# Memory efficiency and resource-rational encoding in sentence processing

**Weijie Xu**
University of California, Irvine
weijie.xu@uci.edu

**Brian Dillon**
University of Massachusetts Amherst
bwdillon@umass.edu

**Richard Futrell**
University of California, Irvine
rfutrell@uci.edu

## Abstract

There is a growing consensus that, in order to serve as models of human language processing, language models (LMs) need to be constrained in their use of memory for context, the analogue to human working memory (WM). Here we take a novel yet simple approach to constraining WM in language models, in a way that reflects models of human cognition where memory is treated as a limited resource and deployed strategically. In order to capture this constraint on memory encoding, we inject noise into the hidden representations of Transformer-based LMs at tunable rates. Then we train the models with a hybrid objective, such that they learn to maximize the performance of next-word prediction subject to explicit constraints on the total encoding precision. We find that explicit WM constraints improve the model's alignment with human reading times. More importantly, we find that the need to manage encoding precision reshapes the nature of the models' context representations, making them more compressed and categorical. Our results show how resource-rational models of WM allocation can be implemented in neural models simply and successfully, and point to a dissociation between WM retrieval mechanisms and the underlying memory representations in models of human sentence processing.

## 1 Introduction

Human sentence processing is incremental, continuously integrating linguistic signals into contextual representations as the sentence unfolds. This incremental process requires the support of working memory (WM), since a previous input is not perceptually accessible anymore at a future time point. However, our WM capacity is strikingly limited, both in terms of how many items can be simultaneously stored (Miller, 1956; Lewis, 1996; McElree, 2006), and in terms of how long it can be faithfully represented before being forgotten (Gibson, 1998; Futrell et al., 2020). Despite considerable empirical support for this cognitive limitation in sentence processing, theoretically, the exact representational nature of WM constraints remains an open question. This paper attempts to answer: What are WM resource constraints? How can we simulate them at the computational level? How do they shape the representational space in sentence processing?

To answer these questions, we propose to integrate WM constraints into Transformer language models (LMs), in a way that allows us to generate and evaluate hypotheses for human WM processes in a stimulus computable fashion (Frank and Goodman, 2025). Previous work has proposed numerous ad-hoc ways to constrain the memory in LMs, including reduced size of the model or training data (Warstadt et al., 2023) and locally biased attention (De Varda and Marelli, 2024; Clark et al., 2025). Here we pursue a more principled approach grounded in resource-rational cognitive theories (Lewis et al., 2014; Lieder and Griffiths, 2020), aiming to advance understanding of both psycholinguistic theories of WM constraint and the impact of our psycholinguistically inspired architectural choice on LMs' performance.

We conceptualize WM constraints focusing on the representational quality of information (McElree, 2006). We argue that greater memory resources result in lower representational uncertainty in memory encodings of linguistic inputs (Ma et al., 2014; Bates and Jacobs, 2020; Bays et al., 2024; Xu and Futrell, 2026). To capture representational uncertainty, we inject noise into Transformer's self-attention value vectors, such that the encoding precision at a key position exponentially decays as a function of its distance with the query position. Given our conceptualization, models with greater WM resources at their disposal can encode linguistic inputs with greater precision in their value vectors. We then train LMs with varying WM constraints in a resource-rational fashion: the models



*Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33603–33618

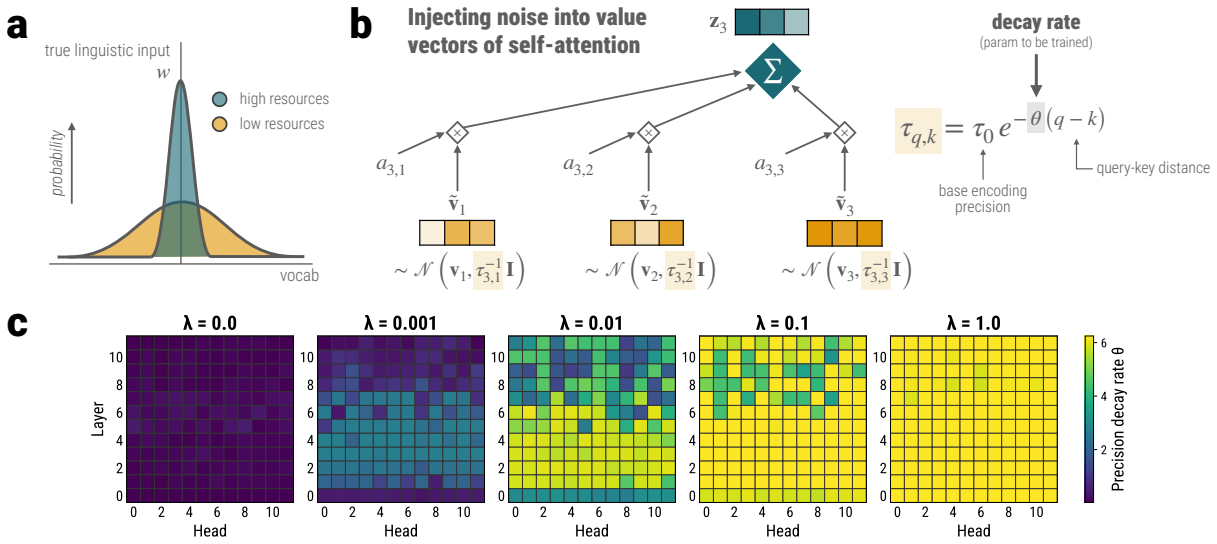


Figure 1: **a.** Greater WM resources result in lower representational noise for the encoded linguistic input. **b.** Injecting Gaussian noise into Transformers' self-attention value vectors $\mathbf{v}$, whose precision decays as a function of query-key distance controlled by $\theta$. **c.** Precision decay rate $\theta$ learned by models across attention heads per layer.

learn to maximize the performance of next-word prediction task under the constraint of limited WM resources for precise encoding (Hahn et al., 2022).

We first observe an asymmetrical decay of encoding precision: lower-layer attention heads undergo faster precision decay under the WM constraint. We then evaluate our models by using their surprisal estimates to predict human reading times (RTs). Consistent with previous studies, we find that explicit WM constraints can help the model's surprisal estimates to better predict RTs. More importantly, the model's output distributions reveal that stronger constraints lead to a more compressed representational space in next-word predictions, resulting in more categorical predictions with less fine-grained distinctions across different contextual inputs, a novel prediction to be examined in future experimental work of human sentence processing.

## 2 Background

### 2.1 WM constraints at the computational level

In order to understand WM processes (Oberauer, 2009), two questions must be addressed: (i) what kinds of information are encoded in memory; (ii) how do we make use of such information? In sentence processing, computational models and theories have focused primarily on the latter, targeting the retrieval mechanism of WM (Lewis and Vasishth, 2005; Lewis et al., 2006). The former has, however, long been the elephant in the room, until a more recent and growing body of work has started to address the representational nature of WM encoding (e.g., Smith and Vasishth, 2020; Hahn et al., 2022; Keshev et al., 2025; Xu and Futrell, 2026).

WM resources, under retrieval-based models, are often conceived as attentional resources, whose distribution determines the extent to which a previous input is retrieved and incorporated into the computation of representation at a future time point (Lewis and Vasishth, 2005; Lewis et al., 2006; Ryu and Lewis, 2021, 2025; Oh and Schuler, 2022). For most studies applying these ideas in LMs, this WM constraint has been interpreted as a locality bias in attention (De Varda and Marelli, 2024; Clark et al., 2025; Mita et al., 2025) or as limited length of context window (Kuribayashi et al., 2022), a design choice that is grounded in the empirical observation where information in the recent past is more salient and better remembered (Gibson et al., 1996; Gibson, 1998; Bartek et al., 2011).[1]

This retrieval-centered theorizing of WM constraint leaves unaddressed how such a constraint influences what we have encoded in memory in the first place. Here we pursue a different approach to conceptualize WM resources and constraints: instead of viewing them as attentional resources, we focus on the representational nature of the encoded information itself, and directly link WM resources to *representational quality* (McElree, 2006).

[1] In an interesting exception, Timkey and Linzen (2023) propose an implementation of WM constraint that is not grounded in the locality bias. Instead, the constraint consists in the limited number of computational modules (that is, the number of attention heads).

Human memory representations are noisy, as the received information always undergoes a certain degree of unavoidable perturbation during transmission in the neural system (e.g., Ma et al., 2014; Futrell et al., 2020; Hahn et al., 2022; Brady et al., 2024; Bays et al., 2024). We thus interpret representational quality as *representational uncertainty*, or the level of noise in memory encoding. As shown in Figure 1a, we argue that greater resources result in lower representational uncertainty, an assumption entertained by more and more recent studies (e.g., Ma et al., 2006; Bays et al., 2009; Ma et al., 2014; Bates and Jacobs, 2020; Bays et al., 2024; Xu and Futrell, 2026).[2]

Our conceptualization of WM resource constraint attempts to address the first question raised at the beginning of this section. That is, we aim to understand the relationship between memory constraints and the content of information encoded in memory, and we offer a computational-level analysis of this question (Marr, 1982).[3] This does not stand in opposition to retrieval-focused work, in that the emergence of retrieval bias may be shaped by the representational quality of what is encoded in the first place. We revisit this point in Section 7.

### 2.2 Efficient use of limited WM resources

What is the optimal strategy to efficiently use limited WM resources in language processing? The answer to this question requires an explicitly defined processing task (Anderson, 1990), and implies a trade-off between the task performance and the amount of available resources (Simon, 1955; Lewis et al., 2014; Gershman et al., 2015; Lieder and Griffiths, 2020). In other words, as a resource-rational language user, the goal is to find what kinds of representation for the encoding of linguistic inputs require the least amount of resources yet good enough to accomplish certain processing task.

Following Hahn et al. (2022), we take next-word prediction as one such processing task. Theoretically, next-word prediction has been argued to be a causal bottleneck for the processing of both lexical and structural features (Hale, 2001; Levy, 2008). Empirically, the great success of language modeling suggests that complex linguistic features can emerge simply from the prediction task (Mahowald et al., 2024; Futrell and Mahowald, 2025). In the resource-rational model of Hahn et al. (2022), inputs in the past are assigned a probability of erasure dependent on their lexical identity and the distance from the current time point. The model learns how to assign this erasure probability to maximize the performance of next-word prediction, subject to a resource limit on how many words are retained. However, this binary erasure in their model (either erased or not) does not capture the intuition that the representation of a past input is often subtly distorted rather than being entirely forgotten.

### 2.3 The current study

To model the link between representational uncertainty and WM resources, we inject Gaussian noise into self-attention value vectors of Transformers. This method lets us manipulate WM constraints: models with greater WM resources have a lower level of noise, allowing them to more precisely encode linguistic inputs. Compared to binary noise manipulations (Futrell et al., 2020; Hahn et al., 2022), Gaussian noise on the distributional representational space of Transformers implements more subtle perturbations on the encoded information. From an information-theoretic perspective, additive white Gaussian noise (AWGN) is a commonly used noise model for communication channels, whose capacity is determined by the signal-to-noise power ratio (SNR). The Gaussian noise we inject into Transformers can therefore be interpreted in terms of modifying the SNR under which the information in each layer is transmitted to the next. In neuroscience, AWGN channels are also widely used to model information encoding in neurons (Stone, 2018). In addition to Gaussian noise injection, we adopt a hybrid training objective: models learn to maximize the performance of next-word prediction with explicit penalty on the model's encoding precision, so that they learn to strike a balance on the trade-off between prediction and memory (Still, 2014; Hahn and Futrell, 2019).

[2] One way to view cognitive *resources* is in terms of the number of perceptual samples available: more perceptual samples are needed to form a more precise encoding distribution (Norris, 2006). Moreover, at the neural level, from the perspective of probabilistic population coding, the overall amplitude of the population activity is inversely proportional to the variance of the represented distribution (Ma et al., 2006), raising a link between the cognitive resources and the lower-level biological energy.

[3] Our use of the term "computational-level" slightly deviates from the one posited by Marr (1982), and our approach is situated at the boundary between the computational and algorithmic levels under Marr's three-level framework. Following the resource-rational approach to cognition (Section 2.2), we view our model as going beyond the computational level (i.e., idealized processing model shaped only by the cognitive function), pushing the theorizing towards the algorithmic level by considering more and more realistic cognitive architecture and resources (Lieder and Griffiths, 2020).

## 3 Modeling framework

### 3.1 Injecting noise into self-attention

As illustrated in Figure 1b, we modify the self-attention mechanism (Vaswani et al., 2017) in Transformer-based LMs such that the noisy version of the value vector of the token $\tilde{\mathbf{v}}_{qk} \in \mathbb{R}^d$ at a previous key position $k$ given the query position $q$ is sampled from a Gaussian distribution

$$\tilde{\mathbf{v}}_{qk} \sim \mathcal{N}\left(\mathbf{v}_{qk}, \tau^{-1}\mathbf{I}\right), \tag{1}$$

where $\tau$ is the encoding precision, which is the inverse of variance $1/\sigma^2$.

We assume an exponential decay on the encoding precision $\tau$, such that the precision of a key token $k$ is a function of its distance with the query position $q$ and is controlled by the decay rate $\theta$

$$\tau_{qk} = \tau_0\, e^{-\theta\,(q-k)}, \tag{2}$$

where $\tau_0$ is the highest achievable precision in memory encoding.[4] The maximal precision $\tau_0$ is a hyperparameter manually specified. The specification in Eq. 2 implies that linguistic inputs can never be perfectly encoded without any noise, and $\tau_0$ is the base level of irreducible noise, which in our model is specified to be encoding precision of the token at query position itself (when $q - k = 0$). This imperfect encoding is not necessarily a flaw in our modeling strategy in that representations are always noisy during neural transmission even at the perceptual level as mentioned above.[5]

The output $\mathbf{z}$ of each attention head at query position $q$ is thus a weighted sum over noisy value vectors of all previous key positions $k$

$$\mathbf{z}_q = \sum_k a_{qk}\tilde{\mathbf{v}}_{qk}, \tag{3}$$

whose weights $\mathbf{a}$ are given by

$$\mathbf{a}_q = \mathrm{softmax}\left(\frac{\mathbf{K}^\top \mathbf{q}}{\sqrt{d}}\right). \tag{4}$$

We then apply a reparameterization for this weighted sum of Gaussian distributions for efficient stochastic optimization. The attention head output $\mathbf{z}$ is decomposed into two components, one deterministic and one capturing the Gaussian noise[6]:

$$\mathbf{z}_q = \mathbf{z}_q^{\mathrm{det}} + \tilde{\mathbf{z}}_q. \tag{5}$$

The deterministic part $\mathbf{z}^{\mathrm{det}}$ is computed as the standard self-attention as if there were no noise. The noisy part is a unit Gaussian distribution $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ scaled by the standard deviation $\sigma$ of $\tilde{\mathbf{z}}$:

$$\tilde{\mathbf{z}}_q = \sigma_q \boldsymbol{\epsilon}, \tag{6}$$

Assuming that the noise components of all previous tokens are independent of each other,[7] the variance of $\tilde{\mathbf{z}}_q$ after rescaling is

$$\sigma_q^2 = \sum_k \left(a_{qk}\sigma_{qk}\right)^2 = \sum_k a_{qk}^2 \tau_{qk}^{-1}. \tag{7}$$

We inject noise into self-attention instead of the output embedding of Transformer blocks for two reasons. First, theoretically, the mechanism of self-attention better resembles WM processes, where input representations at different time points communicate with each other. Second, for our purpose, the noisy representation of a token depends on its position relative to the query position of memory retrieval. This design is naturally realized by self-attention with parallelized computation.

### 3.2 Manipulating WM capacity constraint

In order to manipulate WM constraints, we train the models with a hybrid loss function:

$$\mathcal{L} = \underbrace{-\sum_{\mathbf{y}\in\mathbf{S}}\sum_{t=1}^{T} \log p\left(y_t \mid \mathbf{y}_{<t}\right)}_{\text{cross-entropy loss}} + \lambda \underbrace{\sum_{hltk} \tau_{hltk}\, z^{-1}}_{\text{memory capacity}}, \tag{8}$$

where $z$ is a normalization constant given by

$$z = L\,H\,T\,\frac{(T-1)}{2}. \tag{9}$$

This loss function consists of two components. The first is the cross-entropy loss, which measures the model's performance in the next-word prediction

[4] A small value of 0.0001 is added to every $\tau_{qk}$ in implementation to avoid numerical underflow.

[5] As mentioned earlier, what actually matters for WM constraint is the signal-to-noise ratio (SNR), which is currently manipulated by $\tau$ assuming constant power of the signal. Future work can consider manipulating SNR more directly.

[6] Under our reparameterization, the attention scores $\mathbf{a}$ and the deterministic component in the attention-head output $\mathbf{z}^{\mathrm{det}}$ are independent of noise in the forward pass. The influence of noise on $\mathbf{a}$ and $\mathbf{z}^{\mathrm{det}}$ consists in the backward pass, where the gradients flow into $\mathbf{a}$ and $\mathbf{z}^{\mathrm{det}}$ through the noisy version of $\mathbf{z}$.

[7] This independence assumption is a simplification, and is not necessarily the case in reality, especially given the similarity-based memory interference where the representation of one memory object can be distorted by another one.

task on training examples $\mathbf{S}$. The second component represents the average encoding precision $\tau$ of past tokens over all attention heads $h$ and layers $l$, which corresponds to the amount of WM resources. WM constraint is represented by $\lambda$: higher $\lambda$ means stronger constraint, applying stronger penalty to models with higher encoding precision.[8]

## 4 Experiment 1: Memory capacity and encoding precision decay

### 4.1 Language Models

We train on the `Wikitext-103` dataset (Merity et al., 2016) a sequence of five noisy LMs adapted from the architecture of GPT-2 small (Radford et al., 2019), corresponding to five WM capacity constraints $\lambda = [0, 0.001, 0.01, 0.1, 1]$. The model architecture is the same as GPT-2 small, except that the block size is 512 (instead of 1024 in standard GPT-2 small) and that the self-attention heads include noisy encoding as outlined in Section 3.1.[9] Each attention head at each layer gets its own precision decay parameter $\theta$ to be learned from training data. The highest achievable encoding precision ($\tau_0$ in Eq. 2) is set to 10000 (equivalent to $\sigma = 0.01$). See Appendix A for model training details.

### 4.2 Results and discussion

Figure 1c shows the precision decay rate $\theta$ learned by each attention head per layer. With no constraint on WM capacity ($\lambda = 0$), all heads learn a decay rate close to 0. Intuitively, this is consistent with the memory-surprisal trade-off proposed in Hahn et al. (2021), in the sense that it is useful to use all the information from the preceding context to maximize the accuracy of next-word prediction.

Importantly, as WM constraint gets stronger, there is a clear pattern of asymmetrical decay. With lower-to-mid memory constraint ($0 < \lambda \leq 0.01$), precision decay first begins among lower-layer attention heads, which often take charge of processing morphosyntactic features where information dependency is more local (Vig and Belinkov, 2019). Then, when WM constraint further increases ($\lambda > 0.01$), the decay starts to diffuse to other heads at higher layers, which is often considered to process longer discourse with deeply contextualized representations (Kuribayashi et al., 2025). This prioritization of higher layers mirrors humans' memory strategy, in the sense that humans tend to remember the semantic or discourse gist rather than the exact surface form (Sachs, 1974).

## 5 Experiment 2: Psychometric predictive power on human reading times

Experiment 2 examines the alignment between our models and human reading behaviors. We assume that models better aligning with human reading behaviors should have higher psychometric predictive power for its surprisal estimates on reading times (RTs). Following previous work, the predictive power is quantified as the difference of log-likelihood $\Delta$LL between regression models fit to RTs with and without surprisal terms (Wilcox et al., 2023; Xu et al., 2023; Frank and Bod, 2011). Higher $\Delta$LL indicates better alignment.

### 5.1 Experiment setup

We evaluate the model predictive power on three English reading-time corpora:

(1) Provo (Luke and Christianson, 2018): An eye-tracking corpus with 55 short passages collected from 84 participants.

(2) SPR Natural Stories Corpus (SPRNSC) (Futrell et al., 2021): A self-paced reading corpus with 10 long stories (approx 1000 words each) collected from 181 participants.[10]

(3) A-Maze Natural Stories (MazeNSC) (Boyce and Levy, 2023): Same text material as SPRNSC but collected from 100 participants using A-Maze paradigm (Boyce et al., 2020).

**RT measures.** For Provo, we examine model predictive power on two eye-tracking RT measures separately, namely the first fixation[11] and the total time[12]. We exclude eye-tracking RTs that are faster than 100ms or slower than 2000ms. For SPRNSC, due to the strong spillover effect in the self-paced reading paradigm, we use RTs of both the critical region and the spillover region. For MazeNSC, we focus on the critical region, and exclude RT observations where participants chose a wrong word in their first try. We also exclude RTs that are shorter than 100ms or longer than 3000ms in MazeNSC.[13]

[8] Models and code are available at: https://github.com/weijiexu-charlie/resource-rational-encoding

[9] Our implementation is based on nanoGPT (Karpathy, 2023), whose backbone reflects the architecture of GPT-2.

[10] The NSC corpus intentionally includes rare constructions.

[11] The duration of the first fixation on the word of interest.

[12] The sum of durations of all fixations on a word.

[13] We adopt a wider RT exclusion window for MazeNSC, since the A-Maze paradigm often induces longer RTs com-

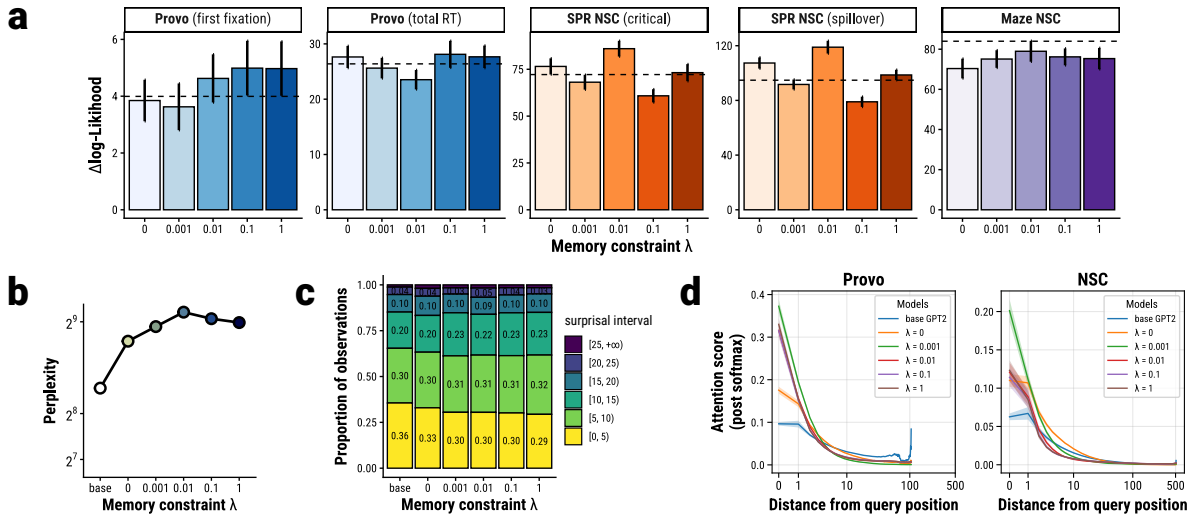


Figure 2: **a.** Psychometric predictive power of word surprisal estimates on RTs (measured as ΔLL) across WM constraints; Bar height is the mean ΔLL over 10 held-out folds with standard errors; Dashed lines are the result of the baseline GPT-2. **b.** Model perplexity on RT corpora. **c.** Distribution of surprisal estimates over six bins for RT corpora. **d.** Attention score over key tokens as a function of query-key distance, aggregated over all query tokens, attention heads and layers; passages in Provo corpus have much shorter context than NSC corpus.

**Surprisal estimates.** Following Goldstein et al. (2022), we use a sliding window paradigm to generate the surprisal estimate of each word in RT corpora, with the maximal context window (511 preceding tokens). In addition to the five noisy LMs trained in Experiment 1, we also train a baseline GPT-2 model on the same dataset without noise injection. All other hyperparameters and training procedures are identical to the noisy models.[14] We examine the predictive power of all six models.[15]

**Regression models.** In order to get ΔLL, for each model, we fit two linear mixed-effects regression models (Baayen et al., 2008) to predict RTs:

$M_0$: Null model including word position in the sentence, word length, and log-transformed word frequency from Speer (2022)

$M_1$: $M_0$ plus terms of word surprisal

To account for the spillover effect typically observed in reading experiments, we also include the word length, word frequency, and the surprisal of two previous words as control predictors. The predictive power is thus measured as the difference of log-likelihood ΔLL between $M_0$ and $M_1$. Regression models are fit using 10-fold cross validation, with ΔLL calculated from the held-out set.[16]

## 5.2 Results and discussion

**Better alignment with human RTs with WM constraint.** As shown in Figure 2a, explicit WM constraint in principle can improve the model-generated surprisals' psychometric predictive power on human RTs. This is especially true for the SPR NSC corpus, where an enhanced predictive power is observed with moderate WM constraint ($\lambda = 0.01$) compared to the baseline GPT-2 without noise injection. We also observe a mild increasing trend of predictive power for the first-fixation durations of the Provo corpus.[17] Moreover, as indicated by Figure 2b, models with explicit WM constraint yields higher perplexity on RT corpora, echoing recent findings where models that are exceedingly accurate in next-word predictions tend to show worse alignment with human reading behaviors (Oh and Schuler, 2023; Oh and Linzen, 2025).[18] It is also worth noting that we

pared to other reading paradigms.

[14]Note that our noisy model with $\lambda = 0$ is not the same as the baseline GPT-2. For our noisy model, even when $\lambda = 0$, there is still a base noise $\tau_0$, and the model still learns a small but non-zero precision decay $\theta$.

[15]For noisy models, due to their internal stochastic process, we generate 100 surprisal estimates for each word and take the average. We also exclude tokens that are punctuations.

[16]We only include random intercept for all regression models, which is the maximal random structure that allows all models to converge without singular fit.

[17]Note that the length of each passage in the Provo corpus is much shorter than in the NSC corpus, which may potentially explain the divergent pattern of predictive power between these two corpora.

[18]However, the enhanced predictive power is not observed in eye-tracking total time, nor in Maze RT. For total time, the lack of a clear pattern is possibly because total time mixes

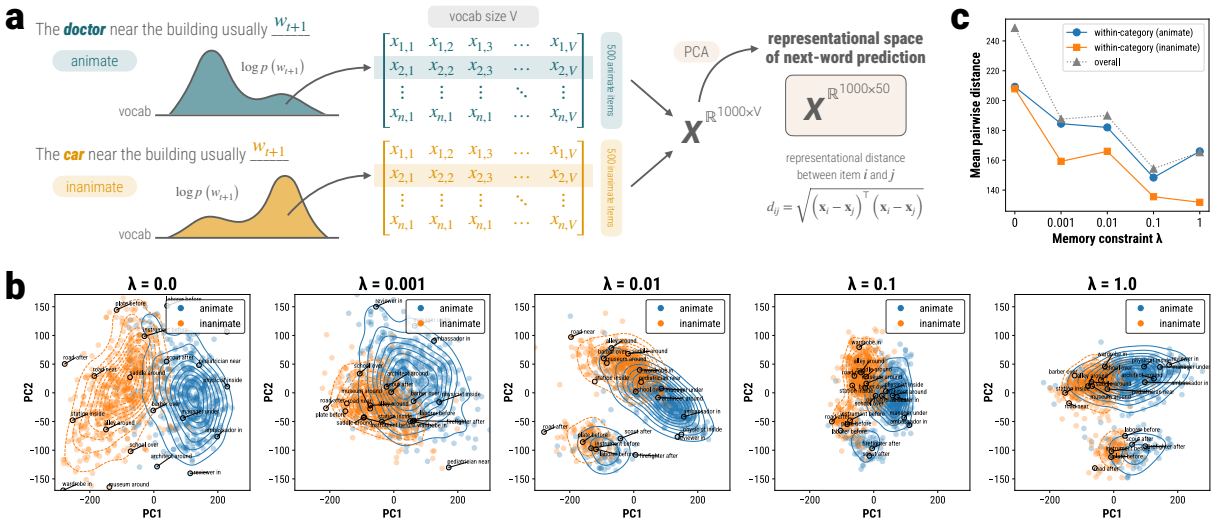


Figure 3: **a.** Generating the representational space of predictions from next-word log probabilities over the vocab; dimensionality reduced using PCA (50 PCs). **b.** Representational space across WM constraints, visualized from the first 2 PCs; each point represents an item of noun phrase (labels include `Noun1` and `Prep`). **c.** Mean Euclidean distance (based on all 50 PCs) with standard error, over all pairs of items and within each animacy category.

do not expect the RT alignment to necessarily increase monotonically as a function of WM constraint, since realistically human WM must not be infinitely constrained, a point we revisit in the Limitations section.

**Less extreme surprisal estimates with memory constraint.** As shown in Figure 2c, the model generates fewer extreme surprisal estimates at the lower end with stronger memory constraint $\lambda$. Instead, more and more estimates with higher $\lambda$ are at the moderate level $[5, 15)$. We speculate that the memory constraint may have encouraged the model to adopt a processing strategy that focuses more on making next-word predictions based on general semantic categories rather than on fine-grained distinctions, a hypothesis that we explore in Section 6.

**Emergent but dissociable locality bias in attention.** Figure 2d shows attention scores on key tokens, aggregated over all heads and layers. Compared to baseline GPT-2, all noisy models have more locally biased attention, with higher weights on local tokens. However, this locality bias does not monotonically increase in WM constraint $\lambda$. Although moderate constraint ($\lambda = 0.001$) indeed encourages the model to assign higher attention scores on local tokens, this effect is weakened for mid-to-higher constraint ($\lambda \geq 0.01$). We speculate that the poorer representational quality due to faster precision decay may have encouraged the model to more uniformly distribute its attention weights as it becomes less certain about which token to attend to in the past context. This pattern potentially points to a dissociation between WM encoding and retrieval, suggesting that stronger WM resource constraint does not necessarily lead to more locally biased retrieval in all situations.

## 6 Experiment 3: Representational space of next-word predictions

Previous studies have argued that LMs tend to make sharper predictions than humans with more fine-grained distinctions across contextual inputs, possibly due to the models' powerful superhuman memory (Oh and Linzen, 2025). This raises the expectation that models with more human-like constraints should make predictions that are less fine-grained, reflecting broader semantic categories. In Experiment 3, we aim to examine this claim, looking at how WM constraint shapes the model's representational space when making next-word predictions. We hypothesize that stronger WM constraint may result in: (i) more compressed representational space, such that different linguistic inputs are encoded more similarly, with more similar linguistic predictions across contexts; (ii) stronger categorical bias, such that different categories of certain linguistic feature (e.g., animate vs. inanimate) are

early and late (re-)reading of a word, potentially indexing multiple different underlying processes (Clifton Jr et al., 2007; Schotter and Dillon, 2025). For A-Maze, this experimental paradigm is typically memory demanding, which potentially explains the reduced predictive power in Maze RT.

represented more distinctly.

### 6.1 Experiment setup

As in Figure 3a, we first construct a set of 1000 items in English in the form of `Det Noun1 Prep Det Noun2 Adv`. The animacy of `Noun1` is manipulated such that 500 items are animate and the other 500 are inanimate.[19] We manipulate the animacy feature for two reasons. First, it has been shown to actively serve as a processing cue in language comprehension (Trueswell et al., 1994). Second, animacy is grammatically marked on pronominal elements in English. Therefore, we consider the animacy feature a potential candidate for the categorical bias to emerge with memory constraint.

Then, using the models trained in Experiment 1, the initial representational space of next-word prediction is constructed from the log probabilities over possible continuations $w_{t+1}$ generated at the `Adv`.[20] In the end, we reduce the dimensionality on this initial representational space using Principle Component Analysis (PCA). The ultimate representational space of next-word predictions is taken from the first 50 PCs.[21]

In order to evaluate this representational space, for each pair of items $i$ and $j$, we calculate their Euclidean distance $d_{ij}$ given by

$$d_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top (\mathbf{x}_i - \mathbf{x}_j)}. \qquad (10)$$

Higher item-to-item distance thus indicates that their next-word predictions are less similar.

### 6.2 Results and discussion

**More compressed representational space with higher WM constraint.** Figure 3b shows the representational space of next-word predictions visualized from the first 2 PCs, and Figure 3c summarizes the item-to-item Euclidean distance based on all 50 PCs.[22] Without explicit WM constraint ($\lambda = 0$), the representational space is more spread out, with higher item-to-item distance, indicating that next-word predictions are relatively distinct from each other across different items. As WM constraint gets stronger, the representational space becomes increasingly concentrated with lower item-to-item distance. This reduced representational distance with higher $\lambda$ suggests that the model's predictions become more similar to each other across items.

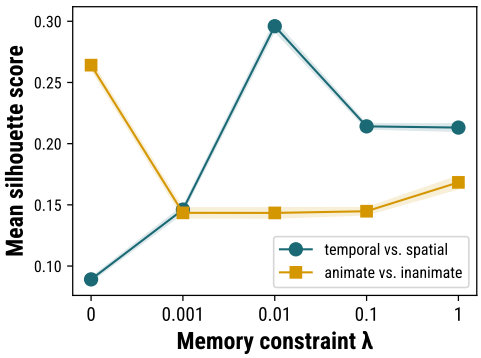


Figure 4: Mean silhouette scores with standard error for the subject-noun animacy (animate vs. inanimate) and the preposition (temporal vs. spatial) categories.

**Emergence of categorical encoding.** Although a categorical bias is not observed for the animacy feature on the subject noun of our stimuli, interestingly, there are indeed separate clusters of next-word predictions that start to emerge with stronger WM constraint $\lambda$, as indicated in Figure 3b ($\lambda \geq 0.01$). Visual inspection reveals that models start to tease apart noun phrases with prepositions *before* and *after* from others.[23] Compared to other prepositions in our stimuli, *before* and *after* tend to function as temporal prepositions, which express temporal relations semantically and more naturally take clausal complements syntactically (Dubinsky and Williams, 1995). This emergent categorical encoding is corroborated by the result of silhouette scores in Figure 4, which is a metric that quantifies how well the two categories can be separated from each other in the representational space (see Appendix C). This result indicates that at least some linguistic features (e.g., prepositions in our case) are encoded more categorically with stronger WM constraint, shifting away from fine-grained semantic distinctions towards higher-level generalization (Alvarez, 2011; Bates and Jacobs, 2020; Brady and Tenenbaum, 2013; Zaslavsky et al., 2018). Future work may consider a more thorough investigation on what kinds of linguistic features tend to be categorically encoded under WM constraints.

[19] `Noun2` and `Adv` are kept identical across all items so these tokens more local to the prediction site do not mask the effect of our manipulation on earlier ones.

[20] Similar to Experiment 2, here for each item we took the average of 100 samples of log probability distributions.

[21] It is worth noting that this is not the *internal* representational space encoded by the model. Instead, what we look into here is the representational space *implied* by the language modeling head output distributions for $w_{t+1}$.

[22] Small standard error in Fig. 3c due to the large number of item pairs.

[23] See Appendix D for the PCA visualization with 200 randomly sampled noun-phrase labels.

## 7 General discussion

In this paper, we introduce a conceptualization of WM constraint at the computational level. Focusing on the task nature of information processing, our model characterizes what kinds of representations would be encoded under WM constraint in order to accomplish the processing task in a good enough way. In information-theoretic terms, we view memory as a communicative channel that transmits information from the past to the future, and its constraint as the information bottleneck during this transmission (Still, 2014; Hahn and Futrell, 2019; Bates and Jacobs, 2020). Our model, therefore, is agnostic to how these representational properties and memory efficiency are realized at the algorithmic level.

As noted above, one potential algorithmic-level implementation of WM efficiency is locally biased retrieval (De Varda and Marelli, 2024; Clark et al., 2025). Intuitively, this retrieval bias may naturally arise from locally biased encoding precision, in that the model may prioritize the attention to more precisely encoded input. In this sense, our focus on representational quality for WM constraint does not argue against accounts focusing on retrieval bias. Instead, in our view, the emergence of locally biased retrieval may be driven by locally biased encoding precision, as what is retrieved ultimately depends on what is encoded in the first place. However, our result also reveals a dissociation: while moderate WM constraints indeed enhance the locality bias in attention weights, this effect diminishes with more stringent constraint. This dissociation suggests that the locality bias in retrieval mechanism alone is not sufficient to fully characterize WM processes. In order to develop an explanatory account for WM efficiency, it is important to target the representational properties more directly, identifying what kinds of information are encoded and why they are encoded for certain processing tasks.

Moreover, recent advances in mechanistic interpretability (Arora et al., 2024), psycholinguistically inspired probing (Hu et al., 2020), as well as our own Experiment 3, all serve as potential methods to examine how the interpretable linguistic features are encoded in our models. We believe that our computational-level manipulation of WM constraints, powered by these methods, provides a promising pathway for analyses on how different interpretable linguistic features are encoded and prioritized for more efficient processing. We believe such an analysis can provide more detailed insight into human memory efficiency, establishing closer connection to traditional linguistic and cognitive theories.

Finally, our focus on the representational quality highlights a connection to model quantization (e.g., Dettmers et al., 2022; Frantar et al., 2022; Dettmers et al., 2023; Lin et al., 2024), which can be viewed as another algorithmic-level realization of WM constraint. As a technique developed to reduce GPU memory demand, quantization approximates continuous values by mapping them to a small set of discrete levels, preserving as much task-relevant information as possible despite certain degree of degradation of precision. From this perspective, stronger WM constraint corresponds to more aggressive quantization: more fine-grained distinctions in the original continuous space are lost, making it increasingly difficult to reconstruct detailed linguistic representations. This connection to quantization thus points to a promising route for translating WM constraint, a classic notion in cognitive psychology, to its implementation in the engineering of human-like artificial intelligent systems.

## 8 Conclusion

In this paper, we manipulate WM constraint at the computational level in a resource-rational fashion, such that models with greater WM resources encode linguistic inputs more precisely. We show that explicit WM constraints improve the model's alignment with human RTs, a pattern that can be possibly explained by more compressed and categorical representational space in next-word predictions. For psycholinguistics, our work provides a new way to use LLMs as a tool to understand WM processes in human sentence processing, highlighting a dissociation between WM retrieval and the underlying representation of the encoded information. For the engineering of human-like artificial intelligent systems, our conceptualization of WM constraint highlights a connection between this classic notion in psychological theories and algorithmic-level techniques such as quantization.

## Limitations

Although our theorizing of WM constraint focuses on the representational nature of memory encoding at the computational level, our model is adapted from the architecture of Transformer-based LMs

(GPT-2 small specifically), and therefore has committed to the algorithmic-level mechanisms implemented in Transformers. For example, the internal representations of Transformers are computed in a parallel manner, a feature that makes model training more efficient. However, this parallel computation is not truly incremental. In human sentence processing, rather than taking the entire sequence as a whole, linguistic inputs are processed word by word and the memory state is updated incrementally at each time point, a process that resembles the information flow in Recurrent Neural Networks (RNNs) (Elman, 1990; Hochreiter and Schmidhuber, 1997). Although traditional RNNs are not efficient in terms of model training due to its sequential nature, this issue has been addressed by recent advances such as State Space Models (SSMs) (Dao and Gu, 2024), making it possible to efficiently train the model while maintaining its sequential nature. Future work may consider using SSMs to model incremental human sentence processing in a more realistic way.

Our model also assumes that the ordinal position relative to the retrieval site is a primitive dimension of attention and decay, an assumption that does not fully align with human sentence processing. Linguistic information is hierarchically structured, where individual units are grouped into constituents that in turn form larger constituents (Chomsky, 1957). As a result, units that are proximate in the linear sequence order can actually take very different hierarchical levels in mental representation. A substantial body of work has established that such hierarchical organization plays a crucial role in language processing (e.g., Caucheteux et al., 2023; Regev et al., 2024; Zhao et al., 2025; Gwilliams et al., 2025). However, in our model, the decay of encoding precision is specified purely as a function of the linear distance between tokens, and therefore is insensitive to the hierarchical structure of linguistic representations.

As mentioned in Section 5.2, the predictive power on human RTs does not necessarily increase monotonically as a function of the model's WM constraint $\lambda$, since the WM capacity in humans should still be powerful enough to support sentence processing. Therefore, it is more likely for $\lambda$ to exhibit the Goldilocks effect, where the predictive power on human RTs peaks at a moderate level of $\lambda$ (e.g., $\lambda = 0.01$ for the SPR NSC corpus). However, there is a lack of linking hypothesis between $\lambda$ and the actual level of human WM constraint, making it difficult to predict which specific value of $\lambda$ should best align with human behaviors.

In the end, the current work is based on GPT-2 small trained on an English-biased dataset (i.e., `wikitext-103`). In order to argue for a domain-general implementation of WM constraints, future work may consider crosslinguistic examination on our models, with more thorough investigation across different model sizes and architectures.

## Acknowledgments

This work is supported by NSF #1947307 to RF, NSF-IIS #2504954 to BD, as well as a Samuel F. Conti Faculty Fellowship from the University of Massachusetts, Amherst. We are also grateful to Jonathan Webster for his excellent research assistance, as well as the reviewers and audience at the ACL Rolling Review and the 2026 HSP Conference on Human Sentence Processing for helpful discussion and feedback.

## References

George A Alvarez. 2011. Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, 15(3):122–131.

John Robert Anderson. 1990. *The Adaptive Character of Thought*. Psychology Press.

Aryaman Arora, Dan Jurafsky, and Christopher Potts. 2024. CausalGym: Benchmarking causal interpretability methods on linguistic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14638–14663.

R Harald Baayen, Douglas J Davidson, and Douglas M Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

Brian Bartek, Richard L Lewis, Shravan Vasishth, and Mason R Smith. 2011. In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(5):1178.

Christopher J Bates and Robert A Jacobs. 2020. Efficient data compression in perception and perceptual memory. *Psychological Review*, 127(5):891.

Paul M Bays, Raquel FG Catalao, and Masud Husain. 2009. The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, 9(10):7–7.

Paul M Bays, Sebastian Schneegans, Wei Ji Ma, and Timothy F Brady. 2024. Representation and computation in visual working memory. *Nature Human Behaviour*, pages 1–19.

Veronica Boyce, Richard Futrell, and Roger P Levy. 2020. Maze made easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, 111:104082.

Veronica Boyce and Roger Levy. 2023. A-maze of Natural Stories: Comprehension and surprisal in the Maze task. *Glossa Psycholinguistics*, 2(1).

Timothy F Brady, Maria M Robinson, and Jamal R Williams. 2024. Noisy and hierarchical visual memory across timescales. *Nature Reviews Psychology*, pages 1–17.

Timothy F Brady and Joshua B Tenenbaum. 2013. A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review*, 120(1):85.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. 2023. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, 7(3):430–441.

Noam Chomsky. 1957. *Syntactic structures*. Mouton de Gruyter.

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.

Charles Clifton Jr, Adrian Staub, and Keith Rayner. 2007. Eye movements in reading words and sentences. *Eye movements*, pages 341–371.

Tri Dao and Albert Gu. 2024. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning*, pages 10041–10071.

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

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. LLM.int8(): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35:30318–30332.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115.

Stanley Dubinsky and Kemp Williams. 1995. Recategorization of prepositions as complementizers: The case of temporal prepositions in english. *Linguistic Inquiry*, 26(1):125–137.

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

Michael C Frank and Noah D Goodman. 2025. Cognitive modeling using artificial intelligence. *Annual Review of Psychology*, 77.

Stefan L Frank and Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22(6):829–834.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. GPTQ: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Richard Futrell, Edward Gibson, and Roger P Levy. 2020. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3):e12814.

Richard Futrell, Edward Gibson, Harry J Tily, Idan Blank, Anastasia Vishnevetsky, Steven T 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.

Richard Futrell and Kyle Mahowald. 2025. How linguistics learned to stop worrying and love the language models. *arXiv preprint arXiv:2501.17047*.

Samuel J Gershman, Eric J Horvitz, and Joshua B Tenenbaum. 2015. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245):273–278.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Edward Gibson, Neal Pearlmutter, Enriqueta Canseco-Gonzalez, and Gregory Hickok. 1996. Recency preference in the human sentence processing mechanism. *Cognition*, 59(1):23–59.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, and 1 others. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380.

Laura Gwilliams, Alec Marantz, David Poeppel, and Jean-Rémi King. 2025. Hierarchical dynamic coding coordinates speech comprehension in the human brain. *Proceedings of the National Academy of Sciences*, 122(42):e2422097122.

Michael Hahn, Judith Degen, and Richard Futrell. 2021. Modeling word and morpheme order in natural language as an efficient trade-off of memory and surprisal. *Psychological Review*, 128(4):726.

Michael Hahn and Richard Futrell. 2019. Estimating predictive rate–distortion curves via neural variational inference. *Entropy*, 21(7):640.

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

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

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1725–1744.

Andrej Karpathy. 2023. nanoGPT. https://github.com/karpathy/nanoGPT.

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. 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, Andrew Howes, and Satinder Singh. 2014. Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science*, 6(2):279–311.

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

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.

Falk Lieder and Thomas L Griffiths. 2020. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43:e1.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems*, volume 6, pages 87–100.

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.

Wei Ji Ma, Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. 2006. Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–1438.

Wei Ji Ma, Masud Husain, and Paul M Bays. 2014. Changing concepts of working memory. *Nature Neuroscience*, 17(3):347–356.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

David Marr. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., New York, NY, USA.

Brian McElree. 2006. Accessing recent events. *Psychology of Learning and Motivation*, 46:155–200.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81.

Masato Mita, Ryo Yoshida, and Yohei Oseki. 2025. Developmentally-plausible working memory shapes a critical period for language acquisition. *arXiv preprint arXiv:2502.04795*.

Dennis Norris. 2006. The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, 113(2):327.

Klaus Oberauer. 2009. Design for a working memory. *Psychology of Learning and Motivation*, 51:45–100.

Byung-Doh Oh and Tal Linzen. 2025. To model human linguistic prediction, make LLMs less superhuman. *arXiv preprint arXiv:2510.05141*.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Tamar I Regev, Colton Casto, Eghbal A Hosseini, Markus Adamek, Anthony L Ritaccio, Jon T Willie, Peter Brunner, and Evelina Fedorenko. 2024. Neural populations in the language network differ in the size of their temporal receptive windows. *Nature Human Behaviour*, 8(10):1924–1942.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

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.

Jacqueline S Sachs. 1974. Memory in reading and listening to discourse. *Memory & Cognition*, 2(1):95–100.

Elizabeth R Schotter and Brian Dillon. 2025. A beginner's guide to eye tracking for psycholinguistic studies of reading. *Behavior Research Methods*, 57(2):68.

Herbert A Simon. 1955. A behavioral model of rational choice. *The Quarterly Journal of Economics*, pages 99–118.

Garrett Smith and Shravan Vasishth. 2020. A principled approach to feature selection in models of sentence processing. *Cognitive Science*, 44(12):e12918.

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

Susanne Still. 2014. Information bottleneck approach to predictive inference. *Entropy*, 16(2):968–989.

James V Stone. 2018. *Principles of neural information theory: Computational neuroscience and metabolic efficiency*. Sebtel Press.

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.

John C Trueswell, Michael K Tanenhaus, and Susan M Garnsey. 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, 33(3):285–318.

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

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and 1 others. 2023. Findings of the babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*.

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.

Weijie Xu and Richard Futrell. 2026. Strategic resource allocation in memory encoding: An efficiency principle shaping language processing. *Journal of Memory and Language*, 146:104706.

Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. 2018. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31):7937–7942.

Junyuan Zhao, Ruimin Gao, and Jonathan R Brennan. 2025. Decoding the neural dynamics of headed syntactic structure building. *Journal of Neuroscience*, 45(17).

## A Model training

**Training procedure.** Models are trained on `Wikitext-103` dataset (Merity et al., 2016) with 16000 training steps. The training procedure

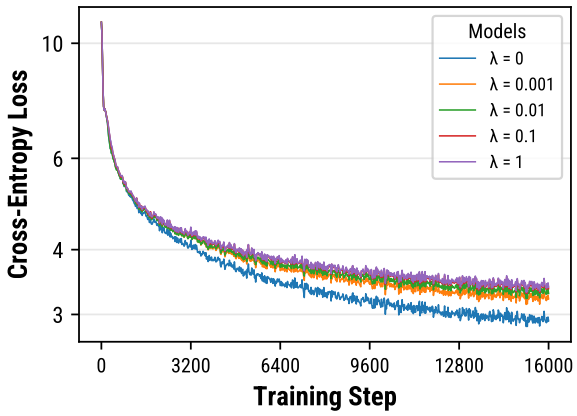


Figure 5: Learning trajectories of model training

largely follows nanoGPT (Karpathy, 2023). Each batch contains 256 examples with sequences of 512 tokens randomly sampled from training data. Models are trained using AdamW optimizer, with a learning rate of 0.0006. The learning rate was linearly warmed up over the first 3% of the maximal training steps (i.e., 500 steps), and was cosine annealed to a minimum of 0.00006. The decay rates $\theta$ are initialized randomly from Gaussian distribution with $\mu = 1$ and $\sigma^2 = 0.001$.

**Learning trajectories.** Figure 5 shows the learning trajectories of model training for the cross-entropy component. With stronger memory constraint $\lambda$, cross-entropy loss decreases more slowly and ends up with higher final cross-entropy loss.

## B Per-layer attention scores for RT corpora in Experiment 2

Figure 6 and 7 show the attention scores on key tokens as a function of key-query distance for NSC and Provo corpus respectively.

## C Silhouette scores in Experiment 3

For each item, we calculate its silhouette score (Rousseeuw, 1987) with respect to the categories of interest. Taking the animacy feature as an example (which only has two categories: *animate* vs. *inanimate*), if item $i$ is from the *animate* category, its silhouette score $s_i$ is given by

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}, \tag{11}$$

where $a_i$ is the mean distance with all other items from the same *animate* category (i.e., within-category distance), and $b_i$ is the mean distance with all items from the *inanimate* category (i.e., cross-category distance).

Then, the separatability between the two categories of a linguistic feature is taken as the mean silhouette score over all items. Higher mean silhouette score thus indicates better separatability.

## D PCA visualization with 200 NP labels in Experiment 3

Figure 8 shows the representational space in Experiment 3 visualized from the first two PCs with 200 randomly sampled noun-phrase labels (including `Noun1` and `Prep`). When $\lambda \geq 0.01$, there emerges a separate cluster for items with temporal prepositions *before* and *after*.

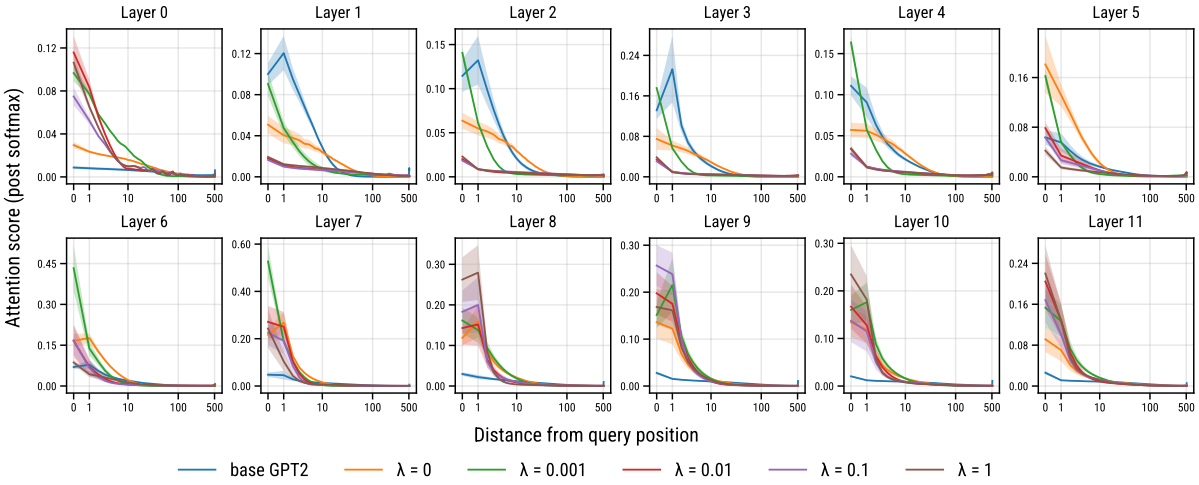


Figure 6: Aggregated attention score per layer on NSC corpus

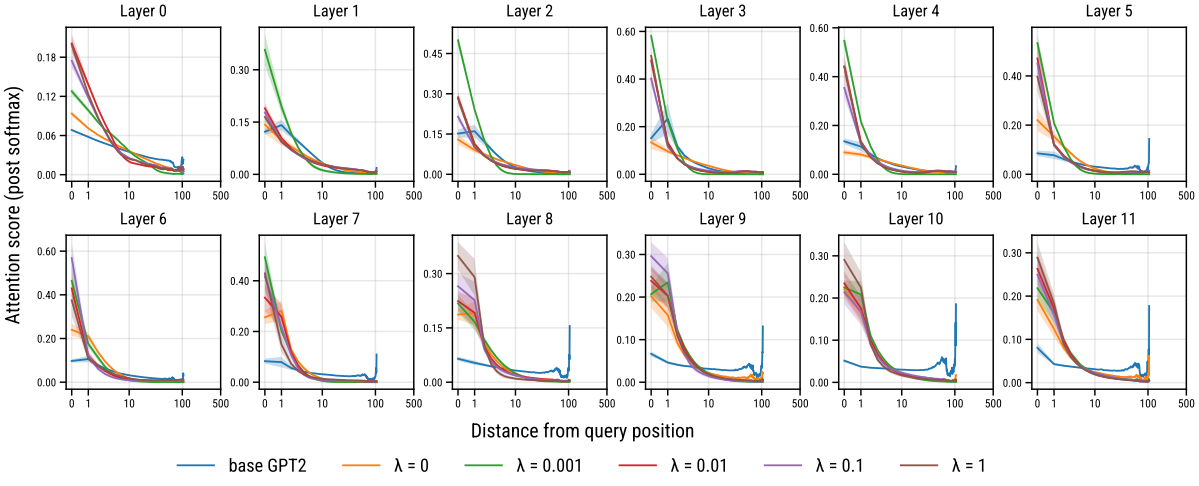


Figure 7: Aggregated attention score per layer on Provo corpus

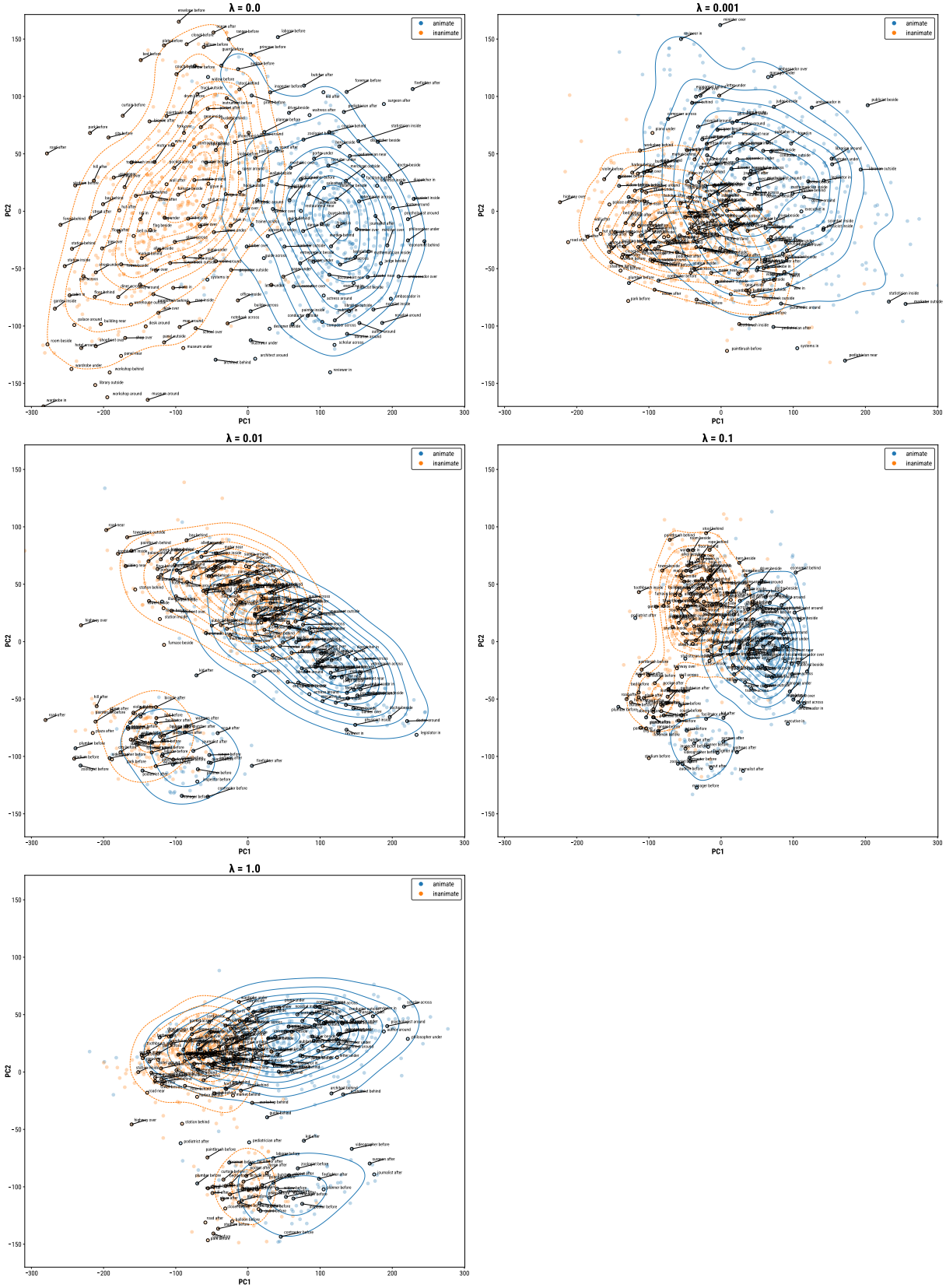


Figure 8: Representational space of next-word prediction visualized from the first two PCs with 200 randomly sampled noun-phrase labels including `Noun1` and `Prep`.