Associative Memory Updates in Language Models Predict Human Reading Times

Christian Clark
Center for Data Science
New York University
christian.clark@nyu.edu

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

Abstract

The self-attention mechanism used in transformer language models implements a form of associative, cue-based memory retrieval, which has been argued to resemble human sentence processing. However, while it is well established that surprisal estimates from transformers are predictive of human comprehension difficulty, it has been less clear how these models' internal memory operations might relate to human sentence processing. We propose to connect the two by measuring transformers' *associative memory updates*, which are calculated as the difference between each word's predicted and actual representation within self-attention. We compute memory updates from three types of language models: (1) standard transformers; (2) CBR-RNN, a cognitively motivated model using a single self-attention head; and (3) CBR-RNN-M, a novel architecture that replaces the self-attention in CBR-RNN with a bounded-size memory cell. Associative memory updates are found to predict naturalistic human reading times over and above surprisal baselines on several psycholinguistic corpora, with evidence for more consistent effects from CBR-RNN and CBR-RNN-M than from standard transformers.

1 Introduction

An important goal in psycholinguistic research is to develop robust broad-coverage models of human sentence processing. Such models are helpful for understanding the steps humans use to incrementally build complex mental representations as they read text passages or listen to speech.

In recent years, autoregressive language models (LMs) such as transformers (Vaswani et al., 2017) have become important tools for this task. Surprisal theory (Hale, 2001; Levy, 2008) posits that the difficulty of comprehending a word w_i is proportional to its surprisal or negative log probability given the preceding context, $-\log \mathrm{P}(w_i \mid w_1, \ldots, w_{i-1})$. Because autoregressive LMs directly estimate surprisals, this creates a natural interface between their outputs and measures of human sentence processing difficulty such as self-paced reading times, gaze durations from eye tracking, or blood oxygenation level–dependent signals from functional magnetic resonance imaging. A wide range of studies have shown strong predictions of these measures from LM-based surprisal measures (e.g., Wilcox et al., 2023; Shain et al., 2024) with many of the strongest estimates coming from transformer-based models (Wilcox et al., 2020; Merkx and Frank, 2021, though cf. Michaelov et al., 2024).

Surprisal theory is intended to describe human sentence processing on the computational level (Marr, 1982), meaning that it characterizes the inputs, outputs, and goals of comprehension but does not try to describe the cognitive representations or operations that achieve these goals (Marr's algorithmic level). Despite strong evidence for a computational-level connection, the degree to which transformers or other LMs characterize broad-coverage sentence processing on an algorithmic level remains less certain. The strongest case for an algorithmic connection between transformers and human sentence processing involves the attention mechanism in transformers, which has been argued to implement a form of cue-based memory retrieval similar to independently motivated theories of human sentence processing (Ryu and Lewis, 2021, 2025; Oh and Schuler, 2022; Timkey and Linzen, 2023). Still, the link between cue-based retrieval in transformers and human comprehension difficulty is not entirely clear, despite some progress towards establishing a link in these prior works.

This paper introduces a new linking hypothesis between memory retrieval in transformers and humans based on the idea of *associative memory updates*. Sentence processing in both humans and transformers involves updates to associations be-

tween cues and targets in memory; we propose that processing difficulty in humans is connected to the magnitude of these updates, and that transformer-like models can be used to measure this. While surprisal can also be understood as a measure of each word's memory update, our proposal differs in that we measure memory updates at the level of transformers' internal representations, i.e., key and query vectors. This may be a useful stepping stone toward connecting transformer-like language models with algorithmic-level psycholinguistic models involving associative memory (e.g., Rasmussen and Schuler, 2018; Keshev et al., 2025).

The experiments in this paper compute associative memory updates from three types of models: (1) standard transformers; (2) CBR-RNN (Timkey and Linzen, 2023), a model which combines a single transformer-style self-attention head with simple recurrent network; and (3) CBR-RNN-M, a model we introduce which replaces the self-attention in CBR-RNN with a fixed-size associative memory cell. Predictors based on memory updates are used to predict human reading times from naturalistic self-paced reading and eye tracking corpora, to test whether the memory updates predict these measures over and above standard baseline predictors such as surprisal. Significant effects on many corpora are observed across all three types of models, with evidence for more consistent effects when memory updates are taken from CBR-RNN and CBR-RNN-M.[1]

2 Related Work

As mentioned in the previous section, Ryu and Lewis (2021, 2025) and Oh and Schuler (2022) are previous works that examine potential connections between transformers' self-attention mechanism and human sentence processing difficulty. These authors propose psycholinguistic predictors based on attention entropy, i.e., the diffuseness of attention to previous tokens, which can potentially serve as an index of memory interference. Oh and Schuler (2022) additionally introduce several distance-based predictors quantifying the shift in attention weights over successive tokens. The memory update measure proposed in the present work also measures memory shifts over successive tokens; however, our measure differs in being computed over the actual memory items (key and value vectors) instead of attention weights, and in being straightforwardly extensible to LMs that use a bounded memory cell (see Experiment 3).

The present work is also similar in spirit to recent work applying models of predictive coding to human sentence processing (Rao and Ballard, 1999; Millidge et al., 2024; Ohams et al., 2026). We propose that memory updates in LMs like transformers, defined according to their word-by-word prediction error, can be linked to processing difficulty.

3 Associative Memory Updates

This section introduces the associative memory update measure which is used as a psycholinguistic predictor in the experiments. To motivate the definition of this predictor, we first give a general background on associative memory models.

3.1 Background: Outer-Product Models of Associative Memory

Models of associative memory describe how *cues* are connected to *targets* in the brain, capturing the familiar idea from Hebbian theory that "neurons that fire together, wire together" (Hebb, 1949; Shatz, 1992). A common method for implementing such a model involves representing cues and targets as vectors, and then defining encoding and retrieval operations in terms of vector and matrix products. For example, given a memory cell $\mathbf{M} \in \mathbb{R}^{d_v \times d_k}$ storing current cue-target associations, the encoding of a new association between cue vector $\mathbf{k} \in \mathbb{R}^{d_k}$ and target vector $\mathbf{v} \in \mathbb{R}^{d_v}$ can be described by a Hebbian learning rule:[2]

$$\mathbf{M}' = \mathbf{M} + \mathbf{v}\mathbf{k}^\top \tag{1}$$

The target $\mathbf{v}$ associated with a cue $\mathbf{k}$ can later be retrieved via multiplication with the memory cell:

$$\mathbf{v} = \mathbf{M}\mathbf{k} \tag{2}$$

If $\mathbf{k}$ is orthogonal to other cue vectors, $\mathbf{v}$ will be recovered perfectly. Otherwise, interference with other stored memories will occur, resulting in a partially distorted target. In psycholinguistics, interference from cues with overlapping features has been used to give associative memory–based accounts of surprisal and center embedding effects (Rasmussen and Schuler, 2018), and of item distortion and confusion effects (Keshev et al., 2025).

[1] Code used for this paper's experiments will be released upon publication.

[2] The variables $\mathbf{k}$ and $\mathbf{v}$ are used to draw attention to the analogy with key and value vectors in transformer attention.

The definition of associative memory updates used in the following experiments is related to an alternative learning rule to Equation (1), known as the Widrow-Hoff delta rule (Widrow and Lehr, 1990):

$$\mathbf{M}' = \mathbf{M} + \beta(\mathbf{v} - \mathbf{M}\mathbf{k})\mathbf{k}^\top \tag{3}$$

This learning rule, which can be interpreted as a form of gradient descent, updates the memory cell based on a computed prediction error: the difference between the desired target $\mathbf{v}$ and the cell's current prediction $\mathbf{M}\mathbf{k}$, which is the target currently associated with the cue $\mathbf{k}$. The parameter β is used to weight the memory update. The delta rule has been used in classic cognitive modeling frameworks like Parallel Distributed Processing (Rumelhart et al., 1986) and the Temporal Context Model (Howard and Kahana, 2002).

We define the *associative memory update* from Equation (3) as

$$\text{MemUpdate}(\mathbf{k}, \mathbf{v}) = \|(\mathbf{v} - \mathbf{M}\mathbf{k})\mathbf{k}^\top\|_\text{F}, \tag{4}$$

where $\|\cdot\|_\text{F}$ is the Frobenius norm. This is a scalar value measuring magnitude of the necessary update to $\mathbf{M}$ in order to pair $\mathbf{v}$ with $\mathbf{k}$.

3.2 Measuring Associative Memory Updates in Transformers

Although standard transformers do not explicitly maintain a memory cell like $\mathbf{M}$, it is possible to compute an associative memory update value, similar to Equation (4), from each attention head at each new token w_i. Specifically, if $\mathbf{k}_i \in \mathbb{R}^{d_k}$ and $\mathbf{v}_i \in \mathbb{R}^{d_v}$ are the current token's key and value vectors and $\mathbf{K}_{<i} \in \mathbb{R}^{(i-1)\times d_k}$ and $\mathbf{V}_{<i} \in \mathbb{R}^{(i-1)\times d_v}$ are matrices containing the key and value vectors of all tokens preceding the current token index i, the memory update at token w_i can be computed as

$$\text{MemUpdate}(\mathbf{k}_i, \mathbf{v}_i) = \tag{5}$$
$$\|(\mathbf{v}_i - \mathbf{V}_{<i}^\top \sigma(\frac{\mathbf{K}_{<i}\,\mathbf{k}_i}{\sqrt{d}}))\mathbf{k}_i^\top\|_\text{F},$$

where $\sigma(\cdot)$ is the softmax function. This memory update measures the gap between w_i's actual value vector $\mathbf{v}$ and the predicted value vector based on the contents of the key-value cache before w_i was observed. We hypothesize that larger values of this memory update will be associated with increased comprehension difficulty in humans.

4 Experiment 1: Memory Updates in Standard Transformers

Experiment 1 tests whether associative memory updates derived from two standard transformer models are predictive of human reading times. The language models used for these experiments are GPT2-small (Radford et al., 2019), which has 124M parameters, and OPT-125M (Zhang et al., 2022). GPT2-small was chosen because previous work shows that its surprisal measures make stronger reading time predictions than surprisal from larger models (Oh et al., 2022; Oh and Schuler, 2023); OPT-125M was used to test whether results observed from GPT2-small replicate in another model instance of similar scale. Both models include 12 layers with 12 attention heads per layer.

4.1 Psycholinguistic Corpora

Five English-language psycholinguistic corpora were used in this experiment: Brown (Smith and Levy, 2013) and Natural Stories (Futrell et al., 2021), which contain self-paced reading (SPR) times; and Dundee (Kennedy et al., 2003), GECO (Cop et al., 2017) and Provo (Luke and Christianson, 2018), which contain eye tracking measures including first-pass (FP) and go-past (GP) durations. Natural Stories includes data from 181 subjects who read a collection of 10 naturalistic stories with a total of 10,256 words. Brown includes data from 35 subjects who read a collection of 13 passages from the Brown corpus with a total of 7,188 words. Dundee includes data from 10 subjects who read a collection of 67 newspaper editorials with a total of 51,501 words. GECO includes data from 14 subjects who read an Agatha Christie novel (Christie, 1920) with a total of 56,441 words. Lastly, Provo includes data from 84 subjects who read a collection of 55 passages including news articles, magazines, and fiction with a total of 2,746 words.

Each corpus was split into fit, exploratory, and held-out partitions, respectively containing 50%, 25%, and 25% of the data points. Fit partitions were used to fit regression models; exploratory partitions were used to select the layers or individual heads that make the strongest reading time predictions; and held-out partitions were used for the final evaluation and significance testing.

4.2 Memory Update Predictors

The memory update predictor introduced in Section 3.2 is defined for a single attention head, but

GPT2-small and OPT-125M each have 144 attention heads. Three methods were therefore considered for deriving memory update predictors from these models. First, an average memory update was computed by averaging the updates of all 144 attention heads. Second, a per-layer memory update was computed for each layer by averaging over its 12 attention heads. Third, memory updates of the individual attention heads were computed. For the latter two methods, memory updates from each layer or individual head were evaluated on the exploratory partitions of the psycholinguistic corpora, and then the best-performing layer or head was tested on held-out data.

4.3 Procedures

Linear mixed-effects (LME Bates et al., 2015) models were fit to each corpus to evaluate whether each memory update measure predicted human reading times above and beyond standard baseline predictors. For SPR corpora, the baseline predictors were word length, word position, unigram surprisal, transformer surprisal (from either GPT2 or OPT, depending on which model's memory updates were being evaluated), surprisal from CBR-RNN, and surprisal from CBR-RNN-M. The latter two surprisal predictors were included to allow a fair comparison between the results from this experiment and those from Experiments 2 and 3, which evaluate memory updates from CBR-RNN and CBR-RNN-M. Evaluations on eye-tracking corpora included all of these baseline predictors plus a predictor marking whether the previous word was fixated. Unigram surprisal was estimated using the KenLM toolkit (Heafield et al., 2013) on the OpenWebText corpus (Gokaslan and Cohen, 2019), which contains 6.5 billion words. All predictors were centered and z-scaled prior to fitting regression models. Models included per-subject random intercepts; random slopes were initially included but removed to prevent convergence issues. The final regression formulas for self-paced reading and eye-tracking measures are given in Appendix A.

The evaluation metrics we report are the change in per-datapoint log likelihood (ΔLL) and the change in mean squared error (ΔMSE) between the baseline and main LME models. To evaluate whether improvements were significant, permutation tests were performed on the per-datapoint squared errors.

4.4 Results

Table 1 presents the main results from the GPT2 and OPT memory update predictors. Along with ΔLL and ΔMSE, this table reports effect sizes, which are the predicted increase in reading times (measured in ms) per standard deviation of the memory update predictors. Although the specific results vary to some degree by LM (GPT2 or OPT) and aggregation method (average over all attention heads, per-layer average, or single attention head), statistically significant improvements are seen in the majority of cases.

When memory updates averaged over all attention heads are used (Tables 1a and 1d), significant effects on Dundee (GP and FP), Provo (GP and FP), and GECO (GP) are seen from both LMs, with OPT's memory update also showing significant effects for the two self-paced reading corpora.

The layers with the best (lowest) average MSE were layer 1 in GPT2 and layer 2 in OPT. Tables 1b and 1e show the results from these layers. Memory updates from both models' best layers make stronger predictions on the Natural Stories corpus than the memory updates from the full-model average. Other results are mixed, with some improvements on the Provo corpus but weakened effects on other eye-tracking corpora.

The best individual attention head from GPT2 (Table 1c), which is in the model's seventh layer, shows a different pattern of results from the other predictors, with strong effects on Brown, Natural Stories, Dundee, and GECO GP, but no significant effects on other corpora. The best head from OPT (Table 1f), located in the second layer, makes especially strong predictions on Natural Stories and Provo GP.

Although our experiments were motivated by the hypothesis that larger memory updates should be associated with increased processing difficulty in humans, the tables show a mix of positive and negative effects. To show how these effects vary over attention heads in the models, Figure 1 plots the effect sizes of all the per-layer predictors. An interesting trend is visible across GPT2 layers: middle and late layers tend to have more positive effects than early layers.[3] OPT shows this trend for some corpora (e.g., Dundee GP), but with less consistency. One hypothesis for the positive trend in effect sizes across GPT2 layers is that the layers'

[3] Linear regression curves fit to the effect sizes over layers in Figure 1a have positive slopes in 8/8 cases.

Corpus	Effect	ΔLL	ΔMSE
Brown SPR	1.74	5.31×10^{-6}	−0.53
NS SPR	1.75	-6.03×10^{-5}	2.31
Dundee FP	2.94	1.74×10^{-4}	−5.87 **
Dundee GP	3.99	1.65×10^{-4}	−21.03 ***
GECO FP	−1.92	-7.99×10^{-6}	0.37
GECO GP	0.49	6.89×10^{-6}	−2.18 ***
Provo FP	−4.66	4.74×10^{-4}	−34.71 ***
Provo GP	−0.38	1.20×10^{-5}	−3.68 ***

(a) GPT2, Update averaged over all heads

Corpus	Effect	ΔLL	ΔMSE
Brown SPR	−0.85	2.65×10^{-6}	−0.14
NS SPR	−5.25	5.76×10^{-4}	−24.00 ***
Dundee FP	1.16	-1.64×10^{-5}	0.56
Dundee GP	1.76	2.25×10^{-5}	−2.82 +
GECO FP	−2.32	-3.72×10^{-5}	1.89 *
GECO GP	−0.84	-1.13×10^{-5}	3.52 ***
Provo FP	−6.62	6.12×10^{-4}	−43.82 ***
Provo GP	−3.38	6.53×10^{-5}	−21.00 ***

(b) GPT2, Strongest layer (Layer 1)

Corpus	Effect	ΔLL	ΔMSE
Brown SPR	6.29	6.43×10^{-4}	−31.15 ***
NS SPR	9.03	1.10×10^{-3}	−44.85 ***
Dundee FP	6.91	9.99×10^{-4}	−33.56 ***
Dundee GP	9.57	7.02×10^{-4}	−87.42 ***
GECO FP	−0.46	3.86×10^{-6}	−0.21
GECO GP	4.58	6.94×10^{-5}	−22.56 ***
Provo FP	1.27	-8.56×10^{-5}	6.29
Provo GP	10.91	-1.16×10^{-4}	38.14

(c) GPT2, Strongest individual head (Layer 7, Head 10)

Corpus	Effect	ΔLL	ΔMSE
Brown SPR	3.45	9.82×10^{-5}	−5.33 *
NS SPR	5.89	1.38×10^{-4}	−5.92 **
Dundee FP	2.93	1.94×10^{-4}	−6.47 ***
Dundee GP	4.15	1.85×10^{-4}	−23.59 ***
GECO FP	−2.01	9.92×10^{-6}	−0.55
GECO GP	1.01	1.35×10^{-5}	−4.28 ***
Provo FP	−6.67	6.12×10^{-4}	−43.88 ***
Provo GP	−4.64	1.17×10^{-4}	−37.25 ***

(d) OPT, Update averaged over all heads

Corpus	Effect	ΔLL	ΔMSE
Brown SPR	5.21	4.22×10^{-4}	−20.37 ***
NS SPR	11.94	1.46×10^{-3}	−58.66 ***
Dundee FP	2.09	1.33×10^{-4}	−4.40 **
Dundee GP	2.14	7.75×10^{-5}	−9.98 ***
GECO FP	−1.37	-9.65×10^{-6}	0.45
GECO GP	1.28	1.65×10^{-5}	−5.30 **
Provo FP	−6.72	5.41×10^{-4}	−38.27 ***
Provo GP	−5.91	1.52×10^{-4}	−49.73 ***

(e) OPT, Strongest layer (Layer 2)

Corpus	Effect	ΔLL	ΔMSE
Brown SPR	4.86	2.38×10^{-4}	−11.72 **
NS SPR	14.94	3.00×10^{-3}	−121.27 ***
Dundee FP	−0.60	-2.50×10^{-5}	0.83
Dundee GP	−1.67	-2.87×10^{-5}	3.88
GECO FP	0.13	-2.76×10^{-7}	0.01
GECO GP	4.53	3.78×10^{-5}	−11.91 **
Provo FP	−7.11	6.91×10^{-4}	−50.59 ***
Provo GP	−10.36	2.58×10^{-4}	−89.82 ***

(f) OPT, Strongest individual head (Layer 2, Head 8)

Table 1: Regression results using memory update predictors from GPT2 and OPT results. NS is short for Natural Stories. Significance levels are * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, with marginally significant results ($p < 0.1$) marked with +.

LM	Source of Update Predictor	χ^2
GPT2	Average over all heads	70.51 ***
GPT2	Strongest layer	49.37 ***
GPT2	Strongest individual head	73.69 ***
OPT	Average over all heads	90.20 ***
OPT	Strongest layer	97.30 ***
OPT	Strongest individual head	62.54 ***

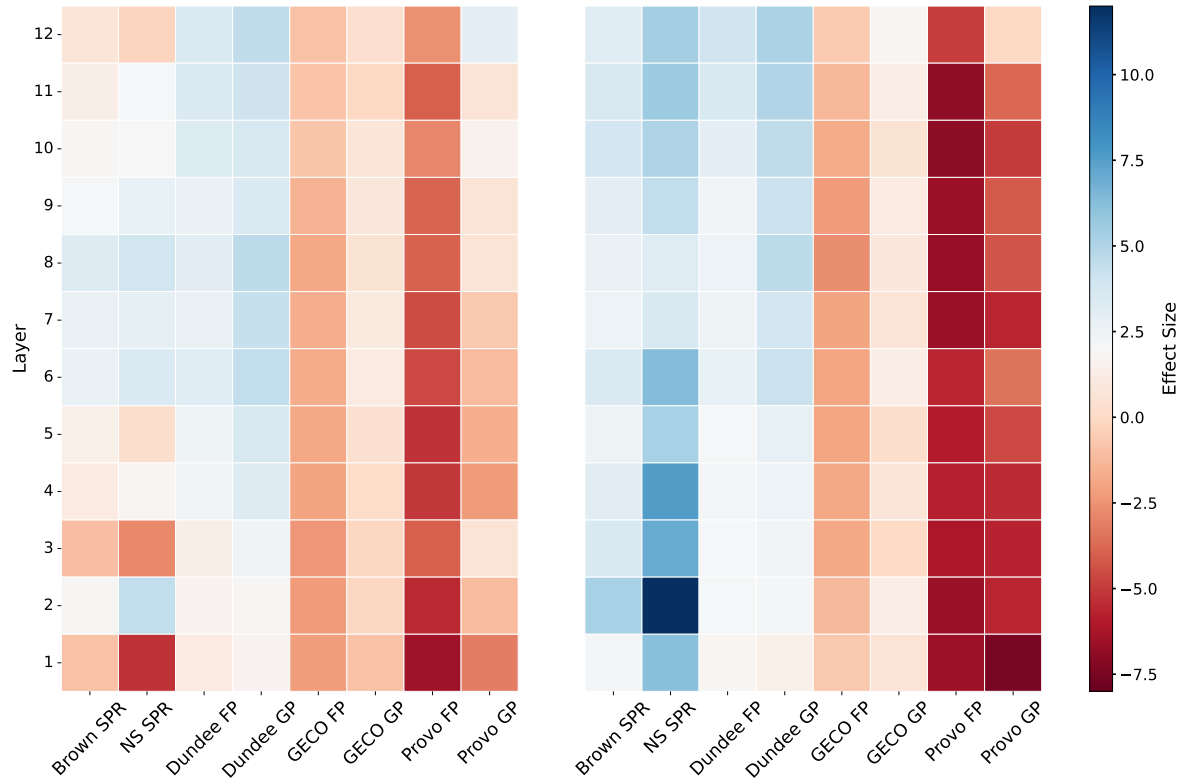
Table 2: Significance results from each transformer-derived memory update predictor, aggregated over psycholinguistic corpora using Fisher's method.

memory updates are sensitive to different types of information—possibly lexical information in earlier layers and structural or semantic information in later layers.[4] There is evidence from other work that surprisal predictors are oversensitive to differences in surface form (Holtzman et al., 2021; Giulianelli et al., 2023) and not sensitive enough to structural updates (Van Schijndel and Linzen, 2021; Huang et al., 2024), so it is possible that memory updates from different model layers compensate for surprisal in different ways.

To give a bird's-eye view of the main results, Fisher's method was used to compute a single test statistic for each of the six GPT2- and OPT-derived memory update predictors, based on the combined significance results across all the psycholinguistic corpora. Table 2 presents the χ^2 statistic for each predictor; each was significant at the Bonferroni-corrected α level of $0.001/8 = 0.000125$.[5]

[4] See Kuribayashi et al. (2025) for evidence that LM layers correspond to different processing stages in humans.

[5] Dividing by 8 accounted for the six memory update pre-



(a) GPT2

(b) OPT

Figure 1: Effect sizes of per-layer memory update predictors from GPT2 and OPT.

5 Experiment 2: Memory Updates in an LM with a Single Attention Head

This experiment tests the effects of memory updates from the Cue-Based Retrieval Recurrent Neural Network (CBR-RNN; Timkey and Linzen, 2023), a cognitively motivated architecture that only uses a single self-attention head. Working with a single attention head brings this model closer to cognitive models of cue-based retrieval such as the Lewis and Vasishth (2005) ACT-R model. It also circumvents the potential issue of attention heads learning specialized, idiosyncratic functions like attention sinks which are seen in larger LMs (Xiao et al., 2024). Because of the relative simplicity of CBR-RNN, its memory updates are predicted to map to human sentence processing difficulty more transparently than updates from standard transformers.

5.1 CBR-RNN Language Model

CBR-RNN combines a simple recurrent neural network (Elman, 1990) with a self-attention mechanism similar to standard transformers. Only one model layer with a single attention head is used, in order to mirror cognitive models of cue-based retrieval such as the Lewis and Vasishth (2005) ACT-R model.

Given the current token w_i, CBR-RNN first calculates an embedding $\mathbf{x}_i = \mathbf{W}_\mathrm{E}\delta_{w_i}$, where $\mathbf{W}_\mathrm{E}$ is an embedding matrix and δ_{w_i} is a one-hot vector representation of w_i. To compute the query vector $\mathbf{q}_i$, the current word embedding and previous hidden state vector $\mathbf{h}_{i-1}$ are multiplied by the weight matrix $\mathbf{W}_\mathrm{Q}$ and then rotary position embeddings (RoPE Su et al., 2024) are applied:[6]

$$\mathbf{q}_i = \mathrm{RoPE}(\mathbf{W}_\mathrm{Q}[\mathbf{x}_i\ \mathbf{h}_{i-1}]) \tag{6}$$

The attention output $\mathbf{a}_i$ is computed by taking a weighted average of all previous value vectors $\mathbf{V}_{<i}$ based on the similarity of previous key vectors $\mathbf{K}_{<i}$ to $\mathbf{q}_i$:

$$\mathbf{a}_i = \mathbf{V}_{<i}^\top \sigma\left(\frac{\mathbf{K}_{<i}\,\mathbf{q}_i}{\sqrt{d_k}}\right), \tag{7}$$

dictors tested in this experiment plus the two additional predictors from Experiments 2 and 3.

[6]The original CBR-RNN model does not use RoPE. We include it because it slightly improves perplexity, and including positional information in memory vectors is common in cognitive models like the Temporal Context Model (Howard and Kahana, 2002).

Note that this differs from standard transformer attention in that the key and value vectors for the current token are not attended to.[7]

The concatenated word embedding, query, attention output, and hidden state are then passed through multiple layers to determine the current key, value, and hidden state vectors:

$$\mathbf{o}_i = \tanh(\mathrm{LN}(\mathbf{W}_\mathrm{O}[\mathbf{x}_i\ \mathbf{q}_i\ \mathbf{a}_i\ \mathbf{h}_{i-1}])) \quad (8)$$

$$\mathbf{k}_i = \mathrm{RoPE}(\mathbf{W}_\mathrm{K}\,\mathbf{o}_i) \quad (9)$$

$$\mathbf{v}_i = \mathbf{W}_\mathrm{V}\,\mathbf{o}_i \quad (10)$$

$$\mathbf{h}_i = \tanh(\mathrm{LN}(\mathbf{W}_\mathrm{H}\,\mathbf{o}_i)) \quad (11)$$

Here, LN denotes layer normalization, and $\mathbf{W}_\mathrm{O}$, $\mathbf{W}_\mathrm{K}$, $\mathbf{W}_\mathrm{V}$, and $\mathbf{W}_\mathrm{H}$ are additional weight matrices. The new key and value vectors are appended to the key-value cache. Finally, the updated hidden state and weight matrix $\mathbf{W}_\mathrm{Y}$ are used to calculate the output distribution over tokens in the vocabulary:

$$\mathbf{y}_i = \sigma(\mathbf{W}_\mathrm{Y}\,\mathbf{h}_i) \quad (12)$$

The implementation of CBR-RNN tested in this experiment uses a vocabulary of 30,000 subword tokens based on byte-pair encoding (Sennrich et al., 2016).[8] Five randomly initialized instances of this model were trained on Wikitext-103 (Merity et al., 2017) for a maximum of 20 epochs; training each model took approximately three days on a single V100 GPU. The model used a hidden dimension of 256 and included a total of 17.4M parameters.

5.2 Procedures

Five pairs of LME models were fit to each corpus following the Experiment 1 procedures. Each model pair used surprisal and memory update predictors from a different randomly initialized instance of CBR-RNN, along with the baseline predictors from Experiment 1.[9] To evaluate whether improvements were significant, a permutation test was performed on the per-datapoint squared errors concatenated over the five pairs of models.

5.3 Results

The results from Experiment 2 are presented in Table 3. Effects, changes in log likelihood, and changes in mean squared error are averaged over the five pairs of LME models. Significant MSE reductions are seen on Brown, Dundee (FP and GP), and GECO (FP and GP). ΔMSE values are generally of a comparable magnitude to those from GPT2 and OPT memory updates (Table 1). Notably, all effects from CBR-RNN memory updates are positive, meaning that larger associative memory updates are consistently associated with increased reading times beyond the baseline predictors. This contrasts with the effects seen from GPT2- and OPT-based predictors, suggesting that CBR-RNN memory updates reflect distinct sources of reading time difficulty beyond surprisal.

Corpus	Effect	ΔLL	ΔMSE
Brown SPR	3.69	3.37×10^{-4}	-16.16 ***
NS SPR	2.73	8.31×10^{-6}	0.00
Dundee FP	9.10	1.49×10^{-3}	-50.47 ***
Dundee GP	10.73	6.68×10^{-4}	-83.91 ***
GECO FP	3.52	5.15×10^{-5}	-2.76 *
GECO GP	2.46	3.31×10^{-5}	-10.89 ***
Provo FP	1.48	2.70×10^{-5}	-2.05
Provo GP	9.84	8.86×10^{-5}	-28.39 +

Table 3: Regression results using memory update predictors from CBR-RNN.

6 Experiment 3: Memory Updates in a Bounded-Memory LM

A potential concern about using transformer-style self-attention to model human sentence processing is that self-attention operates on lossless representations of a large number of previous tokens in the context window. This has been noted as a potential mismatch between transformer memory and human working memory (Kuribayashi et al., 2022; Clark et al., 2025; Oh and Linzen, 2026), as human working memory is known to be tightly constrained. To address this mismatch, Experiment 3 evaluates a model that replaces the self-attention mechanism in CBR-RNN with a bounded associative memory cell; we call the bounded-memory model **CBR-RNN-M**. The implementation of this the memory cell in CBR-RNN-M is inspired by the associative memory models described in Section 3.1, and resembles linear transformers and state space models that have been developed to reduce memory overhead in natural language processing applications (e.g., Katharopoulos et al., 2020; Dao and Gu, 2024; Yang et al., 2024).

Directly implementing an associative memory

[7] In other words, standard attention replaces $\mathbf{K}_{<i}$ and $\mathbf{V}_{<i}$ with $\mathbf{K}_{\leq i}$ and $\mathbf{V}_{\leq i}$ in Equation (7).

[8] The use of BPE also differs from the original CBR-RNN implementation, which uses a word-level model; shifting to a subword-based model improved perplexity on preliminary experiments and also made it possible to avoid using an <unk> token for unknown words, which can distort surprisal and memory update estimates.

[9] This experiment and Experiment 3 use GPT2 surprisal rather than OPT surprisal as a baseline predictor.

cell in CBR-RNN-M also creates a stronger alignment between the models' internal computations and the memory update measure studied in this work. Whereas associative memory updates from GPT2, OPT, and CBR-RNN are extracted as a post-hoc measure, memory updates in CBR-RNN-M are explicitly computed within the model as a step in next-word predictions.

6.1 CBR-RNN-M Language Model

The CBR-RNN-M architecture replaces the softmax-based attention in CBR-RNN (Eq. 7) with a memory access resembling Equations (2) and (3). Specifically, given the current token's query vector $\mathbf{q}_i$, the attention output is calculated as

$$\mathbf{a}_i = \mathbf{M}_{i-1}\mathbf{q}_i, \tag{13}$$

where $\mathbf{M}_{i-1}$ is a memory cell storing associations between the key and value vectors of previous tokens. Later, once $\mathbf{k}_i$ and $\mathbf{v}_i$ have been calculated (Eqs. 9 and 10), the memory cell is updated using a Widrow-Hoff learning rule:

$$\mathbf{M}_i = \mathbf{M}_{i-1} + \beta(\mathbf{v}_i - \mathbf{M}_{i-1}\mathbf{k}_i)\mathbf{k}_i^\top \tag{14}$$

The other processing steps in CBR-RNN-M follow CBR-RNN (Section 5.1). Memory update scores can be straightforwardly computed from this model using Equation (4).

A grid search over values for β parameter in Equation (14), using a subset of the Wikitext-103 data, showed that $\beta = 0.0005$ attained the best perplexity, so this value was used for subsequent LM training on the full corpus. Similarly to Experiment 2, five randomly initialized instances of CBR-RNN-M were trained. Results from alternative β values are given in Appendix B.

6.2 Procedures

The procedures for regression modeling closely followed Experiment 2.

6.3 Results

Table 4 presents the results from Experiment 3. Significant reductions in MSE were observed on all corpora except Provo GP. Similarly to the results from CBR-RNN (Table 3), all effect sizes are positive. Compared to the CBR-RNN results, larger MSE and LL improvements are seen on the two self-paced reading corpora, while improvements on the three eye-tracking corpora are in a similar

Corpus	Effect	ΔLL	ΔMSE
Brown SPR	8.25	1.30×10^{-3}	−60.35 ***
NS SPR	7.12	3.94×10^{-4}	−10.94 ***
Dundee FP	8.34	1.11×10^{-3}	−37.36 ***
Dundee GP	9.71	6.31×10^{-4}	−79.63 ***
GECO FP	2.05	2.81×10^{-5}	−1.48 *
GECO GP	1.63	1.87×10^{-5}	−6.03 ***
Provo FP	2.07	1.16×10^{-4}	−8.75 ***
Provo GP	6.66	4.35×10^{-5}	−13.46

Table 4: Regression results using memory update predictors from CBR-RNN-M.

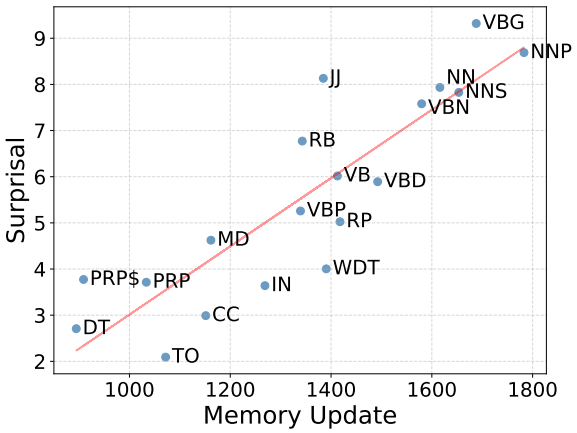


Figure 2: Comparison of average surprisal and memory update values across parts of speech in the Natural Stories corpus.

range. The generally strong reading time predictions from CBR-RNN-M suggest that replacing softmax-based attention with a bounded memory cell resembling linear transformer models (e.g., Katharopoulos et al., 2020) could be a promising direction for psycholinguistic modeling.

7 Analysis of Predictors

To illustrate the similarities and differences between the memory update predictor and surprisal, Figure 2 plots the two predictors' average values for the most frequent parts of speech (POS) in the Natural Stories corpus (Futrell et al., 2021). Surprisal and memory update measures in this plot were both averaged over the five CBR-RNN-M models, and POS labels are from the Penn Treebank tagset (Marcus et al., 1993). In general, there is a strong association between the two predictors' per-POS averages. However, the memory update predictor shows relatively high values for several classes of function words including prepositions (IN, TO), coordinating conjunctions (CC), wh-determiners (WDT), and particles (RP). To test whether this re-

flects a general trend toward larger memory updates for function words, the average difference between z-scaled memory update and surprisal predictors was computed separately for function words and content words.[10] The average difference was significantly higher for the memory update predictor than for surprisal ($p \ll 0.001$, one-tailed Welch's t-test). This is evidence that memory updates are relatively sensitive to words carrying structural information in comparison to surprisal.

The memory update predictors tested in this work typically show correlations of between 0.4 to 0.7 with surprisal predictors, suggesting that the predictors overlap but also may be sensitive to different factors. See Figure 3 in Appendix C for a full comparison of correlations between predictors.

8 Conclusion

We explore a novel connection between LMs and human sentence processing based on associative memory updates. Memory updates are measured both from standard transformers and from models that combine transformer attention with cognitively motivated constraints—including a new model architecture that implements a bounded-size associative memory. Evidence for effects on reading times is found on a range of datasets, with especially promising results from the cognitively motivated models. Focusing on associative memory as a bridge between language models and humans may be helpful for achieving further progress in the development of broad-coverage, algorithmic-level models of human sentence processing.

Limitations

The evaluations in this work are limited to corpora in English containing self-paced reading times and eye-tracking measures. Further work will be necessary to verify the effects of associative memory updates across other languages and measures of processing difficulty (e.g., neural measures from fMRI or EEG). We also restrict our evaluations to two pretrained transformer models and two cognitively motivated architectures; it is possible that different effects will be observed from models with different architectures (e.g., alternative implementations of bounded memory) or of different scales.

[10] The Penn treebank tags classified as function words were DT, IN, PRP, CC, TO, PRP$, MD, WDT, and RP; the tags classified as content words were NN, VBD, JJ, RB, NNS, VB, NNP, VBN, VBG, and VBP. Other tags were not considered due to their low frequency in Natural Stories.

References

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.

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

Christian Clark, Byung-Doh Oh, and William Schuler. 2025. Linear recency bias during training improves transformers' fit to reading times. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7735–7747, Abu Dhabi, UAE. Association for Computational Linguistics.

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.

Tri Dao and Albert Gu. 2024. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

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.

Mario Giulianelli, Sarenne Wallbridge, and Raquel Fernández. 2023. Information value: Measuring utterance predictability as distance from plausible alternatives. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 5633–5653.

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

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*, pages 159–166, Pittsburgh, PA.

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.

Donald O. Hebb. 1949. *The organization of behavior: A neuropsychological theory*. Wiley, New York.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference*

on Empirical Methods in Natural Language Processing, pages 7038–7051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marc W. Howard and Michael J. Kahana. 2002. A distributed representation of temporal context. *Journal of Mathematical Psychology*, 45:269–299.

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR.

Alan Kennedy, James Pynte, and Robin Hill. 2003. The Dundee corpus. In *Proceedings of the 12th European conference on eye movement*.

Maayan Keshev, Mandy Cartner, Aya Meltzer-Asscher, and Brian Dillon. 2025. A working memory model of sentence processing as binding morphemes to syntactic positions. *Topics in Cognitive Science*, 17(1):88–105.

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.

Tatsuki Kuribayashi, Yohei Oseki, Souhaib Ben Taieb, Kentaro Inui, and Timothy Baldwin. 2025. Large language models are human-like internally. *Transactions of the Association for Computational Linguistics*, 13:1743–1766.

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.

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.

Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330.

David Marr. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman and Company.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

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.

James Michaelov, Catherine Arnett, and Ben Bergen. 2024. Revenge of the Fallen? Recurrent Models Match Transformers at Predicting Human Language Comprehension Metrics. In *First Conference on Language Modeling*.

Beren Millidge, Mufeng Tang, Mahyar Osanlouy, Nicol S Harper, and Rafal Bogacz. 2024. Predictive coding networks for temporal prediction. *PLOS Computational Biology*, 20(4):e1011183.

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.

Byung-Doh Oh and Tal Linzen. 2026. To model human linguistic prediction, make llms less superhuman. *Trends in Cognitive Sciences*.

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. 2023. 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.

Chiebuka Ohams, Sathvik Nair, Shohini Bhattasali, and Philip Resnik. 2026. A predictive coding model for online sentence processing. *Journal of Memory and Language*, 146:104705.

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

Rajesh PN Rao and Dana H Ballard. 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87.

Nathan E. Rasmussen and William Schuler. 2018. Left-corner parsing with distributed associative memory produces surprisal and locality effects. *Cognitive Science*, 42(S4):1009–1042.

David E Rumelhart, James L McClelland, PDP Research Group, and 1 others. 1986. *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. The MIT press.

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.

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.

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 (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

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.

Carla J Shatz. 1992. The developing brain. *Scientific American*, 267(3):60–67.

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.

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.

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.

Bernard Widrow and Michael A Lehr. 1990. 30 years of adaptive neural networks: perceptron, madaline, and backpropagation. *Proceedings of the IEEE*, 78(9):1415–1442.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. *Preprint*, arXiv:2006.01912.

Ethan Gotlieb 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.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. *Preprint*, arXiv:2309.17453.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. 2024. Parallelizing linear transformers with the delta rule over sequence length. *Advances in neural information processing systems*, 37:115491–115522.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and 1 others. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

A Regression Formulas

The baseline and main regression formulas for predicting self-paced reading times are given in (15) and (16), respectively:

$$\begin{aligned}\text{RT} \sim\ & \text{WordLength} + \text{WordPosition} \\ & + \text{Surprisal}_{\text{Transformer}} + \text{Surprisal}_{\text{CBR-RNN}} \\ & + \text{Surprisal}_{\text{CBR-RNN-M}} + (1 \mid \text{Subject})\end{aligned} \tag{15}$$

$$\begin{aligned}\text{RT} \sim\ & \text{WordLength} + \text{WordPosition} \\ & + \text{Surprisal}_{\text{Transformer}} + \text{Surprisal}_{\text{CBR-RNN}} \\ & + \text{Surprisal}_{\text{CBR-RNN-M}} + \text{MemoryUpdate} \\ & + (1 \mid \text{Subject})\end{aligned} \tag{16}$$

The baseline and main regression formulas for predicting eye-tracking measures are given in (17) and (18), respectively:

$$\begin{aligned}\text{RT} \sim\ & \text{WordLength} + \text{WordPosition} \\ & + \text{Surprisal}_{\text{Transformer}} + \text{Surprisal}_{\text{CBR-RNN}} \\ & + \text{Surprisal}_{\text{CBR-RNN-M}} + \text{PrevWordFixated} \\ & + (1 \mid \text{Subject})\end{aligned} \tag{17}$$

$$\begin{aligned}\text{RT} \sim\ & \text{WordLength} + \text{WordPosition} \\ & + \text{Surprisal}_{\text{Transformer}} + \text{Surprisal}_{\text{CBR-RNN}} \\ & + \text{Surprisal}_{\text{CBR-RNN-M}} + \text{PrevWordFixated} \\ & + \text{MemoryUpdate} + (1 \mid \text{Subject})\end{aligned} \tag{18}$$

B Additional CBR-RNN-M Evaluations

Table 5 shows results from CBR-RNN-M models with varying values for the β parameter; a higher β means that a larger memory update is made at each incoming word. These evaluations used the same baseline predictors as Experiment 3. A single instance of CBR-RNN-M was trained and evaluated for each β value.

C Correlations Between Predictors

Figure 3 shows the correlations between all predictors that were included in the mixed-effects regression models (except for the predictor indicating whether the previous word was fixated, which was only used in eye-tracking corpora). Correlations were computed on the Natural Stories corpus (Futrell et al., 2021).

Corpus	Effect	ΔLL	ΔMSE
Brown SPR	8.30	1.31×10^{-3}	-61.62 ***
NS SPR	7.86	4.21×10^{-4}	-16.77 ***
Dundee FP	10.46	1.73×10^{-3}	-58.19 ***
Dundee GP	12.72	9.41×10^{-4}	-118.33 ***
GECO FP	0.98	2.76×10^{-6}	-0.18
GECO GP	1.40	9.65×10^{-6}	-3.48
Provo FP	6.09	4.32×10^{-4}	-32.59 *
Provo GP	15.52	2.33×10^{-4}	-71.82

(a) $\beta = 0.01$

Corpus	Effect	ΔLL	ΔMSE
Brown SPR	8.31	1.22×10^{-3}	-57.43 ***
NS SPR	7.75	4.31×10^{-4}	-16.97 ***
Dundee FP	8.92	1.31×10^{-3}	-43.96 ***
Dundee GP	10.31	7.50×10^{-4}	-94.99 ***
GECO FP	2.46	1.38×10^{-5}	-0.69
GECO GP	1.39	1.52×10^{-5}	-4.92 *
Provo FP	2.27	9.76×10^{-5}	-7.70 +
Provo GP	8.42	4.50×10^{-5}	-14.92

(b) $\beta = 0.001$

Corpus	Effect	ΔLL	ΔMSE
Brown SPR	7.32	9.68×10^{-4}	-45.35 ***
NS SPR	6.19	2.18×10^{-4}	-8.08 **
Dundee FP	7.76	8.67×10^{-4}	-29.13 ***
Dundee GP	8.96	5.43×10^{-4}	-68.85 ***
GECO FP	2.55	4.00×10^{-5}	-2.15
GECO GP	1.43	1.24×10^{-5}	-4.26 +
Provo FP	2.51	6.76×10^{-5}	-5.59
Provo GP	6.83	1.50×10^{-5}	-4.51

(c) $\beta = 0.0001$

Corpus	Effect	ΔLL	ΔMSE
Brown SPR	7.02	9.88×10^{-4}	-46.16 ***
NS SPR	6.81	3.95×10^{-4}	-15.10 ***
Dundee FP	7.36	8.59×10^{-4}	-28.73 ***
Dundee GP	9.09	5.39×10^{-4}	-67.93 ***
GECO FP	1.44	2.48×10^{-5}	-1.36
GECO GP	1.33	1.65×10^{-5}	-5.35 +
Provo FP	0.65	1.50×10^{-5}	-1.35
Provo GP	4.74	-4.50×10^{-5}	13.66

(d) $\beta = 0.00001$

Table 5: Regression results using memory update predictors from CBR-RNN-M with varying values for the β parameter.

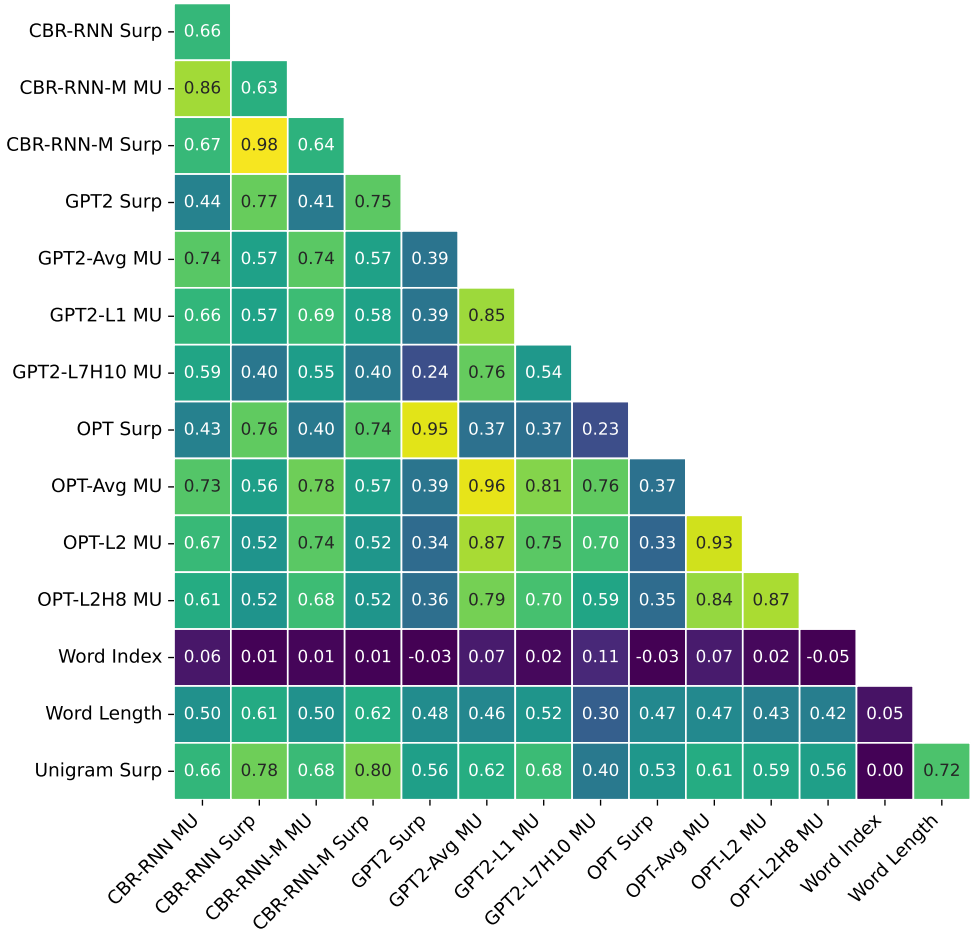


Figure 3: Correlations between predictors used in the regression models. *Surp* and *MU* refer to surprisal and memory update; names like *GPT2-L1* and *GPT2-L7H10* refer to predictors computed over specific layers (L1 = layer 1) or individual attention heads (L7H10 = layer 7, head 10) for Experiment 1.