To bridge this gap, this paper evaluates a modification of the Transformer model that uses ALiBi (Press et al., 2022), a recency bias added to attention scores. Surprisal estimates from a Transformer that includes ALiBi during training and inference show an improved fit to human reading times compared to a standard Transformer baseline. A subsequent analysis of attention heads suggests that ALiBi's mixture of slopes—which determine the rate of memory decay in each attention head—may play a role in the improvement by helping models with ALiBi to track different kinds of linguistic dependencies.

1 Introduction

Expectation-based theories of human sentence processing (Hale, 2001; Levy, 2008) posit that the difficulty of comprehending a word is proportional to its surprisal, i.e. negative log probability, given the preceding context. This creates a natural interface between the task of language modeling, which estimates word probabilities, and modeling human sentence processing. A range of studies (Goodkind and Bicknell, 2018; Wilcox et al., 2020; Merkx and Frank, 2021) have compared surprisal estimates from families of language models (LMs) including n-gram models, recurrent neural networks, and Transformers, generally showing a strong fit between Transformer surprisal and psychometric data such as reading times.



(a) de Varda and Marelli (2024)



(b) Press et al. (2022)

Figure 1: Illustration of two recency bias techniques tested in this work and defined in Equations (2) and (4). Bias matrices (left) are added to raw attention scores (right). Darker colors indicate higher scores. Hyperparameters α, λ, and m control the strength of the bias, and $\sqrt{d}$ scales the $q_i k_j$ values.

The predictive power of surprisal estimates from Transformers raises the question of whether these models internally process language in a way that mirrors human language comprehension. Indeed, the attention mechanism at the heart of the Transformer architecture bears a tantalizing similarity to models of comprehension based on cue-based retrieval (Ryu and Lewis, 2021; Oh and Schuler, 2022; Timkey and Linzen, 2023). However, the fact that a Transformer's context window—typically including hundreds or thousands of tokens—is fully retained in memory when predicting subsequent tokens seems unrealistic for modeling human memory; human working memory has a small capacity, and retrieval from longer-term memory is prone to decay and interference effects that Transformers do not explicitly model.[1]

[1] Ryu and Lewis (2021) do report facilitatory interference effects in GPT-2 (Radford et al., 2019), wherein the presence of a distractor noun decreases surprisal at a target verb in un-

We therefore consider altering Transformers' attention mechanism to include a *recency bias* which upweights keys that are closer to a given query. Such a bias brings Transformers more in line with cognitive models that include some notion of decay or lossy context (Baddeley and Hitch, 1974; Baddeley, 2003; Lewis and Vasishth, 2005; Futrell et al., 2020). The particular form of recency bias we test is based on ALiBi (Attention with Linear Biases), which was originally developed by Press et al. (2022) as a method for helping Transformers to extrapolate beyond their context length. Experiments on psycholinguistic corpora with reading times show that adding this recency bias to a Transformer during training and inference results in improved surprisal estimates compared to a standard Transformer. Additional experiments show that ALiBi's mixture of slopes (decay rates specific to each attention head) plays an important role in this improvement, and suggest that the varied slopes may enable different attention heads to track different linguistic dependencies. Such results could have interesting implications for the implementation of memory decay in models of human language comprehension.

2 Related Work

While neural LMs have been tested for some time as expectation-based models of human sentence processing (e.g. Wilcox et al., 2020; Oh et al., 2022), recent work has more specifically examined how these models' memory representations relate to processing difficulty.

One line of research draws a connection between the self-attention of Transformers (Vaswani et al., 2017) and cue-based retrieval models of sentence processing (e.g. Lewis et al., 2006)[2] and aims to derive measures from Transformer LMs that align with real-time processing behavior. Ryu and Lewis (2021) define an attention entropy metric that quantifies the diffuseness of the attention weights over previous tokens, and shows patterns that are consistent with similarity-based interference observed during the processing of subject-verb agreement. Oh and Schuler (2022) propose a normalized attention entropy metric to control for the number of tokens in the previous context, as well as other predictors that capture the distance between attention weights of consecutive time steps, which are shown to be predictive of naturalistic reading times over a surprisal baseline. Timkey and Linzen (2023) train an LM based on a modified version of the Simple Recurrent Network (Elman, 1991) with one self-attention head that aggregates representations of previous words, which yields attention weights and surprisal that are sensitive to agreement and semantic attraction effects.

A second line of research is concerned with the interaction between memory-based effects and expectation-based effects, and broadly falls under the framework of lossy-context surprisal (Futrell et al., 2020). Recent work has focused on "corrupting" the lossless representations of pretrained Transformer LMs and evaluating the quality of resulting surprisal estimates. Hahn et al. (2022) implement a probabilistic erasure of words based on their frequency and position within the sentence, and show that the resulting surprisal estimates accurately predict the increased reading times at the main verb of deeply embedded sentences. Kuribayashi et al. (2022) constrain LMs' access to the previous context and report improvements in modeling naturalistic reading times of English and Japanese text. de Varda and Marelli (2024) incorporate a softer recency bias into the attention weights of Transformers for surprisal estimates that are more predictive of naturalistic reading times.

While there have been some promising results for modeling interference effects with neural LMs (Ryu and Lewis, 2021; Timkey and Linzen, 2023), this second line of work using simpler recency-based models may provide broadly useful predictions for modeling more naturalistic comprehension. Additionally, as the vast majority of these results are based on post-hoc modifications to pretrained LMs like GPT-2 (Radford et al., 2019), it remains to be seen how various constraints like recency biases influence LMs during training. This work aims to address these gaps by newly training a set of LMs with various recency biases in a controlled setting, and evaluating their surprisal estimates across a wide range of naturalistic reading-time corpora.

3 Background

This section gives an overview of standard Transformer attention and how it is modified to incorpo-

grammatical sentences. However, these effects seem unlikely to provide a general mechanism for humanlike forgetting in Transformers.

[2]This is because the dot product of representations (i.e. Eqn. 1) is used in both models to quantify the degree of similarity (Merkx and Frank, 2021).

rate recency biases in the experiments that follow.

3.1 Transformer Attention Scores

Within an autoregressive Transformer layer (Vaswani et al., 2017), the attention sublayer calculates attention scores based on the scaled dot product between the ith query $\mathbf{q}_i \in \mathbb{R}^d$ and the first i keys $\mathbf{K} \in \mathbb{R}^{d \times i}$:

$$\text{softmax}\left(\frac{\mathbf{K}^\top \mathbf{q}_i}{\sqrt{d}}\right), \tag{1}$$

where $i \in \{1, \ldots, N\}$, d is the dimension of the query and key, and N is the sequence length. The scaling factor $\sqrt{d}$ prevents the magnitude of the dot product from growing too large (Vaswani et al., 2017).

3.2 Recency Bias

The two recency bias techniques tested in this paper—namely, the method from de Varda and Marelli (henceforth "dVM bias") and ALiBi—are illustrated in Figure 1. Both techniques involve modifications to the raw scores in Equation (1), before the softmax is taken. The modifications have the effect of lowering attention scores to tokens more distant from the current query.

The dVM bias is implemented using a vector $\mathbf{b}_i \in \mathbb{R}^i$ such that $\mathbf{b}_i[j] = e^{-\lambda(i-j)}$ for $j \in \{1, \ldots, i\}$. The hyperparameter λ determines the rate of decay. A weighted combination is taken between the bias vector and the raw attention scores:

$$\text{softmax}\left(\alpha \mathbf{b}_i + (1-\alpha)\frac{\mathbf{K}^\top \mathbf{q}_i}{\sqrt{d}}\right). \tag{2}$$

The additional hyperparameter α determines the relative weight of the bias and original attention scores.

ALiBi uses a bias vector $\mathbf{b}'_i \in \mathbb{R}^i$ such that $\mathbf{b}'_i[j] = m \cdot (j - i)$ for $j \in \{1, \ldots, i\}$. The hyperparameter m is a slope determining the rate of decay. Slopes are defined separately for each attention head in a layer. Press et al. (2022) report optimal performance in input length extrapolation from setting the slope of head number h out of H total heads as

$$m_j = 2^{-h \cdot 2^{(-\log_2 H + 3)}} = 2^{-8h/H}. \tag{3}$$

ALiBi bias is directly added to the raw attention scores:

$$\text{softmax}\left(\mathbf{b}'_i + \frac{\mathbf{K}^\top \mathbf{q}_i}{\sqrt{d}}\right). \tag{4}$$

Because the softmax operation involves exponentiation, the linear bias of ALiBi translates to an exponential decay in the final attention scores, consistent with the shape of the decay found in models like ACT-R (Lewis and Vasishth, 2005). However, the terms in the dVM bias become doubly exponential after the softmax;[3] we are not aware of any existing cognitive models that use this form of decay.

4 Experiment 1: Recency Bias During Inference

Experiment 1 considers the effect of incorporating a recency bias into an already trained Transformer LM. Under this approach, the bias is included only at inference time, not during training. This is the technique used by de Varda and Marelli (2024).

We evaluate surprisal estimates from a model with the dVM bias and a model with ALiBi. These are compared against an LM with no recency bias.[4]

4.1 Language Model

The design of the base language model used in this and subsequent experiments follows Pythia language models (Biderman et al., 2023). Pythia LMs are autoregressive, decoder-only Transformer models that vary primarily in their capacity and quantity of training data. The main distinctions between Pythia LMs and other Transformer-based LM families are that Pythia LMs parallelize the computation of the self-attention sublayer and the feedforward neural network, and do not use shared parameters for the embedding and projection matrices.

The specific model configuration was chosen based on optimal settings found in previous work that compared reading-time estimates from Pythia-style models with varying capacities and training data amounts (Oh and Schuler, 2023a). This model configuration uses two layers, four attention heads, and an embedding size of 256, with a total parameter count of 27,335,680. The training data amount was also determined based on the same work; it includes the first 1,000 batches of the Pile (Gao

[3] The softmax operation assigns the final attention score $\frac{e^{x_j}}{\sum_k e^{x_k}}$ to the jth key with raw score x_j. If the raw score is a weighted sum of the dVM bias term $e^{-\lambda(i-j)}$ and the scaled dot product between the query and key, then the final attention score will be proportional to $e^{e^{-\lambda(i-j)}}$.

[4] Code and instructions for replicating this paper's experiments are available at https://github.com/christian-clark/recency-bias.

et al., 2020), a large collection of English-language datasets.[5] Each batch contains 1,024 examples with a sequence length of 2,048, for a total size of 2,097,152 tokens per batch and 2,097,152,000 tokens in all of training. While the model size and the amount of training data is much smaller than contemporary standards for Transformer LMs, this configuration was found to achieve strong fit to reading times (Oh and Schuler, 2023a) and has the additional benefit of allowing quick training of a variety of LMs.

Also following Pythia LMs, the base LM for this experiment uses rotary positional embeddings (Su et al., 2024) and a context window of 2,048 tokens. See Appendix A for additional training details.

4.2 Recency Biases

The LM with the dVM bias uses the hyperparameters $\lambda = 82.86$ and $\alpha = 0.37$, which were the optimal values found by de Varda and Marelli (2024) in a grid search on the Provo corpus (Luke and Christianson, 2018) with GPT2-small as the base LM.[6]

The LM with ALiBi uses the slopes $1/4$, $1/16$, $1/64$, and $1/256$ for the four attention heads in each layer, following Equation (3).

4.3 Corpora

This experiment used reading times from six self-paced reading (SPR) and eye-tracking (ET) corpora, which are described below:

- Brown (Smith and Levy, 2013): SPR times from 35 subjects that read 13 English passages from the Brown Corpus (Kučera and Francis, 1967) consisting of a total of 7,188 words.
- Natural Stories (Futrell et al., 2021): SPR times from 181 subjects that read 10 naturalistic English stories consisting of a total of 10,256 words.
- UCL (Frank et al., 2013): SPR times from 117 subjects and fixation durations from 48 subjects that read isolated sentences extracted from three novels written by aspiring authors, consisting of a total of 4,957 words.

[5]The full Pile collection comprises approximately 300 billion tokens.

[6]The present experiments used these hyperparameters for consistency with de Varda and Marelli (2024), but it is possible that other values of λ and α could perform better on other corpora or LMs.

Corpus/Measure	Fit	Exploratory	Held-out
Brown SPR	59,292	29,671	30,157
Natural Stories SPR	384,905	192,772	192,425
UCL SPR	139,300	70,239	69,753
UCL FP	20,428	10,281	10,310
UCL GP	20,428	10,281	10,310
GECO FP	144,850	72,468	72,574
GECO GP	144,850	72,468	72,574
Dundee SP	155,483	77,809	77,101
Dundee FP	98,115	48,598	48,794
Dundee GP	98,115	48,598	48,794
Provo SP	91,032	45,654	45,404
Provo FP	52,959	26,539	26,640
Provo GP	52,960	26,539	26,640
Total	1,462,717	731,917	731,476

Table 1: Number of observations in each partition of each reading-time corpus.

- GECO (Cop et al., 2017): Fixation durations from 14 monolingual subjects that read the English version of novel *The Mysterious Affair at Styles* (Christie, 1920) that consists of 13 chapters and 56,441 words.
- Dundee (Kennedy et al., 2003): Fixation durations from 10 subjects that read 67 English newspaper editorials consisting a total of 51,501 words.
- Provo (Luke and Christianson, 2018): Fixation durations from 84 subjects that read 55 short English passages ranging between news articles, science magazines, and works of fiction consisting a total of 2,746 words.

For the SPR datasets, the by-word reading times were filtered to exclude those of sentence-initial and -final words and those shorter than 100 ms or longer than 3000 ms. Additionally, the Natural Stories data from subjects who answered fewer than four comprehension questions correctly and the UCL SPR data from sentence-level trials with incorrect answers to comprehension questions were removed.

For the ET datasets, the by-word scan path (SP), first-pass (FP), and go-past (GP) durations were analyzed.[7] These datasets were filtered to remove data points for unfixated words, words following saccades longer than four words, and words at starts and ends of sentences and documents. For the Dundee Corpus (Kennedy et al., 2003) that further provides annotations of positions within lines

[7]Refer to Appendix B for their definitions. The SP duration could not be calculated for the GECO and UCL corpora that do not provide raw eye fixation durations.

and screens, data points corresponding to words at starts and ends of lines and screens were also excluded.

Prior to regression modeling, all datasets were split into fit, exploratory, and held-out partitions of roughly 50%, 25%, and 25% of data points respectively. This partitioning was conducted based on the sum of the subject index and the sentence index[8] in order to ensure that all data points from a subject reading a particular sentence were kept intact in a given partition. The fit partition was used to fit the regression models, and all results are reported on the exploratory partition. The held-out partition was reserved for statistical significance testing, and its use was kept to a minimum to minimize the need for multiple-trials correction. The final number of observations in each partition of each corpus is summarized in Table 1.

4.4 Surprisal Calculation

Each passage of the reading-time corpora described in Section 4.3[9] was tokenized with Pythia LMs' byte-pair encoding (Sennrich et al., 2016) tokenizer and provided as input to each LM to calculate surprisal predictors. In cases where each passage did not fit into a single context window of 2,048 tokens, the second half of the current context window was used to condition the surprisal of the remaining tokens.

A common design choice in subword tokenizers such as that of Pythia LMs is to prepend the whitespace character to tokens, such that they have leading whitespaces. However, if word probabilities are calculated with leading whitespaces (e.g. $\mathsf{P}(\textit{car} \mid \textit{I clean the})$ calculated as $\mathsf{P}(\texttt{␣car} \mid \texttt{I ␣clean ␣the})$), the sum over all word probabilities can exceed one, as the end of the word is not explicitly marked.[10] Therefore, following recent work (Oh and Schuler, 2024; Pimentel and Meister, 2024), word-level probabilities were calculated with trailing whitespaces by factoring the probability of each whitespace and re-allocating it to its preceding token (e.g. $\mathsf{P}(\textit{car} \mid \textit{I clean the})$ calculated as $\mathsf{P}(\texttt{car␣} \mid \texttt{I␣ clean␣ the␣})$), which ensures that the sum over all word probabilities equals one.

4.5 Linear Mixed-Effects Modeling

This experiment fit a set of linear mixed-effects (LME; Bates et al., 2015) regression models to evaluate the influence of different recency biases on Transformer surprisal's fit to human reading times. Following previous work (e.g. Oh and Schuler, 2023b; Wilcox et al., 2023b; Shain et al., 2024), the increase in regression model log likelihood (ΔLogLik) due to including a surprisal predictor over a common baseline regression model was calculated on the exploratory partition of each dataset.

The baseline predictors were word length in characters, index of word position within each sentence, unigram surprisal (both SPR and ET corpora), as well as a whether the previous word was fixated (ET corpora only). Unigram surprisal was estimated using the KenLM toolkit (Heafield et al., 2013) with default smoothing hyperparameters on the OpenWebText Corpus (Gokaslan and Cohen, 2019), which contains about 6.5 billion whitespace-delimited words. On top of these baseline regression models, surprisal at the current word and the previous word was included to capture lingering effects of the previous word (i.e. "spillover" effects; Rayner et al., 1983). Surprisal came from the LMs described in Sections 4.1 and 4.2. All regression models were fit to raw reading times; this assumption of a linear relationship between surprisal and reading times has recently received empirical support (Wilcox et al., 2023b; Xu et al., 2023; Shain et al., 2024).

The random effects structures of LME models were determined by starting with the maximal structure (Barr et al., 2013) and removing the least predictive effect iteratively until the models converged. The final random effects structure for LME models fit to 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. For ET corpora collected from a generally smaller number of subjects, the final random effects structure included random slopes for word position and surprisal of current word, a by-subject random intercept.

4.6 Results

The results of Experiment 1 are presented in Table 2, with ΔLogLik summed over all corpora

[8] If the sum of the subject and sentence number of a data point value is zero or one, modulo four, the data point was assigned to the fit partition; if this value is two, the data point was assigned to the exploratory partition; and if this value is three, the data point was assigned to the held-out partition.

[9] Each sentence was treated as a separate passage for the UCL corpus that contains isolated sentences.

[10] For example, if both $\mathsf{P}(\texttt{␣car} \mid \texttt{I ␣clean ␣the})$ and $\mathsf{P}(\texttt{pet} \mid \texttt{I ␣clean ␣the ␣car})$ have very high probability, the combined probabilities of "␣car" and "␣car pet" in the context "I ␣clean ␣the" can sum to more than one.

Recency Bias	ΔLogLik (↑)
None	**3003**
dVM (de Varda and Marelli, 2024)	2988
ALiBi (Press et al., 2022)	2926

Table 2: Aggregated likelihood results from Experiment 1. For each corpus in Table 1, an LME model was fit to find the improvement in log likelihood (ΔLogLik) from including a surprisal predictor using the indicated recency bias. Per-corpus results were then summed to find the total ΔLogLik. Per-corpus results are in Appendix C.

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

Table 3: Aggregated likelihood results from Experiment 2. The middle two columns mark whether the indicated recency bias was included during both training and inference or only during inference. Per-corpus results are in Appendix C.

listed in Table 1. Surprisal from LMs including either recency bias shows a worse fit to reading times than surprisal from the LM with no recency bias, although the dVM bias performs somewhat better than ALiBi.

We therefore do not generally replicate the improvements reported by de Varda and Marelli (2024). This may be due to differences in experimental setup; for instance, we predict per-subject reading times, while de Varda and Marelli average across subjects. It is also worth noting that the results in their study are somewhat variable, with recency bias leading to improvements in model fit in 3 out of 6 corpora (Fig. 4D in de Varda and Marelli 2024). We also observe variable results across corpora, with small increases in ΔLogLik in several corpora (Brown SPR, UCL SPR, UCL FP, UCL GP, Dundee GP, Provo FP, and Provo GP) offset by decreases in ΔLogLik elsewhere (Appendix C).

5 Experiment 2: Recency Bias During Training and Inference

Only introducing a recency bias during inference, as is done in Experiment 1, may be disadvantageous because it creates a mismatch between training and inference. In addition, the original experiments testing ALiBi on input length extrapolation (Press et al., 2022) include this bias during both training and inference. Experiment 2 therefore compares the LMs tested in the previous experiment with a similar set of LMs that include recency bias during both training and inference.

5.1 Procedures

Two new LMs were trained for this experiment: one including the dVM bias during training and one including ALiBi during training. Following Press et al. (2022), rotary positional embeddings were removed from the LMs trained with recency bias. Otherwise, LM training followed the setup described in Section 4.1, and LME models were fit using the corpora and procedures outlined in Sections 4.3 and 4.5. The newly trained models were compared against the dVM and ALiBi models with inference-only recency bias from Experiment 1.

5.2 Results

Table 3 presents the results, again aggregated over all corpora. Including recency bias during training, instead of inference only, slightly decreases the ΔLogLik from the dVM bias, but dramatically increases performance from ALiBi. When ALiBi is included throughout training and inference, it attains a better ΔLogLik than the LM with no recency bias by a margin of 352 (an improvement of roughly 12%), which is significant at $p < 0.001$ level by a permutation test of squared errors on the held-out partitions aggregated across all corpora.

6 Experiment 3: Uniform ALiBi Slopes

The version of ALiBi tested in the previous experiments uses different bias slopes for each attention head in a given layer, following Equation (3). It might be asked whether this mixture of slopes is necessary for modeling human reading times, or if a single decay rate—as the dVM bias uses—can perform comparably well. Experiment 3 tests this question by evaluating surprisal estimates from a set of LMs with a simplified version of ALiBi in which all attention heads use the same slope.

6.1 Procedures

Surprisal estimates were collected from a total of 14 LMs using ALiBi with uniform slopes, either during inference only as in Experiment 1 or during both inference and training. The uniform slopes tested were $1/256$, $1/64$, $1/16$, $1/4$, 1, 4, and 16. For consistency with the models in Experiments 1

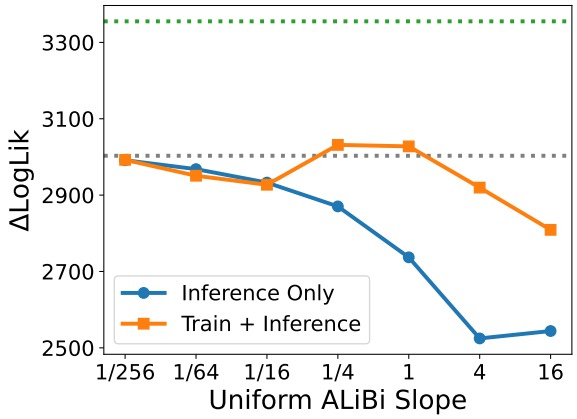


Figure 2: Aggregated likelihood results from Experiment 3. A variant of ALiBi was tested in which all attention heads have the same slope (marked on the x-axis). One set of models included this bias at inference time only (blue line), and the other set included the bias during both training and inference (orange line). The gray dashed line shows the aggregated ΔLogLik from an LM with no recency bias, and the green dashed line shows the same measure from an LM including ALiBi with mixed slopes during training and inference. Per-corpus results are in Appendix C.

and 2, rotary embeddings were included in models with inference-only recency bias, but were removed from models that also included the recency bias during training. Other procedures followed those of Experiments 1 and 2.

6.2 Results

Results from this experiment are reported in Figure 2. Two of the models with uniform ALiBi slopes ($m = 1/4$ and $m = 1$) during both training and inference performed slightly better than the LM with no recency bias (the gray dashed line in Figure 2). However, none of the uniform-slope models matched the performance of the Experiment 2 model that included mixed-slope ALiBi during both training and inference (the green dashed line). These results indicate that including a mixture of slopes is necessary for obtaining optimal reading-time estimates from ALiBi.

7 Experiment 4: Analysis of ALiBi Attention Heads

The previous experiments show that including a mixture of head-specific slopes in ALiBi provides better reading-time estimates than using a single slope across all attention heads. Using mixture of slopes introduces a variable degree of recency bias across heads; we hypothesize that this is helpful for accessing relevant elements in the linguistic context that tend to appear at different distances from the current word.

To test this hypothesis, we explored the sensitivity of the attention heads in the LM from Experiment 2 that includes mixed-slope ALiBi during both training and inference (henceforth `ALiBi-mix-TI`) to three types of semantic dependencies: first arguments (e.g. the relationship between a verb and its subject), second arguments (e.g. the relationship between a verb and its direct object), and coreference (e.g. the relationship between a pronoun and its antecedent). We predicted that different attention heads would be sensitive to first and second argument dependencies (which tend to involve nearby words) and coreference dependencies (which often span longer distances).[11]

7.1 Procedures

The corpus used for this experiment was Natural Stories, which was selected because it had existing annotations with a generalized categorial grammar (Nguyen et al., 2012; Shain et al., 2018) from which semantic dependencies could be extracted. All instances of the three attachment operations—first argument, second argument, and coreference—in which the head of the dependency occurs later than the dependent were identified from the annotations. Instances in which dependents occurred after heads were filtered out, since they are inaccessible in the masked attention used in autoregressive language models. Across the 10 stories in the corpus, there were a total of 2,804 first-argument dependencies, 315 second-argument dependencies, and 1,428 coreference dependencies. Each story entirely fit within the context window of `ALiBi-mix-TI` and therefore no dependencies crossed context windows.

Subsequently, for each dependency type, the mean attention score was calculated for each attention head in `ALiBi-mix-TI`. This score came from averaging over the attention score from the query vector $\mathbf{h}$ corresponding to the head of a dependency, to the key vector $\mathbf{d}$ corresponding to the dependent:

$$\frac{1}{|D|} \sum_{(\mathbf{h},\mathbf{d}) \in D} \text{AttnScore}(\mathbf{h}, \mathbf{d}), \qquad (5)$$

[11] A similar analysis of a model with uniform slope or no recency bias would be more arbitrary, since attention heads in such a model are not distinguished.

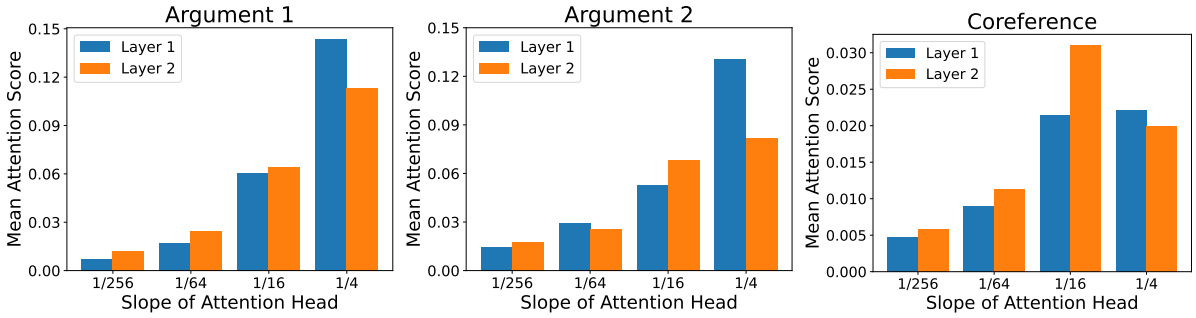


Figure 3: Results from Experiment 4. Mean attention scores for three types of dependencies are presented for each attention head in a model with mixed ALiBi slopes. The evaluated model (`ALiBi-mix-TI`) includes two Transformer layers with four attention heads per layer.

where D is the full set of dependencies of the relevant type and $\mathrm{AttnScore}(\mathbf{h}, \mathbf{d})$ is the attention score between the head and dependent according to the ALiBi formulation in Equation (4). In cases in which the head and/or dependent word spanned multiple tokens, an average was taken between tokens within words before the grand average was taken.

7.2 Results

Figure 3 presents the mean attention scores for the three dependency relations across the eight attention heads in `ALiBi-mix-TI`. For both first- and second-argument dependencies, attention heads with higher slopes (i.e. stronger recency bias) show higher attention scores than heads with lower slopes. For coreference dependencies, however, the first Transformer layer shows similar attention scores from the heads with slopes $1/16$ and $1/4$, and the second layer has the highest mean attention score in the head with slope $1/16$. This suggests that the model makes relatively greater use of an attention head with less decay for longer-distance coreference dependencies.[12] The contrasting attention head behavior across argument and coreference dependencies supports the hypothesis that variable recency bias across heads is helpful for accessing different elements in the context.

8 Discussion

Previous psycholinguistic work has reported conflicting findings about the relationship between the quality of an LM (measured in perplexity) and the psychometric predictive power of its surprisal estimates (Goodkind and Bicknell, 2018; Wilcox et al., 2023a; Oh and Schuler, 2023b). Additionally, while Press et al. (2022) show that ALiBi improves the perplexity of language models, de Varda and Marelli (2024) report that their recency bias degrades language modeling performance. Therefore, a natural question in the context of this work is how the experimental manipulations influence the perplexity of the LM.

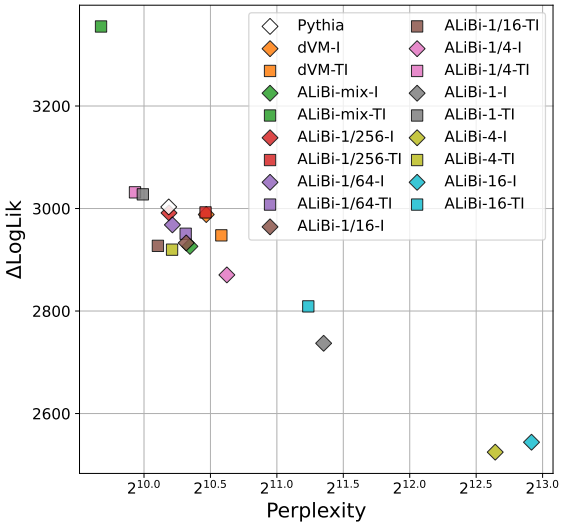


Figure 4: ΔLogLik from LMs as a function of perplexity, both aggregated over all reading-time corpora. In the legend, names ending in *-TI* refer to LMs that include recency bias during training and inference, and names ending in *-I* refer to models with recency bias during inference only.

Figure 4 shows that `ALiBi-mix-TI` achieves lower perplexity than the LM without recency bias, and that the other models tested in this work exhibit a negative relationship between perplexity and ΔLogLik. This suggests that the LMs examined in

[12] The lower overall attention scores for coreference also reflect the longer average dependency distance; a larger number of intervening tokens compete for attention weight.

this work all lie in a regime where more accurate next-word predictions improve the fit of their surprisal estimates to human reading times (Oh and Schuler, 2023a). Moreover, given that the training setup such as the model capacity and training data is identical across conditions, this demonstrates the strong influence of recency biases on the probabilities learned by LMs during training, and opens new possibilities for modeling memory-based effects in sentence processing.

9 Conclusion

This work considers the effect of incorporating recency bias into a Transformer's attention mechanism, as a simple implementation of memory effects suitable for broad-coverage modeling. Improvements in reading-time prediction are observed from ALiBi, a biasing method originally developed for input length extrapolation. Results are strongest when ALiBi is included during both training and inference, and when a mixture of slopes (memory decay parameters) is used across attention heads. Analysis of individual attention heads provides evidence that different slopes may be helpful for tracking shorter- or longer-distance dependencies. These results suggest that incorporating varying rates of memory decay, rather than a single decay parameter as is used in models such as ACT-R (Lewis and Vasishth, 2005), may be a promising direction for developing humanlike models of language processing.

Limitations

The Transformer models evaluated in this study use a single architecture, set of hyperparameters, and training dataset, selected from previous work that tested the influence of model capacity and training data quantity on reading-time predictions (Oh and Schuler, 2023a). With ∼27M parameters, these models are smaller than many of the Transformers used in other areas of natural language processing; it is possible that recency biases could have different effects on larger models.

The alignment between surprisal estimates from Transformer language models and real-time comprehension difficulty presented in this work is based on language model variants trained on English text and data from subjects that are native speakers of English. Therefore, the main findings of this work may not generalize to data collected in other languages. Other possible limitations include the assumption of linear effects of surprisal in regression modeling.

Ethics Statement

This work uses reading-time data collected as part of previously published research (Smith and Levy, 2013; Futrell et al., 2021; Frank et al., 2013; Cop et al., 2017; Kennedy et al., 2003; Luke and Christianson, 2018). Readers are referred to the respective publications for more information on the data collection and validation procedures. As this work focuses on studying the connection between language models and mechanisms underlying human sentence processing, its potential negative impacts on society appear to be minimal.

Acknowledgments

We thank the anonymous reviewers for their helpful feedback. This work was supported by the National Science Foundation grant #1816891. All views expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Alex Andonian, Quentin Anthony, Stella Biderman, Sid Black, Preetham Gali, Leo Gao, Eric Hallahan, Josh Levy-Kramer, Connor Leahy, Lucas Nestler, Kip Parker, Michael Pieler, Shivanshu Purohit, Tri Songz, Phil Wang, and Samuel Weinbach. 2021. GPT-NeoX: Large scale autoregressive language modeling in PyTorch. *Zenodo*.

Alan Baddeley. 2003. Working memory and language: An overview. *Journal of Communication Disorders*, 36(3):189–208.

Alan D. Baddeley and Graham Hitch. 1974. *Working memory*. University of Stirling, Stirling, Scotland.

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68:255–278.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the*

40th International Conference on Machine Learning, volume 202, pages 2397–2430.

Agatha Christie. 1920. *The mysterious affair at Styles*. John Lane. Retrieved from Project Gutenberg.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49(2):602–615.

Andrea Gregor de Varda and Marco Marelli. 2024. Locally biased transformers better align with human reading times. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 30–36.

Jeffrey L. Elman. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7:195–225.

Stefan L. Frank, Irene Fernandez Monsalve, Robin L. Thompson, and Gabriella Vigliocco. 2013. Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, 45(4):1182–1190.

Richard Futrell, Edward Gibson, and Roger P. Levy. 2020. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3).

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi, and Evelina Fedorenko. 2021. The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55:63–77.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint*, arXiv:2101.00027.

Aaron Gokaslan and Vanya Cohen. 2019. OpenWebText Corpus. `http://Skylion007.github.io/OpenWebTextCorpus`.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics*, pages 10–18.

Michael Hahn, Richard Futrell, Edward Gibson, and Roger P. Levy. 2022. A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, 119(43):e2122602119.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696.

Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The Dundee Corpus. In *Proceedings of the 12th European Conference on Eye Movement*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Conference Track Proceedings of the 3rd International Conference on Learning Representations*.

Tatsuki Kuribayashi, Yohei Oseki, Ana Brassard, and Kentaro Inui. 2022. Context limitations make neural language models more human-like. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10421–10436.

Henry Kučera and W. Nelson Francis. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence, RI.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Richard L. Lewis and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419.

Richard L. Lewis, Shravan Vasishth, and Julie A. Van Dyke. 2006. Computational principles of working memory in sentence comprehension. *Trends in Cognitive Science*, 10(10):447–454.

Steven G. Luke and Kiel Christianson. 2018. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2):826–833.

Danny Merkx and Stefan L. Frank. 2021. Human sentence processing: Recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22.

Luan Nguyen, Marten van Schijndel, and William Schuler. 2012. Accurate unbounded dependency recovery using generalized categorial grammars. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 2125–2140.

Byung-Doh Oh, Christian Clark, and William Schuler. 2022. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5:777963.

Byung-Doh Oh and William Schuler. 2022. Entropy- and distance-based predictors from GPT-2 attention patterns predict reading times over and above GPT-2 surprisal. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9324–9334.

Byung-Doh Oh and William Schuler. 2023a. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1915–1921.

Byung-Doh Oh and William Schuler. 2023b. Why does surprisal from larger Transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Byung-Doh Oh and William Schuler. 2024. Leading whitespaces of language models' subword vocabulary pose a confound for calculating word probabilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3464–3472.

Tiago Pimentel and Clara Meister. 2024. How to compute the probability of a word. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18358–18375.

Ofir Press, Noah Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Technical Report*.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. ZeRO: Memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 20.

Keith Rayner, Marcia Carlson, and Lyn Frazier. 1983. The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *Journal of Verbal Learning and Verbal Behavior*, 22(3):358–374.

Soo Hyun Ryu and Richard L. Lewis. 2021. Accounting for agreement phenomena in sentence comprehension with Transformer language models: Effects of similarity-based interference on surprisal and attention. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 61–71.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

Cory Shain, Marten van Schijndel, and William Schuler. 2018. Deep syntactic annotations for broad-coverage psycholinguistic modeling. In *Workshop on Linguistic and Neuro-Cognitive Resources*.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–319.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

William Timkey and Tal Linzen. 2023. A language model with limited memory capacity captures interference in human sentence processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8705–8720.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 6000–6010.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P. Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, pages 1707–1713.

Ethan Gotlieb Wilcox, Clara Meister, Ryan Cotterell, and Tiago Pimentel. 2023a. Language model quality correlates with psychometric predictive power in multiple languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7503–7511.

Ethan Gotlieb Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023b. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Weijie Xu, Jason Chon, Tianran Liu, and Richard Futrell. 2023. The linearity of the effect of surprisal on reading times across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15711–15721.

A Training Procedures of Language Models

All LMs used in the experiments were trained following the procedures of the Pythia LM variants using the GPT-NeoX library (Andonian et al., 2021). Training each model took approximately 3.5 hours on a single 16GB Nvidia V100 GPU. Training batches of 1,024 examples with a sequence length of 2,048 from the Pile (Gao et al., 2020) were provided to each model in the exact same order as

the Pythia models. The Zero Redundancy Optimizer (Rajbhandari et al., 2020) implementation of Adam (Kingma and Ba, 2015) with a learning rate of 0.001 was used to train the model parameters. This learning rate was linearly warmed up over the first 1% of training steps (i.e. 10 steps) and was annealed to a minimum of 0.0001 following a cosine schedule over the remainder of the 1,000 training steps.

B Definition of Eye-Tracking Measures

The following by-word eye-tracking measures were analyzed in this study:

- Scan path (SP) duration: Time taken after entering a word region from the left/right and before entering a different word region to the left/right.
- First-pass (FP) duration: Time taken after entering a word region from the left and before entering a different word region to the left/right.
- Go-past (GP) duration: Time taken after entering a word region from the left and before entering a word region to the right (including all regressive fixations).

C Per-Corpus LMER Results

ΔLogLik values for each tested surprisal variant can be found in Table 4.

LM	Brown	NS	UCL_{SPR}	UCL_{FP}	UCL_{GP}	$Geco_{FP}$	$Geco_{GP}$	$Dundee_{SP}$	$Dundee_{FP}$	$Dundee_{GP}$	$Provo_{SP}$	$Provo_{FP}$	$Provo_{GP}$	Total
Pythia	573.1	582.0	196.4	66.3	53.9	303.1	110.6	126.2	472.0	203.5	5.2	228.8	81.9	3003.1
dVM-I	576.4	554.0	199.9	66.9	54.5	299.0	102.6	122.8	480.4	205.9	4.4	238.2	83.2	2988.3
dVM-TI	556.4	574.0	193.1	57.2	50.2	303.4	108.3	112.3	479.3	198.2	4.6	235.6	75.0	2947.7
ALiBi-mix-I	561.1	551.0	194.6	66.5	53.4	296.8	109.6	122.8	459.5	200.4	4.4	224.0	82.0	2926.1
ALiBI-mix-TI	593.6	643.0	197.1	76.7	63.7	427.5	166.7	133.4	473.3	204.0	5.7	274.9	95.5	3355.1
ALiBi-1/256-I	571.9	577.0	196.3	66.3	53.9	300.9	110.7	125.7	469.7	202.7	5.2	228.9	82.1	2991.4
ALiBI-1/256-TI	558.3	592.0	202.3	56.5	53.6	301.3	106.1	116.1	479.6	201.7	5.0	239.1	80.8	2992.4
ALiBi-1/64-I	569.0	565.0	196.3	66.4	54.0	298.7	110.8	124.2	465.2	201.5	5.2	229.1	82.7	2968.1
ALiBI-1/64-TI	555.3	577.0	202.2	66.5	51.9	296.0	107.7	117.2	459.7	189.8	4.8	239.2	83.5	2950.7
ALiBi-1/16-I	562.1	551.0	195.5	66.5	54.2	298.3	111.3	119.2	456.4	200.7	4.8	228.3	84.5	2932.8
ALiBI-1/16-TI	545.1	545.0	210.9	66.2	60.7	286.8	103.7	118.9	457.0	192.5	4.9	247.2	88.2	2927.2
ALiBi-1/4-I	548.3	537.0	191.4	66.8	54.8	308.3	114.2	111.5	436.7	194.2	2.4	223.3	81.6	2870.4
ALiBI-1/4-TI	550.4	550.0	216.4	68.2	59.3	321.3	125.6	121.8	455.8	201.8	5.9	258.6	96.5	3031.6
ALiBi-1-I	530.3	494.0	181.9	64.2	53.2	322.4	118.3	105.9	408.7	184.8	−0.5	200.9	72.9	2737.0
ALiBI-1-TI	538.1	543.0	219.2	69.3	61.0	333.1	132.3	123.3	464.4	205.2	−13.3	256.9	95.2	3027.7
ALiBi-4-I	523.7	446.0	155.7	50.1	38.0	321.2	115.2	84.8	382.4	171.1	−2.0	177.3	61.2	2524.7
ALiBI-4-TI	543.1	541.0	222.0	66.6	58.0	292.3	108.5	119.6	445.3	195.2	4.5	238.1	85.5	2919.7
ALiBi-16-I	529.1	501.0	146.7	45.9	38.0	321.2	114.2	74.9	384.1	169.1	−1.5	165.8	55.5	2544.0
ALiBI-16-TI	547.3	512.0	206.5	48.9	39.1	320.2	112.7	95.8	452.1	193.5	2.6	217.1	61.3	2809.0

Table 4: ΔLogLik of each surprisal variant tested in Experiments 1–3 on each corpus. *NS* stands for Natural Stories. *Pythia* is the LM with no recency bias. LMs starting with *dVM* use the bias technique from de Varda and Marelli (2024), and LMs starting with *ALiBi* use the eponymous bias technique from Press et al. (2022). Models ending in *-I* include recency bias at inference time only, while those ending in *-TI* include recency bias during both training and inference. Models containing *mix* use the mixture of attention head slopes recommended by Press et al. (2022), and models containing a number (1/256, 1/64, 1/16, 1/4, 1, 4, or 16) use that number as a constant slope across all attention heads.