Aggregated Transformer Attention Measures Predict Reading Times Beyond Surprisal

Lukas Mielczarek, Laura Kallmeyer
Heinrich-Heine-Universität Düsseldorf
Universitätsstraße 1, 40225 Düsseldorf
{lukas.mielczarek, laura.kallmeyer}@hhu.de

Abstract

Transformer language models trained solely for next-token prediction provide strong expectation-based predictors of human reading behaviour via surprisal. A growing literature suggests that additional signals derived from the internal structure of these models – especially from self-attention – may capture memory-related processing difficulty. This paper uses a range of attention-derived structural costs, specifically *attention entropy*, *attention-weighted distance*, *attention-based activation* and *attention-based memory-focus difference*, that are intended to model memory-related processing costs. We define two aggregation schemes for attention patterns across all heads and layers: global metrics computed from an attention distribution aggregated across heads, and head-averaged metrics computed per head and then averaged. For these metrics, we test whether they predict human reading times beyond surprisal: We outline a methodology to evaluate these metrics using linear mixed-effects models on the Provo eye-tracking dataset, controlling for standard predictors such as word length and frequency and including surprisal as a baseline. As a result, we find that unnormalised head-averaged metrics consistently improve prediction beyond surprisal, while this is not the case for globally aggregated and context-scaled variants.

Keywords: transformers, self-attention, reading times, surprisal, attention entropy

1. Introduction

Human reading times (RTs) are strongly influenced by expectation: less predictable words in context incur greater processing difficulty. This relation is formalised as surprisal (Hale, 2001; Levy, 2008), and transformer language models (LMs) provide surprisal estimates (negative log probabilities) that robustly predict human reading behaviour. However, expectation-based accounts do not fully explain processing difficulty (van Schijndel and Linzen, 2021; Arehalli et al., 2023). Memory-based theories attribute additional costs to limitations in encoding and retrieving representations, such as similarity-based interference during retrieval (Lewis, 1993, 1996) and costs associated with long-distance dependencies (Gibson, 2000).

Transformer self-attention offers a potential bridge between these perspectives. Because attention distributes weight over prior tokens, structural properties of attention distributions – such as entropy (Ryu and Lewis, 2021, 2022) or difference in attention distributions between consecutive timesteps (Oh and Schuler, 2022) – have been proposed as proxies for memory-related processing costs. Yet two key issues remain unclear. First, much of the work has focused on attention entropy, omitting formalisations of important theories in the field of sentence processing, questioning the uniqueness of the contribution of entropy. Second, attention in transformers is multi-headed: studies have opted to analyse individual syntactically specialised attention heads or small selections of heads. It is unknown whether all attentional spread in transformer models is associated with processing difficulties (Ryu and Lewis, 2025). We address these issues by asking the following question: **How well do attention-derived structural costs extracted from transformer LMs trained with a pure language-modelling objective predict human reading times beyond surprisal?**

Concretely, we evaluate four memory-motivated attention metrics – entropy, distance, activation, and attention difference – under a unified framework. For each metric, we compare head-averaged and globally aggregated variants, as well as context-length-scaled versions. Using a transformer language model trained solely for next-token prediction, we test whether these metrics predict go-past RTs in the Provo eye-tracking corpus beyond a baseline including frequency, length, surprisal, and spillover effects.

Our main findings are that unnormalised head-averaged metrics consistently improve prediction beyond surprisal, whereas globally aggregated and context-scaled variants show weaker or no effects. Much of the predictive signal overlaps with sentence position, indicating that attention-derived costs are closely tied to expanding contextual memory demands. These findings suggest that transformer attention encodes structural signals aligned with memory-related difficulty, but that their behavioural relevance depends critically on aggregation choices.

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 188–202
16 May 2026.

2. Related Work

Recent work has used LMs to connect theories of human sentence processing with measurable RT predictors. Two lines of theory motivate this: expectation-based (Hale, 2001; Levy, 2008) and memory-based accounts (Gibson, 2000).

Expectation-based theories propose that processing difficulty is driven by the unexpectedness of the next word, which can be opertionalised by *surprisal*. Transformer LMs (Vaswani et al., 2017) yield state-of-the-art estimates of surprisal and reproduce robust correlations between surprisal and human reading behaviour (Wilcox et al., 2023).

Memory-based theories instead attribute difficulty to limitations in encoding and retrieving representations during incremental comprehension. These theories vary in their assumed mechanisms and therefore make different predictions about reading patterns. Two influential approaches are particularly relevant here. First, *cue-based retrieval* theories (Van Dyke and Lewis, 2003) assume that incoming material provides cues that probe memory for previously processed items (words, constituents, referents). When multiple candidates match the cues, retrieval suffers from *similarity-based interference*, yielding increased processing time (Gordon et al., 2002, 2006). Second, *Dependency Locality Theory* (DLT) proposes that integrating new material into an evolving dependency structure incurs higher *integration costs* when the relevant dependencies are long (Gibson, 2000).

Beyond interference and locality, memory-based processing theories also motivate *activation/recency* effects, where representations decay unless reactivated, making later retrieval more costly (Lewis and Vasishth, 2005) , and costs associated with *state change* or reallocation of processing resources, linked in modelling work to shifts in attention focus across time (Oh and Schuler, 2022).

A growing set of studies relates transformer self-attention to memory-based mechanisms. Ryu and Lewis (2021) connect attention to cue-based retrieval and use an attention-based entropy measure to model interference effects. Entropy and distance-based metrics on transformer attention predict RTs across both eye-tracking and self-paced reading datasets (Oh and Schuler, 2022; Ryu and Lewis, 2022). Ryu and Lewis (2025) show that attention entropy aggregated over a subset of heads remains predictive, interpreting heads as parallel retrieval operations.

The present work differs from prior studies by (i) evaluating a broader set of attention-derived, psycholinguistically motivated structural cost measures under unified definitions and (ii) systematically comparing head-wise vs. globally aggregated computation (and corresponding normalisations).

3. Psycholinguistically Inspired Processing Metrics

In the following, starting from attention distributions, we devise measures that are intended to model memory-related psycholinguistic processing factors, which in turn are assumed to explain RTs.

3.1. Attention Distribution

Let the matrices $\mathbf{Q}_h$ contain the queries and $\mathbf{K}_h$ the keys of an input sequence at the h-th attention head. Furthermore, let d_k denote the query/key vector dimensionality. At the i-th position in the sequence, we get the following distribution:

$$\mathbf{A}_h^{[i]} = \text{softmax}\left(\frac{\mathbf{Q}_h^{[i]}\mathbf{K}_h^{\mathsf{T}}}{\sqrt{d_k}} + \mathbf{M}\right) \quad (1)$$

where $\mathbf{M}$ is a causal mask with $\mathbf{M}^{[i,j]} = -\infty$ for $j > i$ and 0 otherwise (Vaswani et al., 2017).

Raw attention $\mathbf{A}_h^{[i,j]}$ indicates how much token i attends to token j in attention head h but the quantities combined in multi-head attention are actually weighted sums of $\mathbf{A}_h^{[i]}$ and a matrix of values generated from the input representations $\mathbf{V}_h$. Furthermore, the output transformation $\mathbf{W}^{\mathbf{O}}$ applied to the concatenated head outputs can be decomposed into per-head subcomponents. This means that the contribution to the output of head h which attention from i to j provides crucially also depends on the magnitude of $\mathbf{V}_h$ and $\mathbf{W}^{\mathbf{O}}{}_h$ (Kobayashi et al., 2020). Therefore, following Oh and Schuler (2022), we take the norm of the weighted value vectors from the attention mechanism. Let $\mathbf{b}^{\mathbf{O}}$ be the optional output transform bias term and m the number of attention heads:

$$\mathbf{N}_h^{[i,j]} := \frac{\mathbf{A}_h^{[i,j]} \cdot \left\|\tilde{\mathbf{V}}_h^{[j]}\right\|_2}{\sum_{k=1}^{i} \mathbf{A}_h^{[i,k]} \cdot \left\|\tilde{\mathbf{V}}_h^{[k]}\right\|_2}, \quad (2)$$

$$\text{with} \quad \tilde{\mathbf{V}}_h^{[j]} = \mathbf{V}_h^{[j]}\mathbf{W}^{\mathbf{O}}{}_h + m^{-1}\mathbf{b}^{\mathbf{O}}.$$

Oh and Schuler (2022) also test a variant normalised with respect to residual connections and layer normalisations. However, since they report no improvement for entropy over the simpler version above, we do not adopt it.

A transformer model can consist of a large number of individual attention heads spread over layers and multi-head attention modules. Inspired by Attanasio et al. (2022), we will quantify the contribution of token j to the transformer output at position i by defining a global attention distribution $\bar{\mathbf{N}}^{[i,j]}$ that is simply the average attention token i assigns to token j taken over all heads in the model.

$$\bar{\mathbf{N}}^{[i,j]} := \frac{1}{m}\sum_{h=1}^{m} \mathbf{N}_h^{[i,j]} \quad (3)$$

3.2. Attention Metrics

The attention-based measures we introduce can be computed in conceptually distinct ways (local/global and normalised/not normalised): First, they can be calculated separately for each attention head using the per-head attention distribution (Equation 2) and then averaged across heads (called *local*). We hope that this approach preserves potential functional specialisation across heads, which has been documented in transformer models (e.g. Voita et al., 2019; Clark et al., 2019). Alternatively, the measures can be computed from the globally aggregated attention distribution (see Equation 3). We refer to this as the *global* variant. Secondly, our attention-derived quantities can be expected to increase with the size of the left context. Thus, we also explore normalised measures that control for trivial sentence-length effects.

Attention Entropy and Similarity-Based Interference Cue-based retrieval theories predict similarity-based interference effects: when several memory representations partially match the retrieval cues, retrieval becomes noisy and slower (Van Dyke and Lewis, 2003; Lewis et al., 2006). If attention implements a soft retrieval mechanism, a sharply peaked attention distribution corresponds to selective retrieval of a single dominant candidate, whereas a diffuse distribution reflects competition among multiple candidates.

We operationalise similarity-based interference as the entropy of the attention distribution based on Ryu and Lewis (2021). For the global attention matrix $\bar{\mathbf{N}}$ for input sequence s, attention entropy at position i is defined as

$$\mathbf{H}^{[i]}_{\text{glob}} := -\sum_{j=1}^{i} \bar{\mathbf{N}}^{[i,j]} \log_2 \bar{\mathbf{N}}^{[i,j]}. \tag{4}$$

For the local variant, entropy is computed separately for each head h and then averaged:

$$\begin{aligned} \mathbf{H}^{[i]}_{h} &:= -\textstyle\sum_{j=1}^{i} \mathbf{N}^{[i,j]}_{h} \log_2 \mathbf{N}^{[i,j]}_{h} \\ \mathbf{H}^{[i]}_{\text{loc}} &:= \tfrac{1}{m} \textstyle\sum_{h=1}^{m} \mathbf{H}_h(i,s). \end{aligned} \tag{5}$$

Higher entropy reflects greater competition among partially matching items and is therefore predicted to correlate positively with reading times.

Distance and Dependency Locality Dependency Locality Theory (Gibson, 2000) (DLT) proposes that integration cost increases with the distance between dependent elements measured by the number of intervening referents (usually nouns and verbs). Establishing longer dependencies requires maintaining representations across more intervening material, leading to greater processing cost.

Since attention assigns continuous weights to prior positions, it does not provide a discrete set of retrieved items. Furthermore, transformer LMs do not overtly distinguish between referents and non-referents. Rather than imposing a threshold to identify "retrieved" tokens and classifying dependents, we define a continuous proxy for integration distance via the expected backward distance under the attention distribution:

$$\mathbf{D}^{[i]} := \sum_{j=1}^{i} \mathbf{N}^{[i,j]}(i-j). \tag{6}$$

Larger values indicate longer-distance retrieval and are predicted to correlate with increased processing cost under DLT.

Because expectation is linear, averaging head-wise expected distances yields the same value as computing the expectation from the globally averaged distribution. We do not distinguish two versions for this measure.

Activation and Memory Decay In models of sentence processing based on the cognitive architecture ACT-R (Anderson et al., 2004; Vasishth and Lewis, 2004), activation decreases as a function of time since last retrieval, raising subsequent retrieval cost (Lewis and Vasishth, 2005). Representations decay over time unless they are reactivated.

To approximate this, we define a recency trace over attention matrices: For an matrix $\mathbf{X}$, the degradation function degr is defined recursively as:

$$\begin{aligned} \text{degr}(\mathbf{X})^{[1,j]} &:= 0 \\ \text{degr}(\mathbf{X})^{[i,j]} &:= (1-\mathbf{X}^{[i-1,j]})\ \text{degr}(\mathbf{X})^{[i-1,j]} \\ &\quad +(1-\mathbf{X}^{[i-1,j]}). \end{aligned} \tag{7}$$

Intuitively, if position j is not attended to at $i-1$, its recency trace increases at i; otherwise, the trace is partially reset. The activation-based cost at i is then defined as the expected recency under the current attention distribution (Equation 8). We compute $\mathbf{W}_{\text{glob}}$ using $\bar{\mathbf{N}}$ and $\mathbf{W}_{\text{loc}}$ by computing $\mathbf{W}_h$ per head and averaging across heads.

$$\mathbf{W}^{[i]} := \textstyle\sum_{j=1}^{i} \mathbf{N}^{[i,j]} \text{degr}(\mathbf{N})^{[i,j]} \tag{8}$$

This predictor is large when attention is allocated to items that have not been recently attended, corresponding to retrieval of low-activation representations. Importantly, it captures temporal recency rather than purely structural distance: a representation may be linearly close yet inactive, or distant yet frequently refreshed.

Attention Difference and Reallocation If contextual representations are stored sequentially in memory, reallocating attention across distant positions may incur additional processing cost. In attention-based models, each timestep redistributes retrieval weight over prior tokens, effectively shifting a soft pointer along a linear memory. We quantify this

shift as the expected displacement between consecutive attention distributions.

For positions $i > 1$, we define the reconfiguration cost as the expected absolute displacement between the attention distributions at $i-1$ and i:

$$\mathbf{J}^{[i]} := \sum_{k=1}^{i-1} \sum_{j=1}^{i} \mathbf{N}^{[i-1,k]} \mathbf{N}^{[i,j]} |j-k|. \quad (9)$$

For completeness, we set $\mathbf{J}^{[1]} := 0$.

Intuitively, this quantity is small when attention remains concentrated on similar memory positions across timesteps, and increases as attention shifts toward more distant tokens. Unlike Earth Mover's Distance (EMD) as used by Oh and Schuler, 2022, our formulation captures overall dispersion in retrieval focus rather than the minimal work required to transform one distribution into the other. In particular, it assigns non-zero cost to dispersed but identical distributions, penalising diffuse attention.

As with our other attention-based metrics, we compute both local and global variants.

3.3. Normalisation

Since all proposed metrics increase with larger left contexts assuming uniform attention distributions, we define length-normalised variants that divide each measure at position i by its maximal attainable value. This maps each metric to the interval $[0, 1]$.

For attention entropy, we adopt the method of Oh and Schuler (2022): the maximum occurs under a uniform distribution over $\{1, \ldots, i\}$ and is therefore $\log i$. For expected backward distance, the maximum occurs when all attention is on the leftmost position, yielding a cost of $i-1$. For activation-based cost, the degradation trace of a position that has never been attended grows linearly with its distance from the current position. The maximal activation cost is thus likewise obtained by attending fully to the oldest, maximally degraded position, yielding $i-1$. Finally, for the difference metric, the maximum occurs when attention shifts between opposite extremes of memory in consecutive steps. In this case, the absolute displacement equals $i-1$. Thus, in all three cases, the normalised metric is gained by dividing the measure by $i-1$ (analogous to Oh and Schuler, 2022's normalisation of EMD). We call the normalised measures $\tilde{\mathbf{H}}, \tilde{\mathbf{D}}, \tilde{\mathbf{W}}$ and $\tilde{\mathbf{J}}$.

4. Methodology

4.1. Language Model Training

We trained a transformer LM comparable in architecture and parameter count to `GPT-2 medium`[1] (Radford et al., 2019) and based on the implementation of Mielczarek et al. (2025).

[1] *Generative Pretrained Transformer*

	Value
parameter count	482 590 464
d_{model}	768
vocab size	222 319
f_{FF}	9
dropout	
– embedding	0.14
– feed-forward	0.32
– residual	0.17
learning rate	1.92×10^{-4}

Table 1: Selection of hyperparameters for training the final model.

Split	Train	Validation
loss	2.93	3.45
perplexity	18.70 (70.39)	31.45 (72.96)
final epochs	8.8	

Table 2: Benchmarks on the final model. In parentheses, the corresponding perplexities of `GPT-2 medium` on the sentence-level are provided).

Prior research on attention patterns for RT prediction has focused on the `GPT-2 small` transformer variant (Ryu and Lewis, 2022; Oh and Schuler, 2022; Ryu and Lewis, 2025) and it remains unclear whether the reported effects generalise to transformers with more attention heads. This motives our use of a larger model. The rationale for training a new model was to take into account only intra-sentential contexts, whereas `GPT-2` was trained using a wide context window that crosses sentence boundaries. Since the theories motivating our attention measures predominantly make predictions with respect to sentence processing and are evaluated on the sentence level in prior literature, we aimed to control for extra-sentential confounders.

The same motivation underlies the use of word-level tokenisation instead of Byte Pair Encoding (Gage, 1994) where psycholinguistic theory often employs the simplifying assumption that whole words constitute cognitive units. Punctuation is removed and all words are uncapitalised. Embedding dimensionality is reduced from 1024 to 768 due to the restricted availability of computational resources. Note that our model has a larger overall dimensionality than `GPT-2 medium` (483M vs 355M) due to our large embedding matrix. Furthermore, a hyperparameter search was performed prior to full training. The key hyperparameters that differ from `GPT-2 medium` are listed in Table 1.

All training was conducted on the pre-processed Wikitext-103-v1 dataset (Merity et al., 2016) consisting of over 100 million tokens from Wikipedia. Since we investigate sentence processing, the training contexts are reduced to the sentence-level and a beginning-of-sequence token is included. We en-

force a maximum sentence length of 60. The final model was trained until convergence using early stopping based on validation performance. The selected checkpoint minimised validation perplexity. Table 2 contains information on the final model.

4.2. Psycholinguistic Evaluation

For answering the research question, we will extract the memory-based measures defined in Section 3 from attention scores generated by our language model on a psycholinguistic dataset. Then, we will test whether these measures have an effect on human reading times for that dataset, above and beyond baseline factors that are well-known to influence sentence processing.

Dataset We evaluate predictors on the Provo eye-tracking corpus (Luke and Christianson, 2018). It is based on eye-measurements of 84 native English speakers who were instructed to read 55 paragraphs from diverse sources such as news articles, science magazines and fiction. The paragraphs were presented to the participants in a random order and contained 2.5 sentences each (Luke and Christianson, 2018). From the available metrics in Provo, we use *go-past time* (GPT), which is the sum of all fixation durations including regressive fixations starting with the first fixation on that word up to the point of leaving to the right (Radach et al., 2008). We choose this measure since it is the standard for investigations of high-level meaning integration processes (Radach and Kennedy, 2013) which we assume to be most indicative of any type of memory effects.

Words that were skipped by the participant are not included in the analysis since we assume that they do not cause integrative processes. Where items in Provo correspond to several tokens in our language model, they are tokenised correspondingly and the extracted measures are averaged afterwards.

Linear Mixed-Effects Models We investigate the potential effect of attention-based metrics on GPT with *linear mixed-effects models* (LMEMs), using the `lme4` implementation in `R` (Bates et al., 2015). To assess the significance of a predictor, we compared two nested linear mixed-effects models using a *likelihood ratio test* (LRT). The full model included the predictor as a fixed effect, whereas the reduced model omitted it. Random effects for the predictor to test are included in both models. This allows us to test whether its effect significantly exceeds random group-level variability. Model comparison was conducted via a χ^2 test on the likelihood ratio. The p-value quantifies the probability of obtaining a likelihood ratio statistic at least as large as observed, assuming the reduced model (i.e. no effect of the predictor). We adopt standard practice in rejecting the null hypothesis if p is below the significance threshold of $\alpha = 0.05$.

We also provide the delta in *Akaike Information Criterion* (AIC) and *Bayesian information criterion* (BIC) (see Chapter 2 in Burnham and Anderson, 1998) and the estimated effect sizes.

To maintain generalisability of our models and prevent Type I error inflation (false positives), we aim for a maximal random effects structure, following Barr et al. (2013). However, as a maximal structure is not supported by our data and we encounter issues with the fitting process, we employ the simplification method from Bates et al. (2018).

Baseline Effects We establish a baseline model structure motivated by known factors contributing to RTs. Word frequency and word length are known to be strong predictors (Rayner, 1998). We calculate log frequency using the `wordfreq` Python module Speer (2022). We also include surprisal, i.e. the negative log probability of the respective word given its left context restricted to the sentence-level generated by our LM as a base predictor and we check whether to include features tracking temporal reading progression: sentence and word position inside the paragraph and word position inside the sentence. All predictors are standardised (z-transformed) to allow for faster model fitting.

For the grouping variables, besides the per-participant effects, we consider two hierarchies for nested effects: `lemma` $\rightarrow$ `word` (form) and `paragraph` $\rightarrow$ `sentence` $\rightarrow$ `abs.pos`. Here `abs.pos` corresponds to one specific token.

After finding a maximal model for the three baseline predictors and the given grouping variables, we check for possible *spillover*/*lag*-effects. Processing slowdown is often delayed in reading time data (Ehrlich and Rayner, 1983). We need to account for this in order to ensure that our analysis is not compromised by spurious correlations. Therefore, we consider shifted versions of the predictors, e.g. for the word at position i, we check if we need to take into account word length, frequency and surprisal at $i-1$ (or further). This is done according to the standard procedure: Given spillover window s for a variable, we not only use the value assigned to the current word but also those of the s preceding ones. s is chosen by iteratively performing LRTs for larger spilled-over variants as long as significance is given (cf. Xu et al., 2023). We do this separately for each baseline predictor to prevent overspecification.

5. Results

First, we will outline the results of the LMEM selection process. Then, we will discuss the results of the likelihood ratio tests for the attention measures. Finally, a follow-up analysis will investigate the robustness of our results against spurious effects due

	Estimate	SE
intercept	282.01	5.36
frequency	-7.77	1.95
length	23.11	2.12
surprisal	8.07	1.21
length.1	-7.30	1.23
surprisal.1	4.65	1.22
surprisal.2	2.34	0.85

Table 3: Baseline LMEM effect estimates and standard errors.

to word position.

5.1. Baseline Model

In the following, we describe the baseline effects structure resulting from the model selection process. The complete dataset has 212 436 observations. Conservative cleaning, following standard practices as in Eskenazi (2024); Marsden et al. (2018) (removal of skipped-over words, cut-offs 80–3000 ms, outlier removal by 2.5 SD, removal of sentences with words unknown to the LM tokeniser), leads to 130 151 remaining observations.

Temporal features are not included because the significance threshold was not reached. The `lemma` grouping variable was also discarded due to singularity issues when including random coefficients and non-significance of its random intercept.

Visual inspection confirms that the assumptions that the LRT relies on (linearity of the relationship between predictors and dependent variable, normality of the residuals, homogeneity of variances and normality of the random effects) are met to an acceptable degree. The QQ-plot reveals that the residuals are somewhat right-skewed. We suspect that we could mitigate this by applying a log transformation to the dependent. However, this would reduce interpretability of the model and contradict prior knowledge about the linearity of our base effects, e.g. the linear relationship between surprisal and RTs. Therefore, we rely on the relatively large amount of data to allow reliable estimates.

Additionally, the analysis reveals some irregularities for very low log frequencies. Manual inspection indicates that these extreme values are likely caused by character encoding errors. Therefore, we remove items with values <2 (= occurring less than 100 times per 1 billion words) from the analysis, removing another 550 data points.

The spillover identification process suggests the inclusion of length up to a window of 1 and surprisal with a window of 2. This excludes the first two words per sentence (9240 data points).

Table 3 shows the effect estimates of the baseline model. Correlations are accounted for in the subject grouping variable.

Model Designs Including attention measures required additional simplifications of the random effects structure. In all cases, correlations were not supported by the data. Furthermore, for some candidates, specific random coefficients had to be omitted. The estimated models are given in Table 4 with the first row containing the baseline model without subject-level correlations.

Repeated Testing For each dataset, four candidates are tested with modes weighted/unweighted each and three with global/local features, resulting in 14 statistical tests on the same dataset. This increases the probability of us reporting at least one false positive result and falsely reporting that attention patterns predict RTs. Therefore, we correct the p-values via the *Benjamini–Hochberg* (BH) procedure (Benjamini and Hochberg, 1995).

5.2. Attention-Based Predictions

Overview Table 5 summarises fixed-effect estimates (β; per 1 SD increase), likelihood-ratio tests, raw p-values, BH-corrected q-values, and information-criterion differences (ΔAIC, ΔBIC) for attention-based predictors of go-past time (GPT) relative to their respective baseline models.

Across specifications, three patterns emerge: (i) head-averaged (local), unnormalised metrics consistently improve fit beyond surprisal and lexical controls; (ii) globally aggregated metrics show weaker and less reliable improvements; (iii) context-length scaling largely removes (and sometimes reverses) effects.

Head-Averaged (Local) Metrics All four unnormalised head-averaged metrics – entropy, distance, activation, and difference – significantly improved fit relative to the baseline ($q < 0.05$). Fixed effects were positive and similar in magnitude across predictors ($\beta \approx 3$ ms per SD), indicating that higher attention-derived memory cost is associated with longer go-past times.

Improvements in AIC were modest but consistent across predictors, suggesting small but systematic gains in explanatory power beyond surprisal, word length, frequency, and their spillover terms. Effect sizes are less than half the estimated effect of surprisal (see Table 3). Notably, this contrasts with findings on self-paced reading where attention entropy can exceed surprisal in predictive strength (Oh and Schuler, 2022).

Table 7 provides correlation coefficients for the baseline predictors and the unscaled local attention measures. Correlations for the other predictor candidates can be found in Appendix A. Based on these values, multicollinearity in the LMEMs is not a concern. The attention measures (not included together in a joint LMEM) are strongly correlated,

Candidates	Formula
$\mathbf{H}_{\text{loc}}, \mathbf{D}, \mathbf{W}_{\text{loc}}, \mathbf{J}_{\text{loc}}, \mathbf{W}_{\text{glob}}, \mathbf{J}_{\text{glob}}$	`GPT ~ 1 + freq + len + surp + len.1 + surp.1 + surp.2 + cand + (1 + freq + len + len.1 + surp + cand || subject) + (1 + cand || word) + (1 + surp + surp.1 || paragraph) + (1 + cand || sentence) + (1 + freq + len.1 + cand || abs.pos)`
$\mathbf{H}_{\text{glob}}, \tilde{\mathbf{W}}_{\text{loc}}, \tilde{\mathbf{W}}_{\text{glob}}$	`-(cand | abs.pos)`
$\tilde{\mathbf{H}}_{\text{loc}}$	`-(cand | word)`
$\tilde{\mathbf{D}}, \tilde{\mathbf{H}}_{\text{glob}}$	`-(cand | word) - (cand | sentence)`
$\tilde{\mathbf{J}}_{\text{loc}}$	`-(cand | word) - (surp.1 | paragraph) - (cand | abs.pos)`
$\tilde{\mathbf{J}}_{\text{glob}}$	`-(cand | word) - (cand | abs.pos)`

Table 4: Identified LMEMs for testing the attention-based measures. The second column supplies the LME model descriptions used for individually testing the candidate metrics given in the first column. The term `cand` stands for the respective candidate metric. Rows 2–6 contain the components removed from the maximal model fitted for LRT given in row 1. The notation is explicit – intercepts were not removed. Given in `lme4` syntax. `freq` = frequency, `len` = length, `surp` = surprisal. "`.1`" and "`.2`" signify lagged versions of predictors. "`||`" means that correlations between random effects are not accounted for.

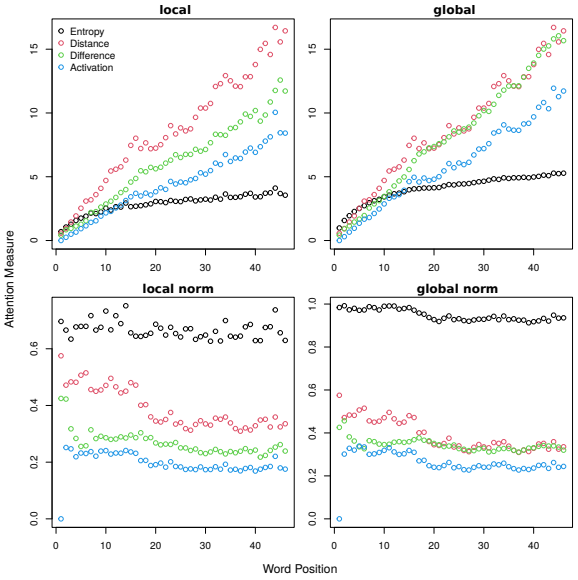

Figure 1: Example attention measures for the sentence *"When a nucleus has all of its proton shells or neutron shells loaded to full capacity, the layers can align so well that each shell can spoon intimately with its attractive neighbors, forming a more compact sphere that fits well within the strong nuclear force's area of influence."*

with the highest correlations of 0.99 occurring between attention distance and activation, and between difference and activation. This suggests that these measures capture closely related aspects of attention structure.

Globally Aggregated Metrics When metrics were computed from the globally averaged attention distribution, results were less consistent. Activation and attention difference remained significant after correction ($q < 0.05$), with positive coefficients comparable to the local variants, whereas global entropy did not reach significance and showed a smaller effect size than local entropy (2.29 vs 2.97).

Inspection of the distributions suggests a mechanistic reason for the weaker global results: after averaging heads, the attention distribution becomes highly diffuse, pushing entropy toward its maximum, thereby reducing variability. Fig. 1 illustrates this for a long sentence: local metrics vary substantially, while global entropy closely tracks its maximum.

These observations indicate that averaging attention *before* metric computation can wash out head-specific structure that is informative for GPT. If attention is interpreted as a retrieval-like mechanism, the contrast between local and global variants suggests that a single pooled "retrieval process" is an oversimplification of how transformers allocate attention. Additionally, our global attention distribution $\bar{\mathbf{N}}$ may be influenced by noisy uninformative attention heads whose contribution to the model is actually low following their output weight $\mathbf{W}^O$. However, our measure takes all attention heads into account equally, without weighing them with respect to their overall magnitude. This may manifest in semi-static noise when computing attention measures on a per-head basis in the local setting (possibly being a contributor to the linear trend observed in Fig. 1), but destroy the global distribution.

At the same time, these findings do not warrant the strong claim that attention *is* cognitive retrieval. A more conservative linking hypothesis is supported: certain attention-derived structural properties correlate with GPT in a direction consistent with memory-based accounts, but the behaviourally relevant signal appears concentrated in specific components as evidenced by the success of Ryu and Lewis (2021, 2022) in using only specific heads, rather than the fully pooled distribution. Architecturally, this suggests that some retrieval-like operations contribute to internal computation without behavioural consequences. It remains a challenge to unify the cognitively plausible view of a single

retrieval mechanism (cf. Timkey and Linzen, 2023) with the multiple parallel and hierarchical attention operations in transformers if we want to explore psycholinguistic hypotheses using these models.

Context-Length Scaling In contrast to the unnormalised predictors, none of the context-length-scaled variants produced significant improvements beyond the baseline after correction. Moreover, several scaled predictors exhibited negative coefficients, indicating shorter predicted GPT with increasing scaled memory cost. The bottom plots in Fig. 1 hint towards a recency bias in attention with the measures declining throughout the sentence (i.e. sublinear growth with word position for distance, activation and difference). The absence of positive scaled effects suggests that GPT aligns more with absolute growth in contextual demands than with proportional attention allocation within an expanding window as suggested by Oh and Schuler (2022).

Magnitude of Improvement Across significant models, ΔAIC values (Table 5) indicate modest but consistent improvements over the baseline. Given the large number of observations, small ΔAIC reductions correspond to systematic likelihood gains; nevertheless, the overall magnitude of improvement remains small relative to the dominant contribution of surprisal. Thus, attention-based metrics explain additional variance in eye-tracking GPT beyond expectation-based and lexical factors, but surprisal remains the dominant predictor, consistent with the findings of Ryu and Lewis (2025).

Robustness to Sentence Position Because the unnormalised attention metrics increase with sentence position, we tested whether their improvements reflect shared variance with word position rather than an independent contribution. Adding position (linear) and log(position) to the maximal baseline model (first row in Table 2) yielded significant positive effects (linear effect: 3.35, p=0.011; log effect: 2.51, p=0.015), confirming a general increase in GPT across sentences not captured by surprisal, length or frequency. Now, controlling for position, the improvements of the attention-based metrics were substantially reduced: only entropy showed a marginal improvement in raw p-values for log(position) resulting in a flipped position effect of -6.85 and an entropy effect of 9.64. However, no effects survived multiple-testing correction (separate correction under the new stricter specification, Table 6).

This suggests that part of the predictive signal in the unnormalised metrics overlaps with generic context growth. Note that this does not imply that attention metrics are spurious. Increased memory costs caused by a larger context might be the cause of the position effect in the first place.

Dependency length Lastly, we correlate $\mathbf{D}$ with dependency lengths extracted from the annotated UD EWT corpus (Silveira et al., 2014). For each word, we computed the sum of distances to all tokens in its left context that are in a dependency relation with it (i.e. its dependents and, where applicable, its head). We find a modest Pearson correlation of 0.22. Manual inspection shows that a large number of short dependency length items were assigned high attention distances (see Appendix B). This could reflect contextualisation needed for next-token prediction in the absence of a recurrent representation.

6. Discussion

In summary, attention-derived structural metrics over all transformer heads provide modest but systematic improvements in predicting reading times beyond surprisal. However, their explanatory contribution depends critically on aggregation choices and overlaps substantially with sentence position.

These results support a cautious view: transformer attention patterns exhibit properties consistent with memory-related processing difficulty, but they should not be equated directly with human retrieval mechanisms. Current results lack evidence for sufficiently distinguishing the precise nature of this effect. Understanding how architectural biases and training objectives give rise to these correlates remains an important direction for future research.

Moreover, additional analyses are necessary to disentangle the contribution of sentence position to RTs from potential memory effects. As the estimated attention effects are already very small and given that our measures seem to naturally correlate strongly with sentence position, reliably detecting a memory effect beyond sentence position might be difficult to achieve with corpus-based methods and require significantly larger amounts of data.

On the other hand, our choice of eye-tracking metric might have contributed to this correlation. GPT includes the duration of regressions to preceding words. If regressions are used to reread words that have been imperfectly detected or forgotten (see Booth and Weger, 2013), a large context would imply the existence of more items of this kind and thus lead to more regressions, increasing the GPT measure.[2] Against this backdrop, reconducting the analysis for lower-level features such as first fixation duration and gaze duration might be advisable. Alternatively, moving-window self-paced reading times would provide a measure that is not confounded by regressive eye movements.

Second, the aggregation choice should be revisited: rather than uniformly averaging heads, weighting heads by their functional contribution (e.g. via output projections) may better preserve

[2]We thank an anonymous reviewer for this hint.

Metric	Variant	Scaling	Effect	SE	χ^2	p	q	ΔAIC	ΔBIC
entropy	local	no	2.97	1.14	6.51	0.011	0.033	-5	5
		yes	1.96	1.13	2.97	0.085	0.099	-1	9
	global	no	2.29	1.13	2.97	0.041	0.072	-2	7
		yes	-1.90	1.04	3.35	0.073	0.093	-1	9
distance		no	3.17	1.26	6.07	0.014	0.033	-4	5
		yes	1.96	0.98	4.39	0.036	0.072	-2	8
activation	local	no	3.15	1.25	6.13	0.013	0.033	-5	5
		yes	-2.04	1.08	3.45	0.063	0.088	-2	9
	global	no	3.16	1.26	6.04	0.014	0.033	-4	6
		yes	-2.09	1.04	3.97	0.046	0.072	-2	8
difference	local	no	3.21	1.24	6.43	0.011	0.033	-4	5
		yes	-1.67	1.06	2.38	0.123	0.133	0	9
	global	no	3.51	1.28	7.20	0.007	0.033	-5	5
		yes	-0.47	0.98	0.00	0.935	0.935	2	11

Table 5: Effects of attention-based predictors on GPT in LMEMs. For each metric, we report the fixed-effect estimate (β; per 1 SD increase), standard error (SE), likelihood-ratio χ^2, raw p-value, BH-corrected q-value, and information-criterion differences relative to the corresponding baseline model (ΔAIC, ΔBIC). Highlighted q-values indicate effects surviving multiple-testing correction (q<0.05).

Metric	Baseline	p	q
entropy	position	0.898	0.898
	log(position)	0.029	0.208
distance	position	0.424	0.484
	log(position)	0.086	0.208
activation	position	0.224	0.358
	log(position)	0.104	0.208
difference	position	0.332	0.443
	log(position)	0.086	0.208

Table 6: p-values and BH-corrected q-values for the effect of local unnormalised predictors controlled for word position.

	GPT	freq	len	surp	$\mathbf{H}_{\text{loc}}$	$\mathbf{D}$	$\mathbf{W}_{\text{loc}}$
freq	-0.16						
len	0.18	-0.76					
surp	0.13	-0.61	0.49				
$\mathbf{H}_{\text{loc}}$	0.02	-0.06	0.09	0.03			
$\mathbf{D}$	0.00	0.03	0.02	-0.07	0.93		
$\mathbf{W}_{\text{loc}}$	0.01	0.00	0.04	-0.03	0.94	0.99	
$\mathbf{J}_{\text{loc}}$	0.01	-0.02	0.06	-0.01	0.93	0.98	0.99

Table 7: Pearson correlations between baseline predictors and local unscaled metrics computed for the observations included in the LMEM analysis.

behaviourally relevant structure. Third, to reduce reliance on post-hoc linking assumptions and maybe to disentangle position effects and various candidates for memory costs, a possible path forward would be to incorporate memory-relevant constraints into modelling itself – either by training objectives that penalise costly attention patterns, by developing process-level models that predict RT as a function of model-internal computation or by developing dynamic models of memory access as discussed by Ryu and Lewis (2025).

Finally, caution is advised when generalising our results to the broad class of transformer-based LMs. Previous work relating attention patterns with RTs has predominantly focused on the `GPT-2 small` variant (Oh and Schuler, 2022; Ryu and Lewis, 2022, 2025). However, given that the predictive power of next-word surprisal decreases for larger transformers (Oh et al., 2022), this raises the question of how model size and architecture influence the predictiveness of other metrics, such as attention entropy. To our knowledge, this question remains unexplored, with the notable exception of Timkey and Linzen (2023), who successfully model interference effects using a single-head architecture. Although we observe modest effects beyond surprisal for a larger model than those used in prior work, these findings do not support strong generalisations. Our differences in results might be due to aggregation scheme or architecture.

However, taken together with the reliable effects reported in previous research, our findings are consistent with the possibility that the predictiveness of aggregate metrics decreases as the number of attention heads increases, potentially due to a growing proportion of heads whose computations do not resemble the retrieval operations posited by psycholinguistic theory. Whether these heads play other cognitively interpretable roles or capture linguistic regularities in alternative ways remains unclear, as does whether increases in the number of heads arising from additional layers and those due to a greater number of parallel heads have comparable effects.

7. Limitations

Several limitations should be acknowledged.

First, the analysis is correlational and conducted on a single eye-tracking dataset. It does not demonstrate a causal role of attention mechanisms in human sentence processing and generalisation to other corpora is not guaranteed.

Second, attention weights as computed here are only indirect proxies for representational contribution and may not fully capture the functional role of information flow within the model due to not accounting for residual connections and layer norms.

Third, the transformer architecture permits parallel retrieval operations via multi-head attention, whereas traditional cue-based retrieval theories assume strong constraints for human memory that allow for only a very limited number of elements to be retrieved when processing a word. The cognitive interpretation of multi-head attention therefore remains speculative.

8. Acknowledgements

We would like to thank Benoit Crabbé for his valuable suggestions and Milica Gašić for sharing computational resources. We also thank the three anonymous reviewers for their helpful feedback. Part of this study was conducted on the Heinrich Heine HILBERT HPC.

9. Code Availability

The `R` notebook with the full linear-mixed effects model selection process and the language model training code are available at `https://github.com/filemon11/MITransformer`.

10. Bibliographical References

John R. Anderson, Daniel Bothell, Michael D. Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. 2004. An integrated theory of the mind. *Psychological review*, 111 4:1036–60.

Suhas Arehalli, Brian Dillon, and Tal Linzen. 2023. Syntactic Surprisal From Neural Models Predicts, But Underestimates, Human Processing Difficulty From Syntactic Ambiguities. ArXiv:2210.12187 [cs].

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based Attention Regularization Frees Unintended Bias Mitigation from Lists. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1105–1119. Association for Computational Linguistics.

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(3):255–278.

Douglas Bates, Reinhold Kliegl, Shravan Vasishth, and Harald Baayen. 2018. Parsimonious mixed models.

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.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Robert W. Booth and Ulrich W. Weger. 2013. The function of regressions in reading: Backward eye movements allow rereading. *MEMORY & COGNITION*, 41(1):82–97.

Kenneth P. Burnham and David R. Anderson. 1998. *Model Selection and Inference. A Practical Information-Theoretic Approach*, first edition. Springer Books. Springer, Newy York, New York, USA.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Kate Ehrlich and Keith Rayner. 1983. Pronoun assignment and semantic integration during reading: eye movements and immediacy of processing. *Journal of Verbal Learning and Verbal Behavior*, 22(1):75–87.

Michael A. A. Eskenazi. 2024. Best practices for cleaning eye movement data in reading research. *Behavior Research Methods*, 56(3):2083–2093.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal archive*, 12:23–38.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, pages 94–126. The MIT Press, Cambridge, Massachusetts, USA.

Peter C. Gordon, Randall Hendrick, Marcus Johnson, and Yoonhyoung Lee. 2006. Similarity-based interference during language comprehension: Evidence from eye tracking during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(6):1304–1321.

Peter C. Gordon, Randall Hendrick, and William H. Levine. 2002. Memory-Load Interference in Syntactic Processing. *Psychological Science*, 13(5):425–430.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *NAACL '01: Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001*, pages 1–8, Morristown, New Jersey, USA. Association for Computational Linguistics.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

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

Richard L. Lewis. 1993. *An architecturally-based theory of human sentence comprehension*. Ph.D. thesis, CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE.

Richard L. Lewis. 1996. Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research*, 25(1):93–115.

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 sciences*, 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.

Emma Marsden, Sophie Thompson, and Luke Plonsky. 2018. A methodological synthesis of self-paced reading in second language research. *Applied Psycholinguistics*, 39(5):861–904.

Lukas Mielczarek, Timothée Bernard, Laura Kallmeyer, Katharina Spalek, and Benoit Crabbé. 2025. Modelling expectation-based and memory-based predictors of human reading times with syntax-guided attention. In *Proceedings of the Second Workshop on the Bridges and Gaps between Formal and Computational Linguistics (BriGap-2)*, pages 52–71, Düsseldorf, Germany. Association for Computational Linguistics.

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

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. ArXiv:2212.11185 [cs].

Ralph Radach, Lynn Huestegge, and Ronan Reilly. 2008. The role of global top-down factors in local eye-movement control in reading. *PSYCHOLOGICAL RESEARCH-PSYCHOLOGISCHE FORSCHUNG*, 72(6):675–688.

Ralph Radach and Alan Kennedy. 2013. Eye movements in reading: Some theoretical context. *QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY*, 66(3, SI):429–452.

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

K Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *PSYCHOLOGICAL BULLETIN*, 124(3):372–422.

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, Online. Association for Computational Linguistics.

Soo Hyun Ryu and Richard L. Lewis. 2022. Using transformer language model to integrate surprisal, entropy, and working memory retrieval accounts of sentence processing. *35th Annual Conference on Human Sentence Processing*.

Soo Hyun Ryu and Richard L. Lewis. 2025. Memory for prediction: A Transformer-based theory of sentence processing. *Journal of Memory and Language*, 145:104670.

Robyn Speer. 2022. rspeer/wordfreq: v3.0. [Online; accessed 26-January-2025].

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, Singapore. Association for Computational Linguistics.

Julie A. Van Dyke and Richard L. Lewis. 2003. Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, 49(3):285–316.

Marten van Schijndel and Tal Linzen. 2021. Single-Stage Prediction Models Do Not Explain the Magnitude of Syntactic Disambiguation Difficulty. *Cognitive Science*, 45(6):e12988.

Shravan Vasishth and Richard L. Lewis. 2004. Modeling sentence processing in ACT-R. In *Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together*, IncrementParsing '04, pages 82–87, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. 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, Singapore. Association for Computational Linguistics.

11. Language Resource References

Luke, Steven G. and Christianson, Kiel. 2018. *The Provo Corpus*. Open Science Framework.

Merity, Stephen and Xiong, Caiming and Bradbury, James and Socher, Richard. 2016. *WikiText-103: A Large Language Modeling Corpus*. Salesforce Research.

Natalia Silveira and Timothy Dozat and Marie-Catherine de Marneffe and Samuel Bowman and Miriam Connor and John Bauer and Christopher D. Manning. 2014. *Universal Dependencies English Web Treebank*. Universal Dependencies.

A. Predictor Correlations

length.1	-0.03	0.07	-0.03	-0.07	0.00	0.03	0.04	0.05		
surprisal.1	-0.00	0.06	-0.07	0.00	-0.07	-0.05	-0.04	-0.02	0.55	
surprisal.2	0.01	0.04	-0.04	0.05	-0.08	-0.06	-0.06	-0.06	-0.07	-0.00
	GPT	frequency	length	surprisal	$\mathbf{H}_{\mathrm{loc}}$	$\mathbf{D}$	$\mathbf{W}_{\mathrm{loc}}$	$\mathbf{J}_{\mathrm{loc}}$	length.1	surprisal.1

Table 8: Pearson correlations between spillover versions of baseline predictors and the remaining predictors for the likelihood ratio tests of the local unscaled attention metrics, computed for the observations included in the tests.

$\tilde{\mathbf{H}}_{\mathrm{loc}}$	0.05	-0.24	0.13	0.27	-0.16	-0.13	-0.07			
$\tilde{\mathbf{D}}$	-0.04	0.16	-0.15	-0.13	-0.08	-0.06	-0.01	0.31		
$\tilde{\mathbf{W}}_{\mathrm{loc}}$	-0.02	0.04	-0.08	0.00	-0.02	-0.02	-0.00	0.51	0.86	
$\tilde{\mathbf{J}}_{\mathrm{loc}}$	0.02	-0.19	0.09	0.22	0.02	0.01	-0.03	0.39	0.51	0.68
								$\tilde{\mathbf{H}}_{\mathrm{loc}}$	$\tilde{\mathbf{D}}$	$\tilde{\mathbf{W}}_{\mathrm{loc}}$
$\mathbf{H}_{\mathrm{glob}}$	0.01	-0.01	0.06	-0.04	0.03	-0.05	-0.07			
$\mathbf{D}$	0.00	0.03	0.02	-0.07	0.03	-0.05	-0.06	0.95		
$\mathbf{W}_{\mathrm{glob}}$	0.00	0.03	0.03	-0.06	0.04	-0.04	-0.06	0.95	1.00	
$\mathbf{J}_{\mathrm{glob}}$	0.01	-0.00	0.05	-0.03	0.04	-0.02	-0.05	0.93	0.99	0.99
								$\mathbf{H}_{\mathrm{glob}}$	$\mathbf{D}$	$\mathbf{W}_{\mathrm{glob}}$
$\tilde{\mathbf{H}}_{\mathrm{glob}}$	0.00	-0.08	0.03	0.02	-0.11	-0.12	-0.06			
$\tilde{\mathbf{D}}$	-0.04	0.16	-0.15	-0.13	-0.08	-0.06	-0.01	0.82		
$\tilde{\mathbf{W}}_{\mathrm{glob}}$	-0.04	0.17	-0.16	-0.15	-0.06	-0.04	0.01	0.86	0.97	
$\tilde{\mathbf{J}}_{\mathrm{glob}}$	0.02	-0.13	0.06	0.16	0.00	0.01	-0.06	0.50	0.51	0.42
	GPT	frequency	length	surprisal	length.1	surprisal.1	surprisal.2	$\tilde{\mathbf{H}}_{\mathrm{glob}}$	$\tilde{\mathbf{D}}$	$\tilde{\mathbf{W}}_{\mathrm{glob}}$

Table 9: Pearson correlations between candidate metrics (local scaled, global unscaled and global scaled) and the other predictors for the observations included in the LMEM analyses.

B. Dependency length correlations

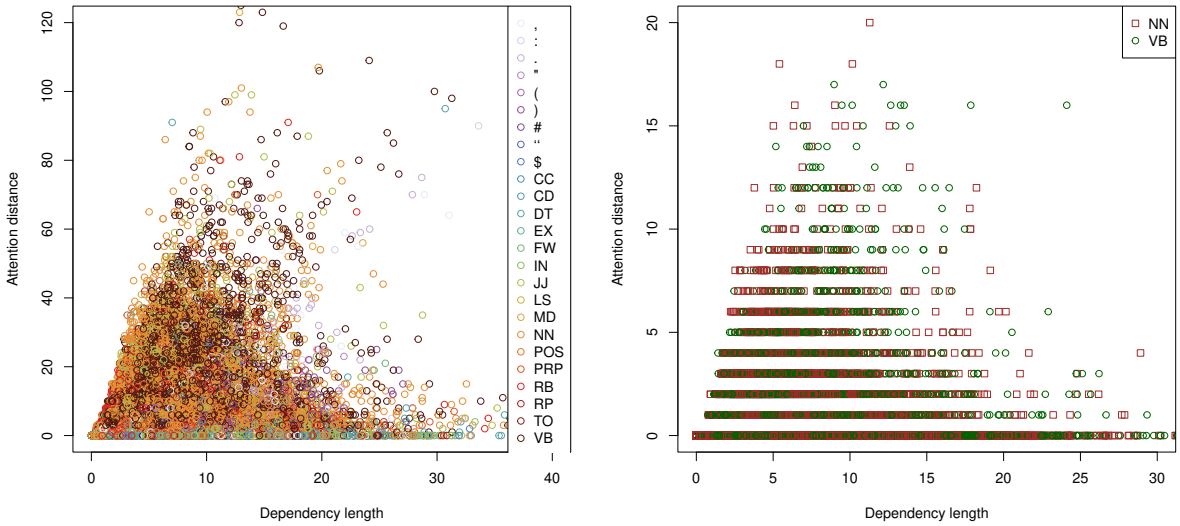

Figure 2: Dependency lengths extracted from the UD English Web Treebank train split plotted against local attention distance generated by our LM. Left plot: dependency length for a token was calculated by summing the surface distances to all items left of it that are connected to it via a relation (both dependents and head). Right plot: distance for a relation was estimated by the number of intervening nouns/verbs and only nouns/verbs were assigned a cost. A few outliers are not shown. Pearson correlation coefficient for the second setting: 0.13.

C. Model Training

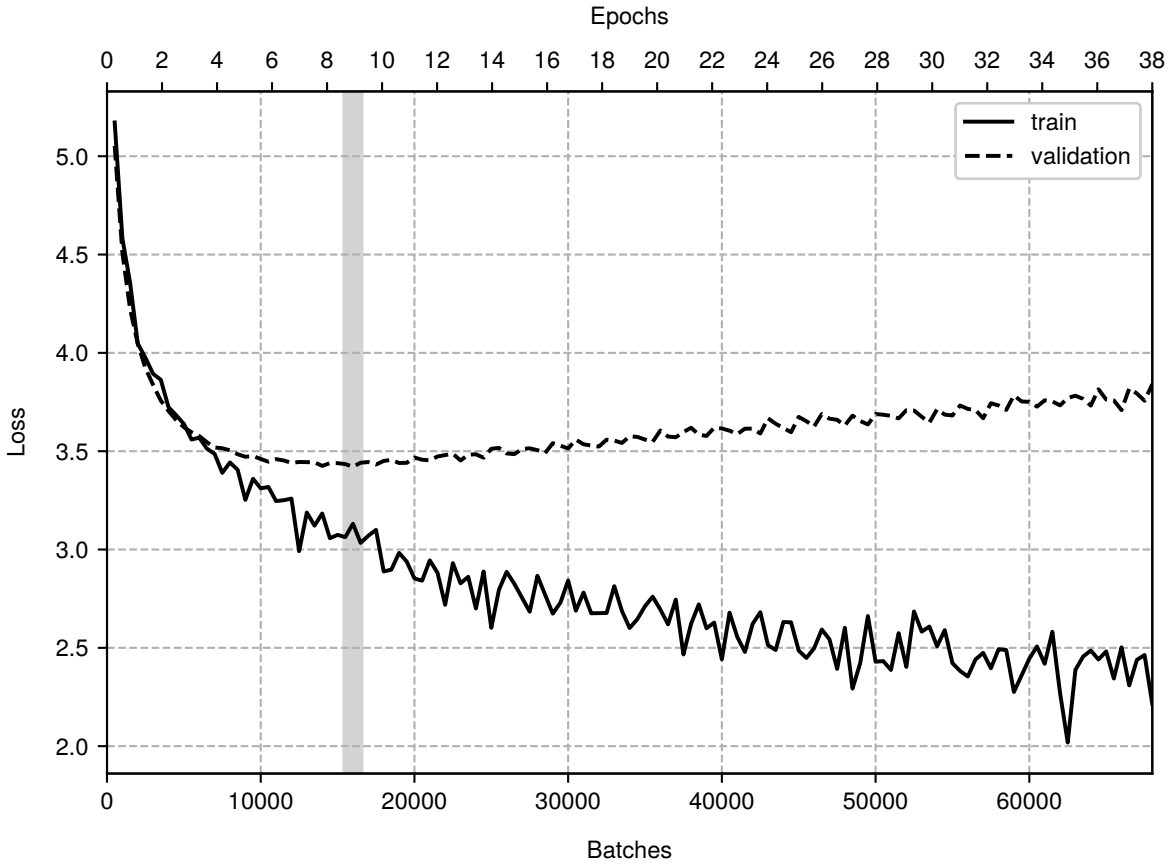

Figure 3: Development of the loss on the training and the development split of the Wikitext corpus during training. The gray vertical bar at 8.8 epochs indicates the region of minimal validation loss which determines the model checkpoint used for psycholinguistic evaluation. One epoch consists of 3 574 204 training sentences. Training was only performed on sentences with a minimum of 3 and a maximum of 60 words. The validation split contains 15 312 sentences.

Resources	Value
batch size	1920 (160 · 12 gradient accumulation)
evaluation interval	every 500 batches
resources	HHU HPC Hilbert
devices	4 Nvidia A100 GPUs
threads (dataloader workers)	8 per process
processor	AMD EPYC 75F3, 2.95 GHz
Peak RAM usage	89.98 Gb DDR4
duration until minimum validation loss reached	10.53 h

Table 10: Model training resources

D. Marginal Effects Plots

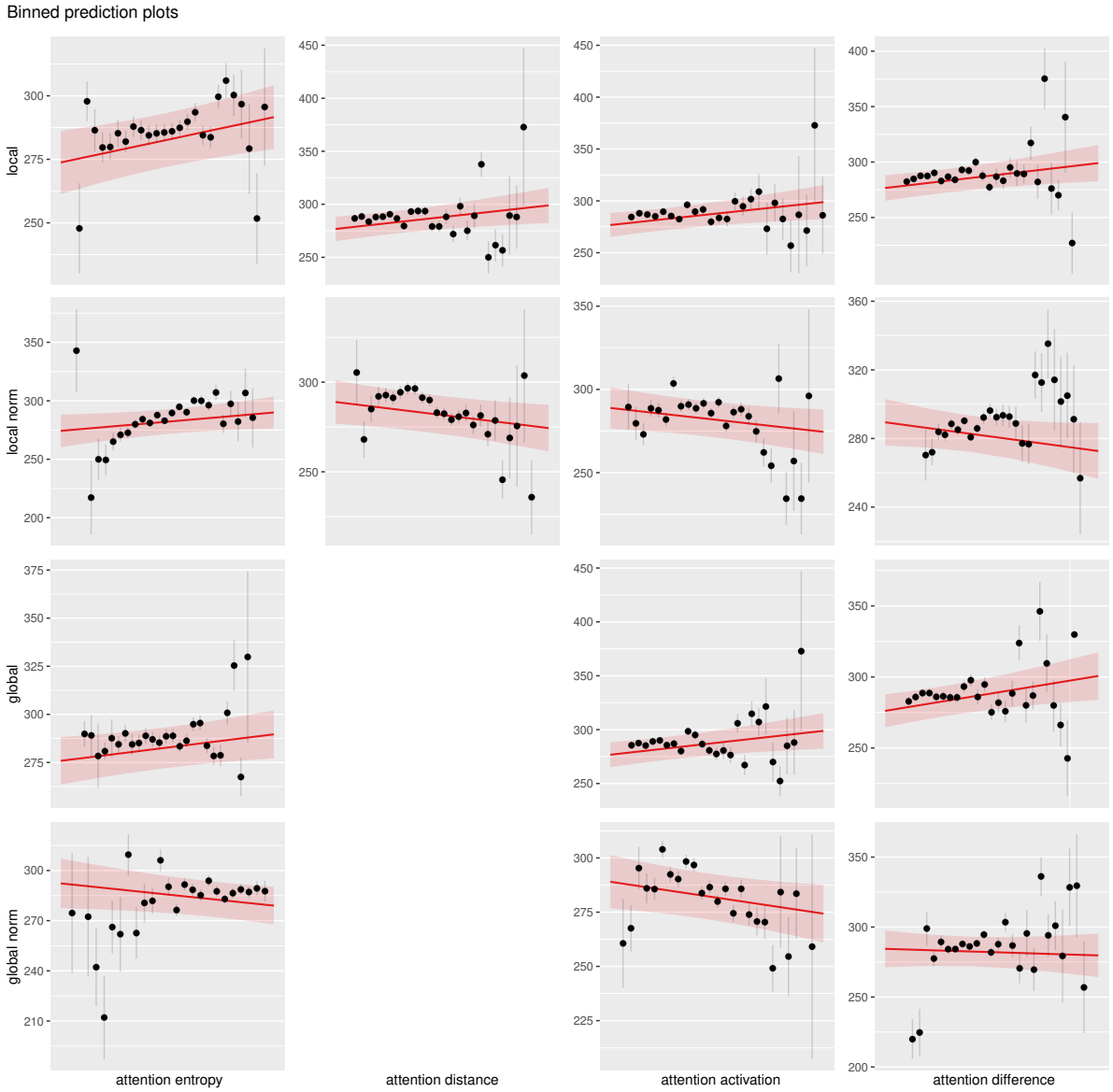

Figure 4: Population-level predicted go-past times as a function of the attention measures (x-axis), based on fixed effects from the mixed-effects models, with other covariates held constant at typical values. Points show mean observed go-past times within evenly spaced bins of the predictor. Shaded areas and error bars indicate 95% confidence intervals.