Working Memory Constraints Scaffold Learning in Transformers under Data Scarcity

Pranava Madhyastha[△,∞] Dagmar Adamcová[⊕]

[△]*City, University of London*
[∞]*The Alan Turing Institute*
[⊕]*Grounded Machines*

Abstract

We investigate the integration of human-like working memory constraints into the Transformer architecture and implement several cognitively inspired attention variants, including fixed-width windows based and temporal decay based attention mechanisms. Our modified GPT-2 models are trained from scratch on developmentally plausible datasets (10M and 100M words). Performance is evaluated on grammatical judgment tasks (BLiMP) and alignment with human reading time data. Our results indicate that these cognitively-inspired constraints, particularly fixed-width attention, can significantly improve grammatical accuracy especially when training data is scarce. These constrained models also tend to show a stronger alignment with human processing metrics. The findings suggest that such constraints may serve as a beneficial inductive bias, guiding models towards more robust linguistic representations, especially in data-limited settings.

1 Introduction

The dominant self-attention mechanism in Transformer-based models (Vaswani et al., 2017) diverges profoundly from established cognitive theories of human language processing. A key area of divergence lies in how information is accessed and maintained over time. Human comprehenders rely on working memory, a short-term mental buffer that temporarily stores and manipulates information during language processing. This memory system is understood to be capacity-limited (Miller, 1956; Cowan, 2001), subject to temporal decay (Baddeley, 2000), and influenced by serial position effects like primacy and recency (Glanzer and Cunitz, 1966). In stark contrast, standard Transformer self-attention allows near-uniform access to all tokens within a potentially very large context window, lacking inherent architectural biases that reflect these fundamental human cognitive constraints. While language models like GPT-2 have proven useful in cognitive modelling, correlating well with measures such as reading times and neural activity (Goodkind and Bicknell, 2018; Wilcox et al., 2020; Madhyastha et al., 2023), this success may occur despite, rather than because of, architectural alignment with human processing limitations.

This discrepancy motivates the central research question of our work: can integrating architectural constraints inspired by human working memory (henceforth refered as WM) produce models that not only exhibit more human-like linguistic behaviour but also learn more efficiently? To this end, our contribution is threefold. First, we implement and systematically compare four distinct, cognitively-motivated attention mechanisms: (1) a strict, fixed-width attention window to model capacity limits; (2) exponential and (3) logistic decay mechanisms to model recency bias; and (4) a primacy-recency model to capture serial position effects. Second, we train these models entirely from scratch on developmentally plausible datasets of 10 million and 100 million words, allowing these constraints to shape the learning process from its inception. Third, we conduct a multi-faceted evaluation, assessing both grammatical competence on the BLiMP benchmark (Warstadt et al., 2020) and alignment with human processing data and the nature of the internal representations learned by the models.

Our results show that imposing attention constraints, particularly restrictive fixed-width attention windows, serves as a powerful inductive bias that yields significant benefits. In low-data settings, constrained models substantially outperform the standard GPT-2 baseline in grammatical accuracy, confirming that such biases are highly effective when data is scarce. We also see that these models produce surprisal values that are markedly more predictive of human processing data, suggesting a

closer alignment with the cognitive dynamics of comprehension. While this performance gap narrows with more training data, psychometric alignment in fact degrades for all models at the larger scale, consistent with a growing body of evidence that the language-modelling objective and human comprehension are not asymptotically convergent (see for e.g., Oh and Schuler, 2023; Shain et al., 2024; de Varda et al., 2024). Attention visualisations and structural probing further clarify why these constraints help. Limiting the context forces the model to develop specialised, linguistically interpretable attention heads and a more explicit internal encoding of syntactic structure. The unconstrained baseline shows no comparable specialisation, with attention patterns that remain diffuse across layers.

2 Background

In psycholinguistics, verbal WM constraints are well-documented experimentally. For example, long or structurally complex sentences have been shown to impose a burden on memory resources (Caplan and Waters, 2013; Just and Carpenter, 1992), and comprehenders appear to manage this load through incremental processing that prioritises local dependencies, with representations of earlier material becoming progressively less accessible (see Levy, 2008; Hahn et al., 2022; Futrell et al., 2020; Gibson, 1998, *inter alia*). A strong claim, advanced by Christiansen and Chater (2016) and others (e.g., Kirby, 1999), is that the structure of human language itself reflects these constraints: features ranging from preferred word orders to limits on centre-embedding may be adaptations to a working memory bottleneck. Experimental support in this regard comes from typological evidence that dependency lengths are systematically minimised across languages (Futrell et al., 2015).

Recent work in computational psycholinguistics has begun to probe how these constraints might be reflected in language models, often within the developmentally plausible data scales (10–100M words) emphasised by the BabyLM Challenge (Warstadt et al., 2023), which enables investigation of cognitive constraints in regimes more comparable to human language acquisition. Ryu and Lewis (2021) show that pre-trained GPT-2 implicitly captures similarity-based interference effects, and Timkey and Linzen (2023) advocate simpler architectures grounded in cue-based retrieval theories (Van Dyke and Lewis, 2003). A more direct line of work modifies attention itself: De Varda and Marelli (2024) apply an exponential decay bias to a pre-trained GPT-2 model and report improved reading-time prediction; Kuribayashi et al. (2022) show that truncating context at inference improves cognitive alignment of surprisal; Clark et al. (2025) train with ALiBi (Press et al., 2021), a linear bias that softly downweights distant tokens. Janik (2023) reports that standard Transformers exhibit weak primacy and recency effects emergently, though these are scale-sensitive and appear to be statistical artefacts of training.

These interventions on attention share a common motivation but each leaves the constraint operating at the periphery of the model. Post-hoc decay biases (De Varda and Marelli, 2024) and inference-time context truncation (Kuribayashi et al., 2022) modify the output of an already-trained system whose representations were shaped without any such constraint. Soft positional biases such as ALiBi (Clark et al., 2025; Press et al., 2021) are present during training but only discourage rather than explicitly prevent long-range attention, which may allow the model to bypass the constraint when useful. Emergent effects in unconstrained models (Janik, 2023) are scale-sensitive and fragile.

In light of these limitations, we integrate working memory constraints directly into the GPT-2 architecture from the outset of training. This contrasts with approaches based on implicit learning or post-hoc modification, which leave the underlying representations unconstrained. Specifically, we investigate how imposing constraints inspired by prominent findings in human working memory research (including capacity limitations, temporal decay, and serial position effects) impacts model performance and language processing.

3 Methods

We implement four distinct attention mechanisms directly into the GPT-2 architecture to simulate human-like WM constraints. Each of these mechanisms is designed to model specific features of human WM.

3.1 Fixed Window Attention

A hallmark feature of human working memory is its limited storage capacity. While classic accounts attempted to quantify this capacity as a discrete number of items or “chunks", arriving at numbers

between four and nine (Miller, 1956; Cowan, 2001), modern theories however have moved toward more dynamic models. These accounts posit that capacity is not a fixed unit count but is instead a skill that emerges from language comprehension and production processes, viewing verbal WM as the activated portion of linguistic long-term memory. This perspective suggests that the ability to maintain and order verbal information is intrinsically linked to an individual's language experience and proficiency, rather than relying on separate, discrete storage buffers (Schwering and MacDonald, 2020; MacDonald, 2016; Buchsbaum and D'Esposito, 2019). Nevertheless, we find it useful to simulate this very restrictive capacity constraint as a discrete cut-off in order to isolate local dependencies. We do this by implementing a fixed-width attention window.

In this approach, the attention calculation for each token is restricted to a *fixed-size window* of a set of fixed preceding tokens. For a token at position i, attention is only computed over tokens in the range $[\max(0, i - W + 1), i]$, where W is the fixed window size. This is implemented using an attention mask M_{window} that prevents access to tokens outside the window:

$$M_{window}^{(i,j)} = \begin{cases} 0, & \text{if } \max(0, i - W + 1) \leq j \leq i \\ -\infty, & \text{otherwise} \end{cases} \tag{1}$$

The attention weights are then calculated as: $a'_{ij} = a_{ij} + M_{window}^{(i,j)}$, where a_{ij} are the original attention weights. This sets the attention weights for tokens outside the window to $-\infty$ (and thus zero after `softmax` normalisation), while tokens inside the window retain their original weights. The mechanism imposes a hard capacity limit at each layer, forcing the model to operate within a strictly local context. The selection of fixed window sizes for our models is directly motivated by influential findings in cognitive science concerning the capacity limits of human WM. A window size of $k = 4$ is informed by contemporary research, such as that of Cowan (2001), who hypothesised that short-term memory has a capacity of approximately four 'chunks' of information. The window sizes of $k \in 5, 7, 9$ are derived from Miller's (1956) seminal observation concerning "the magical number seven, plus or minus two", which represents the classic estimate for the number of items an individual can hold in immediate memory. In the context of our experiments, we treat a token as a fundamental information chunk.

In the context of NLP, a form of fixed window attention is a core component of models designed for long documents, such as Longformer (Beltagy et al., 2020) and Block-Sparse Transformer (Child et al., 2019). While these models are motivated by efficiency, our fixed window attention is motivated by mirroring the limited capacity of human WM, suggesting a cognitive basis for such architectural choices in NLP.

3.2 Primacy-Recency Attention

Human memory recall has been hypothesised to exhibit primacy and recency effects (Glanzer and Cunitz, 1966; Morrison et al., 2014, *inter alia*). Items presented at the beginning (primacy) and end (recency) of a list are typically better recalled than items in the middle. In the context of language processing, this suggests that initial and final parts of a sequence might hold disproportionate importance in shaping the overall representation. We incorporate constraints of this kind through a primacy-recency attention mechanism. This mechanism adds a position-dependent bias to the attention weights that emphasise both the initial and final tokens in the sequence. It learns two parameters, w_{primacy} and w_{recency}, which are initialized to 0.5 during training. We calculate primacy weights p_i and recency weights r_i for each position i in a sequence of length L:

$$p_i = \frac{e^{-i/L}}{\sum_{j=0}^{L-1} e^{-j/L}} \tag{2}$$

$$r_i = \frac{e^{-(L-1-i)/L}}{\sum_{j=0}^{L-1} e^{-(L-1-j)/L}} \tag{3}$$

where i is the position index (starting from 0). Primacy weights decay exponentially from the beginning of the sequence, while recency weights decay exponentially from the end. Both sets of weights are normalised to sum to one. The final bias b_i for each position is a weighted combination of primacy and recency weights: $b_i = w_{primacy} \cdot p_i + w_{recency} \cdot r_i$, where $w_{primacy}$ and $w_{recency}$ are learnable weights that control the relative contribution of primacy and recency biases. These biases are then added to the attention weights as: $a'_{ij} = a_{ij} + b_j$. We note here that the bias b_j is added based on the *key* position j (i.e., the position of the token being attended to in the attention mechanism). Our intention with this position-dependent bias is to encourage the model to attend

more strongly to tokens at the beginning and end of the sequence, reflecting primacy and recency effects from psycholinguistic theories. We also separately run an ablation with exclusively primacy and recency based attention to understand the impact of each of these mechanisms. While less directly related to computational efficiency in long sequence processing, the primacy-recency attention mechanism aligns with the broader trend in NLP towards incorporating positional information in more sophisticated ways than simple positional embeddings. For instance, relative positional embeddings (Shaw et al., 2018) and complex positional encodings (Su et al., 2021) aim to capture richer positional relationships. In some way, this attention modification helps focus on more positional information to emphasize the structural importance of sequence beginnings and endings.

3.3 Exponential Decay Attention

Inspired by recent psycholinguistic theories that highlight the interplay between linguistic expectations and WM constraints in human language processing (see Smith and Levy, 2013; Gibson, 1998; Hahn et al., 2022, *inter alia*), we also consider an exponential decay attention mechanism. This modification is directly motivated by the work of De Varda and Marelli (2024), who propose biasing Transformer models to prioritize local linguistic context, simulating a lossy representation of distant contextual information in human sentence processing. Here, the exponential decay attention mechanism modulates the standard attention weights by incorporating a decay factor that diminishes the influence of tokens based on their temporal distance. The modified attention weight a'_{ij} between token i and token j is calculated as:

$$a'_{ij} = (1 - \alpha)a_{ij} + \alpha e^{-|i-j|\cdot\lambda} \tag{4}$$

where a_{ij} represents the original dot-product attention weight, λ is the decay rate, and α is a mixing parameter. The exponential term $e^{-|i-j|\cdot\lambda}$ introduces a bias favouring attention to closer tokens, effectively implementing a recency effect by exponentially reducing the contribution of more distant tokens. Following De Varda and Marelli (2024), we adopt the hyperparameters $\lambda = 82.86$ (corresponding to `decay_rate` in our implementation) and $\alpha = 0.37$. These values were identified as optimal in their grid search using GPT-2-small on the Provo corpus (Luke and Christianson, 2018). De Varda and Marelli (2024) demonstrated that applying a post-hoc exponentially decaying attention bias to a pretrained GPT-2 model improved its correlation with human reading times. While these results are useful, this approach modifies an already developed system rather than allowing constraints to shape the learning process from the outset. To remedy this, our methodology involves training the customised GPT-2 model from scratch with the exponential decay attention mechanism inherently integrated into its architecture. This approach will allow the model learns to process language under the constraint of locality-biased attention from the outset. We hypothesise that training with this constraint from the beginning may lead to a more concordant and effective integration of the psycholinguistic principle, as the model architecture is aligned with the intended processing mechanism throughout the learning process, rather than having the bias imposed after the model has already learned with a different attention paradigm. This approach has clear parallels within the field of NLP, where locality-sensitive attention is explored for efficiency reasons when handling long sequences. This is evident in models such as Performer (Choromanski et al., 2020) and in Transformers that use linear biases (Katharopoulos et al., 2020; Press et al., 2021).

3.4 Logistic Decay Attention

While exponential decay offers a straightforward way to model temporal forgetting, its immediate and sharp decline may not accurately reflect how WM handles recent verbal material. This is particularly relevant when considering the recency effect, where the most recent information remains highly accessible for a short period before its recall probability diminishes (Glanzer and Cunitz, 1966). To reconcile this with the concept of a limited memory span, we introduce a second temporal decay mechanism: logistic decay. Unlike an exponential bias, which applies a consistent rate of forgetting from the outset, a logistic function imposes a non-linear S-shaped curve. We hypothesise that this curve can more realistically represent the dynamics of human WM by providing a short period of sustained high accessibility for recent tokens before a rapid drop-off in influence, thereby combining elements of both a discrete capacity cut-off as reflected in our fixed-window attention and a more nuanced decay function. The logistic decay mechanism modulates attention weights based on the temporal distance

between tokens using a psychometric function. For tokens at positions i and j, the attention weight modification is computed as:

$$w_{ij} = \frac{1}{1 + e^{k \cdot (d_{ij} - m)}} \tag{5}$$

where $d_{ij} = \max(1, i - j + 1)$ represents the distance between tokens, k is the steepness parameter controlling the sharpness of the decay curve, and m is the midpoint parameter determining the distance at which attention weight equals 0.5. The final attention weights are calculated by multiplicatively combining the original attention scores with the logistic prior: $a'_{ij} = a_{ij} \cdot w_{ij}$. We set $k = 0.4$ and $m = 12.0$ as default parameters, establishing a psychologically motivated attention profile where tokens within approximately 5 positions maintain relatively strong (high) attention weights, while more distant tokens experience rapid attention decay. The logistic decay attention mechanism exhibits several key characteristics that distinguish it from other approaches. First, it maintains relatively stable attention weights for nearby tokens (distance $< m$), followed by a rapid transition to low attention weights for distant tokens (distance $> m$). This creates a more pronounced boundary between accessible and inaccessible memory, aligning with accounts of discrete WM span. Second, unlike fixed window attention which implements a hard cutoff, logistic decay provides a smooth but steep transition, avoiding the potential discontinuities associated with binary attention masking.

4 Experimental Setup

This section details our experimental setup. We introduce the set of models evaluated, the datasets employed for training, and the overall framework for our analysis, which focuses on training language models from scratch with modified attention mechanisms. Our experiments are conducted on the GPT-2-small architecture (Radford et al., 2019). Our core models are based on the standard GPT-2 configuration, but incorporate custom attention mechanisms implemented by modifying the `GPT2Attention` module. We also train two baseline models using the default GPT-2 configuration provided by the Hugging Face `transformers` library (**?**).

4.1 (Pre-)Training Corpora

Our primary interest lies in evaluating language models under conditions that are more cognitively plausible in terms of data scale than typical large language model pretraining. Therefore, we utilize the BabyLM dataset (Warstadt et al., 2023). The BabyLM Challenge itself drew inspiration from the scale of data available during human language acquisition. Specifically, we use the training portions of "Strict-Small" (10 million words) and "Strict" (100 million words) training subsets provided by the BabyLM Challenge (Warstadt et al., 2023). These datasets comprise text from sources considered potentially relevant to child language exposure, including Simple English Wikipedia, children's books from Project Gutenberg, CHILDES transcripts, the British National Corpus, OpenSubtitles, and the Switchboard Dialog Act Corpus. We note that all of our models also use GPT2Tokenizers which are trained specifically on 10 million and 100 million words based corpora separately.

4.2 Training Configuration

For all models, we used the GPT-2 small architecture as a base. Training was performed using the AdamW optimizer with a learning rate of $5e^{-5}$, a batch size of 64, and a weight decay of 0.01. Models were trained for 5 epochs with a batch size of 50. Gradient clipping was applied with a maximum norm of 1.0. These settings were kept consistent across all model variants to ensure a fair comparison. These hyperparameters are similar to the range of empirical setups common in Warstadt et al. (2023).

4.3 Tasks

We evaluate the trained models on two distinct tasks designed to probe different aspects of their linguistic capabilities and cognitive plausibility.

BLiMP The first task assesses the sensitivity of the models to English grammatical structure using the Benchmark of Linguistic Minimal Pairs (BLiMP) (Warstadt et al., 2020). BLiMP consists of numerous sub-tasks, each targeting a specific linguistic phenomenon. Every example in BLiMP presents a "minimal pair": one sentence that is grammatically acceptable and another that is unacceptable, with the two differing only minimally (often by a single word or morpheme). A language model is considered correct on a given minimal pair if it assigns a *higher probability score* (hence low surprisal) to the acceptable sentence compared to the unacceptable one. Success across BLiMP

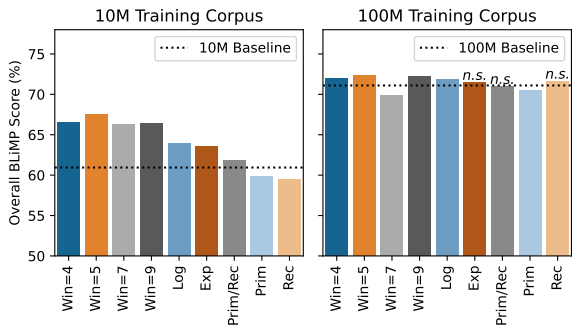


Figure 1: Performance comparison of GPT-2 self-attention modifications on BLiMP tasks for 10M (left) and 100M (right) parameter models. Dashed lines indicate the performance of the baseline GPT-2 model (without any attention modifications) trained from scratch on the corresponding dataset size. All models show statistically significant differences from baseline ($p < 0.001$) except where marked as not significant (n.s.).

tasks indicates that the model has learned representations consistent with fine-grained grammatical distinctions in English.

Psychometric Benchmark The psychometric data we use are drawn from established experimental paradigms (detailed in de Varda et al. 2024) with measurements averaged across several neural and behavioural indices of cognitive processing on a set of 1725 sentences. This includes a) Eye-Tracking Data with measures such as First Fixation Duration, which reflects early lexical access, and later-stage integrative measures like Gaze Duration, Go-Past Time, and Right-Bounded Time; b) Self-Paced Reading Time which is a controlled measure of reading speed, influenced by a range of semantic and syntactic factors; c) Event-Related Potentials which are neural signals that provide a fine-grained temporal view of language processing. These include the N400 component, which is modulated by meaning processing; the P600, indicative of syntactic reanalysis and integration; and various Left Anterior Negativity components (LAN and ELAN) associated with phrase structure building.

5 Results

5.1 Overall observations on grammaticality

We first examine the overall performance of our modifications on attention mechanisms across 10 million and 100 million word corpora (Figure 1). We observe a clear trend in the low-data setting. We see that all models with our modified attention mechanisms (apart from the Primacy and Recency ablations) demonstrate a substantial and statistically significant improvements ($p \leq 0.001$) in average BLiMP accuracy compared to the baseline model. While the baseline scores approximately 61% average accuracy on the task, models with fixed attention windows, which impose the most stringent constraints, achieve markedly higher scores of around 68%. This result highlights the benefit of architectural inductive biases derived from WM principles, particularly when training data is scarce.

However, this advantage diminishes when models are trained on the larger 100M-word dataset. In this higher-data regime, the baseline's performance improves markedly to an accuracy of approximately 71%, narrowing the performance gap considerably. Notably, models with exponential or primacy-and-recency constraints on attention show no statistically significant difference from the baseline. With a sufficient volume of data, the standard attention mechanism evidently recovers much of the capability required for this task. Despite this, the best-performing constrained models maintain a small but statistically significant edge, indicating that their inductive biases remain beneficial even at a larger scale.

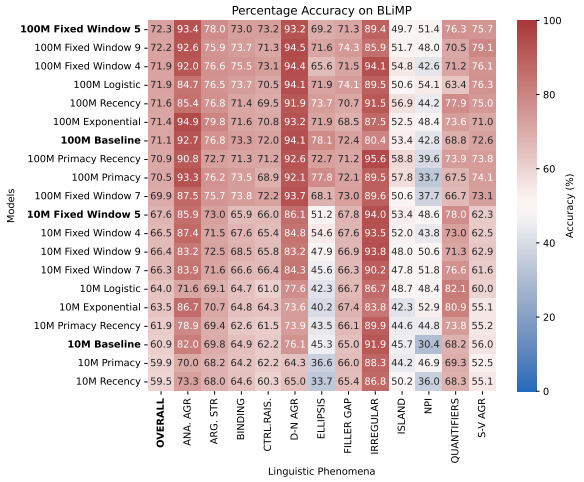


Figure 2: Model performance across linguistic phenomena evaluated on the BLiMP task. Models are sorted by Overall score in a descending order; 50% indicates chance performance. Best performing 10M and 100M models are highlighted along with the corresponding Baselines.

5.2 Performance across linguistic phenomena

In order to understand how our modifications to attention influence performance on specific linguis-

tic tasks, we analysed the results across individual BLiMP sub-tasks. In Figure 2, we categorise the different tasks into a set of classes motivated by the analyses in Warstadt et al. (2020). Consistent with previous findings, our models perform best on phenomena related to morphological agreement, which often rely on local dependencies. Across both 10 million and 100 million data regimes, all models achieve reasonably high accuracy on Determiner-Noun Agreement (which focuses on number agreement, e.g., that chair vs. that chairs), Irregular Forms, and Anaphor Agreement (where the focus is on reflexive pronoun agreement, e.g., girls insulted themselves vs. herself). We note that the fixed-window models, which explicitly enforce locality, excel here. We also find that most models suffer on more abstract syntactic and semantic constraints. Performance on Island Effects (restrictions on syntactic movement) and NPI Licensing (the requirement for words like ever to be in a negative context) is the lowest across the board, often only marginally better than chance. This confirms that these phenomena, which predominantly require sensitivity to complex structural and logical scope, represent a persistent challenge for language models, and our architectural modifications do not offer a simple solution to assist models on these tasks.

On the other hand, the results for Argument Structure, which governs a verb's ability to appear with certain arguments (e.g., disturbing a person vs. boasting a person), are particularly interesting. Performance is generally boosted by locality constraints across both data regimens. We see that the models that have fixed window size of about 5 and exponentially decaying attention constraints (which tend to promote highly local attention structures) seem to generally perform better[1]. We also observe that Ellipsis emerges as a phenomenon highly dependent on data scale. At 10M, all models perform exceptionally poorly, with accuracies in the 33-54% range. However, at 100M words, performance improves dramatically into the 65-78% range across all models. This sharp increase suggests that the generalisations required to correctly resolve ellipsis are not readily captured with limited data, but indeed become accessible with more exposure.

An important overall finding is that highly constrained, simple models often outperform the less-constrained baseline, particularly on complex syntactic and semantic challenges. Our null hypothesis before experimentation was that heavily constrained models may significantly hamper performance in transformers, especially since a large body of work in contemporary NLP looks into methods for moving away from locality bias (Tay et al., 2020; Zaheer et al., 2020, *inter alia*). We find the performance of small, rigid attention windows on phenomena that are not strictly local to be the most surprising result. For instance, fixed window 5 is broadly one of the best performing models in complex sub categories. Especially in Argument Structure where smaller fixed window models obtain significantly higher performance, where the sentences usually focus on structures which govern a verb's relationship with its arguments. This is noteworthy because these relationships can be complex and are not solely determined by adjacent words, yet this highly local model captures them effectively. Similarly, these models perform well on Binding, which involves the structural relationship between a pronoun and its antecedent. One might expect this to require a wider context, but the small fixed window proves highly effective, outperforming the baseline. On the other hand, leaky models, both Exponential and Logistic bias based models, which allow attention to "leak" across the entire context while prioritising recent information, show interesting trends, however these models are inferior in performance compared to stricter and smaller fixed window based models.

Primacy/Recency Ablation We wanted to further understand the efficacy of primacy and recency based formulation and carried out an experiment where we trained models separately with primacy and recency biases. We present these results in Figures 1 and 2. We notice, unsurprisingly, that Recency based bias in attention tends to perform better with local structures while Primacy tends to have a slight edge on tasks with dependency structures that require access to longer distances. While both Primacy and Recency generally tend to perform worse in comparison to the baseline model, however, increasing data tends to substantially help the model with Recency based bias. This tends to correlate with fixed window models with strict local attention constraints. Finally, while the model based on Primacy and Recency is not among the best models, however in most cases, it has a tendency to even out the divergence between models

[1]We also see that the performance on Argument Structure especially is remarkably similar to GPT2-large pretrained model (Warstadt et al., 2020).

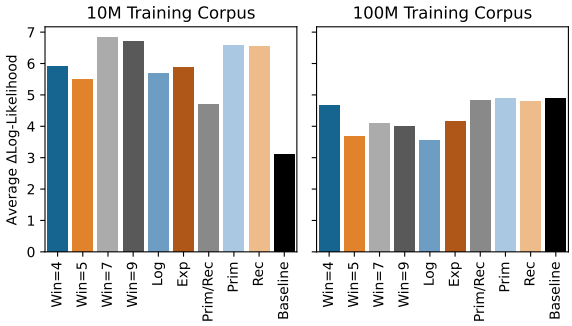


Figure 3: ΔLog-Likelihood averaged across psychometric measures for each model.

trained separately with Primacy or Recency.

5.3 Analysing the model behaviour with human processing data

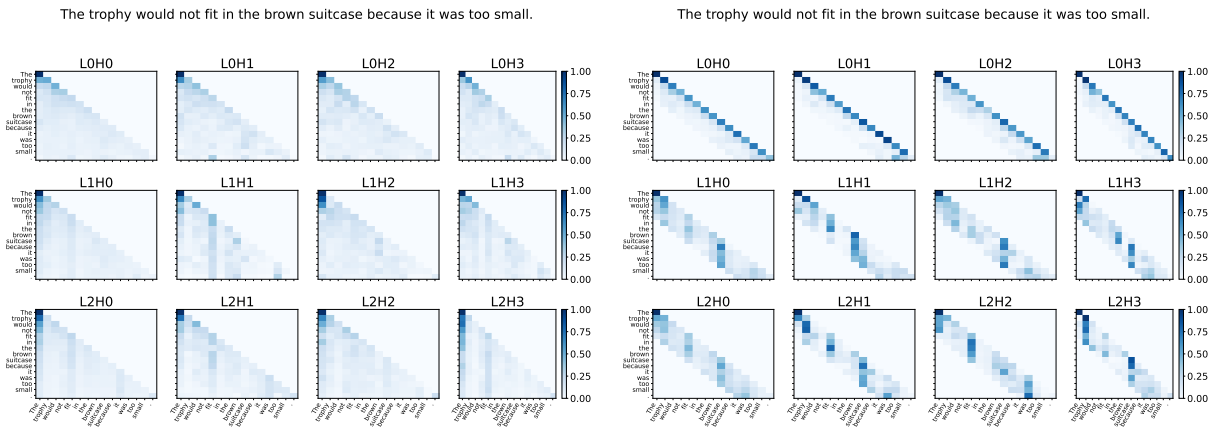
Would our cognitively inspired attention constraints correlate with human processing data? To answer this we examine the alignment of surprisal and psychometric data using averaged difference of log-likelihood (Figure 3). Here, ΔLog-Likelihood measures the additional explanatory power offered by surprisal values over a base model without surprisal. We then average this across psychometric corpora. That is, the difference in log-likelihood (ΔLog-likelihood) represents the improvement in statistical model fit when surprisal is added as a predictor, compared to a base statistical model that includes only covariates like word length and frequency. Therefore, a higher ΔLog-likelihood indicates that a language model's surprisal values are a better predictor of human cognitive effort during sentence comprehension. The 'Baseline' bar in the figure refers to the ΔLog-Likelihood achieved by GPT-2 model with unmodified attention trained on corresponding corpora, serving as our primary model for comparison.

We observe a significant trend emerging in the low data regime. Nearly all models with modified attention mechanisms produce surprisal values that are substantially more predictive of human processing data than the unmodified Transformer baseline. The baseline model achieves a relatively low average ΔLog-Likelihood of approximately 3.2, whereas the top-performing models with fixed-window attention (especially 7 and 9), as well as Primacy and Recency biases, achieve scores nearly twice as high. This result suggests that in data-constrained settings, architectural constraints that mimic principles of human WM serve as a powerful and effective inductive bias, guiding the model to learn representations that are more closely aligned with human cognitive processes.

On the other hand, the advantage of these explicit architectural biases diminishes considerably when the models are trained on the larger, 100M-word dataset. The standard baseline model's performance improves, with its average ΔLog-Likelihood increasing marginally. More importantnly, the performance gap between the baseline and the modified models narrows significantly. This pattern suggests that with an order of magnitude more data, the standard self-attention mechanism is better able to approximate the necessary behaviours and produce surprisal values that correlate well with human processing metrics. However, we note here that even for all the constrained attention models and the baseline (with unmodified attention) the ΔLog-likelihood is lower in the 100M setting compared to the 10M setting. A compelling hypothesis for this seemingly counter-intuitive trend relates to the divergence between the language modelling objective and the cognitive processes underlying human comprehension. As models are trained on more data, they become increasingly specialised at predicting the statistical patterns of the training corpus. This high degree of specialisation may cause their expectations to diverge from the more generalised predictions that humans make which is also illustrated in recent work (Oh and Schuler, 2023; de Varda et al., 2024; Shain et al., 2024, *inter alia*).

5.4 Understanding attention distribution

As an initial, exploratory step toward understanding the inductive biases in our models, we examine the internal representations along two complementary dimensions: the attention patterns within individual heads, and the syntactic structure recoverable from the contextual embeddings. Figure 4 compares the attention patterns of two models trained on the 10M corpus, the baseline GPT-2 with unmodified self-attention and a constrained model with a fixed-window mechanism (window size=5), on the sentence "The trophy would not fit in the brown suitcase because it was too small", a classic test case for pronoun resolution. We focus the figure on early layers, though the pattern is consistent across higher layers. We observe significant diffences. The fixed-window model shows an immediate and sharp focus on local context, with heads in the initial layer already concentrating on



(a) 10M Baseline

(b) 10M Fixed Window 5

Figure 4: Attention weights distributions for the sentence "The trophy would not fit in the brown suitcase because it was too small." across individual heads in the first three layers of the 10M Baseline and Fixed Window 5 models.

the immediately preceding token. This specialisation sharpens in later layers: heads `L2H0` and `L2H1` come to focus on the core subject-verb-object structure ('trophy', 'fit', 'suitcase'), a pattern reminiscent of the telegraphic speech observed in children between 18 and 36 months (Brown, 1973); `L2H2` specialises in verbs ('fit', 'was') and `L2H3` in nouns ('trophy', 'suitcase'). The baseline model mostly shows dissimilar patterns where its heads remain diffuse across layers, attending to less interpretable combinations of function and content words without evident functional differentiation. The fixed-window constraint appears to force a structured division of labour amongst the attention heads, producing a representation of syntax that is more explicit than anything that emerges in the unconstrained model.

Following Hewitt and Manning (2019) and its recent extension by Someya et al. (2025), we further train a diagnostic linear projection to reconstruct unlabelled dependency trees from each model's contextual embeddings, reporting Unlabelled Unrooted Attachment Score (UUAS) by dependency relation. The fixed-window model achieves higher UUAS than the baseline across all five relations tested, with the largest gaps on core grammatical relations such as `nsubj` and `dobj`, which capture the link between a verb and its arguments. The advantage emerges early in the network and is most pronounced at intermediate layers. Full per-layer and per-relation results are in Appendix B.

The architectural constraint seems to shape the local attention patterns within heads, producing interpretable functional specialisation. It also tends to alter the global geometry of the representational space such that syntactic dependencies become more linearly recoverable.

6 Conclusions

We investigated the integration of cognitively inspired working memory constraints into the Transformer architecture, comparing fixed-width attention windows, exponential and logistic decay mechanisms, and primacy and recency biases against an unmodified GPT-2 baseline trained on developmentally plausible corpora of 10M and 100M words. Across grammatical competence on BLiMP, alignment with a convergent battery of neural and behavioural processing measures, and analyses of the models' internal representations, the results converge on a single picture. Constraints aligned with the limits of human working memory function as a potent inductive bias, yielding gains in data efficiency, predictive alignment with human comprehension, and the development of linguistically interpretable internal structure. The advantage is most pronounced in the data-limited regime that most closely resembles the conditions under which humans acquire language, and the psychometric alignment of all models, both constrained and unconstrained, degrades as training data increases, consistent with growing evidence that the language-modelling objective and human comprehension diverge as models grow larger. These findings push against the contemporary trends towards longer contexts and weaker inductive biases, and suggest that hard cognitive constraints actively scaffold learning rather than hindering it.

Limitations

We believe several bounds on the present work deserve explicit acknowledgement. Our investigation is deliberately confined to the developmentally plausible regimes of 10M and 100M words, and we make no claim about what happens at the trillion-token scale of contemporary frontier models. The trajectory of psychometric alignment we observe between 10M and 100M, in which alignment degrades rather than improves with more data, is at least consistent with the possibility that the specialised structure our constrained models acquire reflects an inductive bias absent from the unconstrained architecture rather than a pattern recoverable from data alone, but our experiments don't currently dont provide strong experimental evidence that this is the case with all dominant models. Our implementation of working memory is also a deliberate simplification. Working memory is not a rigid, position-indexed buffer. It is perhaps a dynamic, content-addressable system shaped by similarity-based interference among held items (Van Dyke and Lewis, 2003) and by linguistic experience itself (Schwering and MacDonald, 2020; MacDonald, 2016). In our experiments, we isolate one dimension, namely the locality and decay of accessible context, and the development of attention mechanisms reflecting interference and content-addressable retrieval is the natural next step. Our empirical evaluation is also restricted to English, and the locality constraints that benefit our models here may interact quite differently with languages whose dependencies are mediated by say morphological case, by free word order, or by head-final configurations.

A more interesting class of limitations comes from the results themselves. While constrained models perform well across most BLiMP categories, their performance on Island Effects remains only marginally above chance, mirroring the unconstrained baseline. This raises the genuine possibility that locality constraints, useful as they are for the acquisition of local syntactic dependencies, may impede the acquisition of phenomena that require sensitivity to global syntactic structure. Whether cognitive constraints have natural analogues that operate over abstract *structural* rather than linear distance is, to us, among the more interesting directions this work opens up. The most fundamental bound, however, is one our study shares with nearly all contemporary computational psycholinguistics. Human language is acquired, produced, and comprehended in densely multimodal contexts where gesture, prosody, gaze, and pragmatic interaction contribute centrally to meaning (see Holler and Levinson, 2019, for a broader perspective). Working memory in everyday language use is integrated across these multiplicities of modalities and record of experiences. Our work, like the overwhelming majority of language modelling research, operates within a unimodal text-only paradigm, and the cognitive plausibility achievable within that paradigm is correspondingly bounded. Whether the architectural constraints we study have analogues that would prove useful in genuinely multimodal language models is perhaps the most interesting question this work leaves open.

Acknowledgements

This work was supported in part by the Alan Turing Institute under Fundamental Research (Project No. PP00029).

References

Alan Baddeley. 2000. The episodic buffer: a new component of working memory? *Trends in cognitive sciences*, 4(11):417–423.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Roger Brown. 1973. *A first language: The early stages*. Harvard University Press.

Bradley R Buchsbaum and Mark D'Esposito. 2019. A sensorimotor view of verbal working memory. *Cortex*, 112:134–148.

David Caplan and Gloria Waters. 2013. Memory mechanisms supporting syntactic comprehension. *Psychonomic bulletin & review*, 20:243–268.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

arXiv preprint arXiv:2009.14794.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, and David Belanger, Lucy Colwell, and Adrian Weller. 2020. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*.

Morten H Christiansen and Nick Chater. 2016. The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and brain sciences*, 39:e62.

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.

Nelson Cowan. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, 24(1):87–114.

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, Bangkok, Thailand. Association for Computational Linguistics.

Andrea Gregor de Varda, Marco Marelli, and Simona Amenta. 2024. Cloze probability, predictability ratings, and computational estimates for 205 english sentences, aligned with existing eeg and reading time data. *Behavior Research Methods*, 56(5):5190–5213.

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, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

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

Murray Glanzer and Anita R Cunitz. 1966. Two storage mechanisms in free recall. *Journal of verbal learning and verbal behavior*, 5(4):351–360.

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

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 Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Judith Holler and Stephen C Levinson. 2019. Multimodal language processing in human communication. *Trends in cognitive sciences*, 23(8):639–652.

Romuald A Janik. 2023. Aspects of human memory and large language models. *arXiv preprint arXiv:2311.03839*.

Marcel A Just and Patricia A Carpenter. 1992. A capacity theory of comprehension: individual differences in working memory. *Psychological review*, 99(1):122.

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.

Simon Kirby. 1999. *Function, selection, and innateness: The emergence of language universals*. OUP Oxford.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Steven G Luke and Kiel Christianson. 2018. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50:826–833.

Maryellen C MacDonald. 2016. Speak, act, remember: The language-production basis of serial order and maintenance in verbal memory. *Current Directions in Psychological Science*, 25(1):47–53.

Pranava Madhyastha, Ye Zhang, and Gabriella Vigliocco. 2023. Are words equally surprising in audio and audio-visual comprehension? *arXiv preprint arXiv:2307.07277*.

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.

Alexandra B Morrison, Andrew RA Conway, and Jason M Chein. 2014. Primacy and recency effects as indices of the focus of attention. *Frontiers in human neuroscience*, 8:6.

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.

Ofir Press, Noah A Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*.

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

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

Steven C Schwering and Maryellen C MacDonald. 2020. Verbal working memory as emergent from language comprehension and production. *Frontiers in human neuroscience*, 14:68.

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.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Taiga Someya, Ryo Yoshida, Hitomi Yanaka, and Yohei Oseki. 2025. Derivational probing: Unveiling the layer-wise derivation of syntactic structures in neural language models. *arXiv preprint arXiv:2506.21861*.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *Preprint*, arXiv:2104.09864.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2020. Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

William Timkey and Tal Linzen. 2023. A language model with limited memory capacity captures interference in human sentence processing. *arXiv preprint arXiv:2310.16142*.

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

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. Curran Associates, Inc.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 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*, pages 1–34, Singapore. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Ethan Wilcox, Peng Qian, Richard Futrell, Ryosuke Kohita, Roger Levy, and Miguel Ballesteros. 2020. Structural supervision improves few-shot learning and syntactic generalization in neural language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4640–4652, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297.

A Architectural Parameters and Training Configuration

A.1 Base Architecture

All models in this study use the GPT-2 small architecture as a base, with the following parameters: 12 transformer decoder layers, 768-dimensional token embeddings and hidden states, 12 parallel attention heads per layer, and a maximum context window of 1024 tokens. The combination yields approximately 124 million trainable parameters per model. The results are averaged over 3 runs.

A.2 Per-Mechanism Parameter Justifications

Fixed Window Attention. We evaluate window sizes $W \in \{4, 5, 7, 9\}$. The choice of $W = 4$ is motivated by Cowan (2001), who proposed an upper bound of approximately four chunks on the contents of focal attention in short-term memory under controlled conditions. The choices of $W \in \{5, 7, 9\}$ are derived from Miller (1956), whose "magical number seven, plus or minus two" represents the classical estimate of the number of items maintainable in immediate memory. In the present setting, we treat each token as a fundamental information chunk, acknowledging that this is a simplification of human chunking, which operates over linguistically meaningful units of variable size.

Exponential Decay Attention. We adopt $\lambda = 82.86$ and $\alpha = 0.37$, the values reported as optimal by De Varda and Marelli (2024) from a grid search using GPT-2 small on the Provo corpus (Luke and Christianson, 2018). We retain these values without re-tuning in order to enable direct methodological comparison with the post-hoc application reported in that work.

Logistic Decay Attention. We use $k = 0.4$ as the steepness parameter and $m = 12.0$ as the midpoint distance. These values establish a profile where tokens within approximately five positions retain high attention weight before the decay accelerates.

Primacy and Recency Attention. The mechanism includes two learnable scalar parameters, w_{primacy} and w_{recency}, which control the relative contribution of primacy and recency biases. Both are initialised at 0.5 at the start of training.

A.3 Training Configuration

All models are trained from scratch using the AdamW optimiser with a learning rate of 5×10^{-5}, batch size of 64, and weight decay of 0.01. Training proceeds for 5 epochs with gradient clipping at a maximum norm of 1.0. These hyperparameters follow the empirical setups commonly used in the BabyLM Challenge (Warstadt et al., 2023) and are kept consistent across all model variants to ensure fair comparison. Tokenisation uses the standard GPT-2 byte-pair encoding, with separate tokenisers trained on the 10M and 100M corpora respectively to match the data scale of each setting.

B Structural Probing Analysis

Our structural probing analysis applies the derivational probing framework of Someya et al. (2025), which extends the structural probe of Hewitt and Manning (2019) by making explicit the layer-wise derivation of syntactic structures in neural language models. The method trains a diagnostic linear projection $B \in \mathbb{R}^{k \times d}$ over a model's contextual embeddings $h_i \in \mathbb{R}^d$, such that the squared L_2 distance between projected embeddings approximates the tree distance between the corresponding tokens in the syntactic dependency parse:

$$d_B(h_i, h_j)^2 = (B(h_i - h_j))^T (B(h_i - h_j)) \quad (6)$$

The probe is trained to minimise the difference between d_B and the true parse-tree distance d_T over a training set of parsed sentences. From the resulting distance matrix at evaluation time, we recover the unlabelled syntactic dependency structure using a minimum spanning tree algorithm, and we report the Unlabeled Unrooted Attachment Score (UUAS) by relation type. UUAS measures the proportion of edges in the recovered tree that match an edge in the gold parse, ignoring direction and dependency labels.

We probe representations at every layer of the model (layers 0–11 for GPT-2 small) and report results for the five most frequent dependency relations in our evaluation corpus: `nsubj` (nominal subject), `dobj` (direct object), `prep` (prepositional modifier), `attr` (attribute), and `root` (sentence root). Probes are trained on gold-parsed sentences from the English Web Treebank, with embeddings extracted from each model in inference mode (no gradient flow back to the language model). Probe dimensionality is set to $k = 64$, optimised with Adam at a learning rate of 10^{-3} for 30 epochs, with early stopping on validation loss. We evaluate on a held-out test partition.

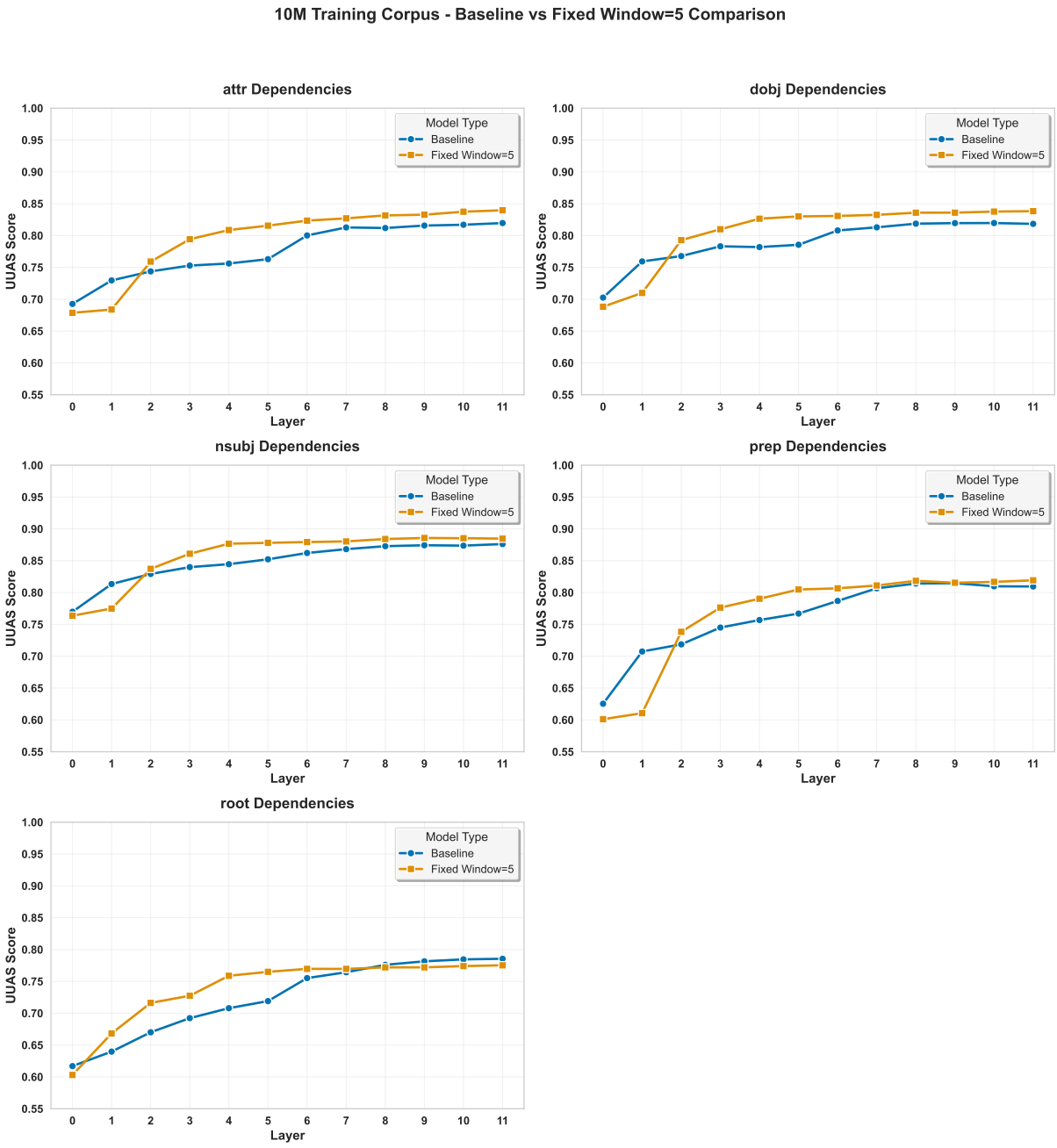


Figure 5: UUAS Score comparison between the Baseline and Fixed Window 5 models trained on 10M words.

We believe that this structural probing complements behavioural evaluation in a specific way. BLiMP measures whether a model assigns higher probability to grammatically acceptable sentences than to unacceptable ones, but does not directly tell us whether the model has internalised the structural relations that underlie the contrast. A model could in principle perform well on agreement minimal pairs by tracking surface co-occurrence statistics rather than by representing the binding relation between subject and verb. In a way, structural probing addresses this gap by asking whether the geometry of a model's representations encodes the grammatical structure of its inputs in a recoverable form. A higher UUAS indicates that syntactic dependencies are linearly decodable from the embedding space, which we take as evidence that the model has acquired a more explicit internal representation of structure rather than an implicit, behaviourally adequate proxy.

B.1 Results by Dependency Relation

Figure 5 reports UUAS for the baseline and fixed-window models trained on the 10M corpus, plot-

ted as a function of layer depth. Several patterns are worth noting. First, the fixed-window model achieves consistently higher UUAS than the baseline across all five relations, with the gap emerging in the early-to-middle layers and persisting across the remainder of the network. The advantage is most pronounced for `nsubj` and `dobj`, where the fixed-window model reaches UUAS values approximately 0.05 above the baseline at several intermediate layers. These two relations capture the core argument structure of clauses, which is precisely the domain in which the fixed-window model also outperforms the baseline most dramatically on BLiMP. The convergence of the behavioural and probing evidence is notable, and suggests that the BLiMP gains observed for the fixed-window model are underwritten by genuinely better structural representations rather than by surface-level heuristics.

Second, both models converge to similar UUAS scores at the deepest layers for several relations, particularly `prep` and `root`. This is consistent with the general finding in structural-probing literature that the latest layers of language models tend to be optimised for the prediction objective rather than for representational explicitness, and that intermediate layers often carry more directly recoverable structural information (Hewitt and Manning, 2019; Tenney et al., 2019). The fixed-window model's advantage is therefore most visible precisely where syntactic structure is most explicitly available, namely the early-to-middle layers.

Third, the fixed-window model develops its structural representations earlier in the network. For `nsubj`, `dobj`, and `prep`, the fixed-window model's UUAS at layer 2 is comparable to the baseline's UUAS at layer 4 or 5. This earlier emergence of structure is consistent with our attention-distribution analysis (Section 5.4), which suggests that the fixed-window model's heads specialise for syntactic functions already in the early layers. Together, these results suggest that the architectural constraint pushes the model to discover syntactic abstractions sooner in the processing pipeline, freeing later layers for higher-order operations.

B.2 Relation to the Attention-Head Analysis

The probing results provide a quantitative complement to the qualitative attention-head analysis in Section 5.4. Where the attention-head visualisations show that individual heads in the fixed-window model specialise for identifiable linguistic functions, the probing results show that this specialisation translates into representational geometry in which syntactic relations are linearly recoverable. The two analyses converge on the same underlying claim that the fixed-window constraint shapes both the local attention patterns within heads and the global geometry of the resulting representational space.

We however caution that structural probing remains an indirect measure. A high UUAS shows that syntactic structure is recoverable from the embeddings, but does not establish that the model uses this structure in any computationally meaningful way during prediction. The convergence between probing results and behavioural performance on BLiMP is suggestive. We treat the probing analysis as one of several converging lines of evidence rather than as a standalone proof of structural representation.

C Detailed BLiMP Analysis by Phenomenon

The main text reports BLiMP performance aggregated by linguistic category (Section 5.2). Here we provide a finer-grained analysis at the level of individual sub-tasks, drawing on the per-phenomenon heatmap shown in Figure 6. The analysis surfaces several patterns that do not emerge clearly at the category level and that bear on the interpretation of where architectural constraints help, where they do not, and why.

The largest constrained-versus-baseline gaps appear on phenomena that combine two properties: a syntactically structured relationship between elements that are not always linearly adjacent, and a heavy dependence on training data for the acquisition of that relationship. Argument structure phenomena highlight this pattern. Sub-tasks such as `animate_subject_passive`, `animate_subject_trans`, and `causative` test whether the model has internalised the selectional restrictions verbs impose on their arguments. The fixed-window models with $W = 5$ achieve accuracies on these sub-tasks at the 10M scale that often match or exceed the corresponding 100M baseline, suggesting that locality is doing genuine work in scaffolding verb-argument acquisition rather than merely capturing surface co-occurrence. This pattern is consistent with the argument structure of English clauses being predominantly local, with the verb's arguments typically appearing within a few tokens of the verb itself.

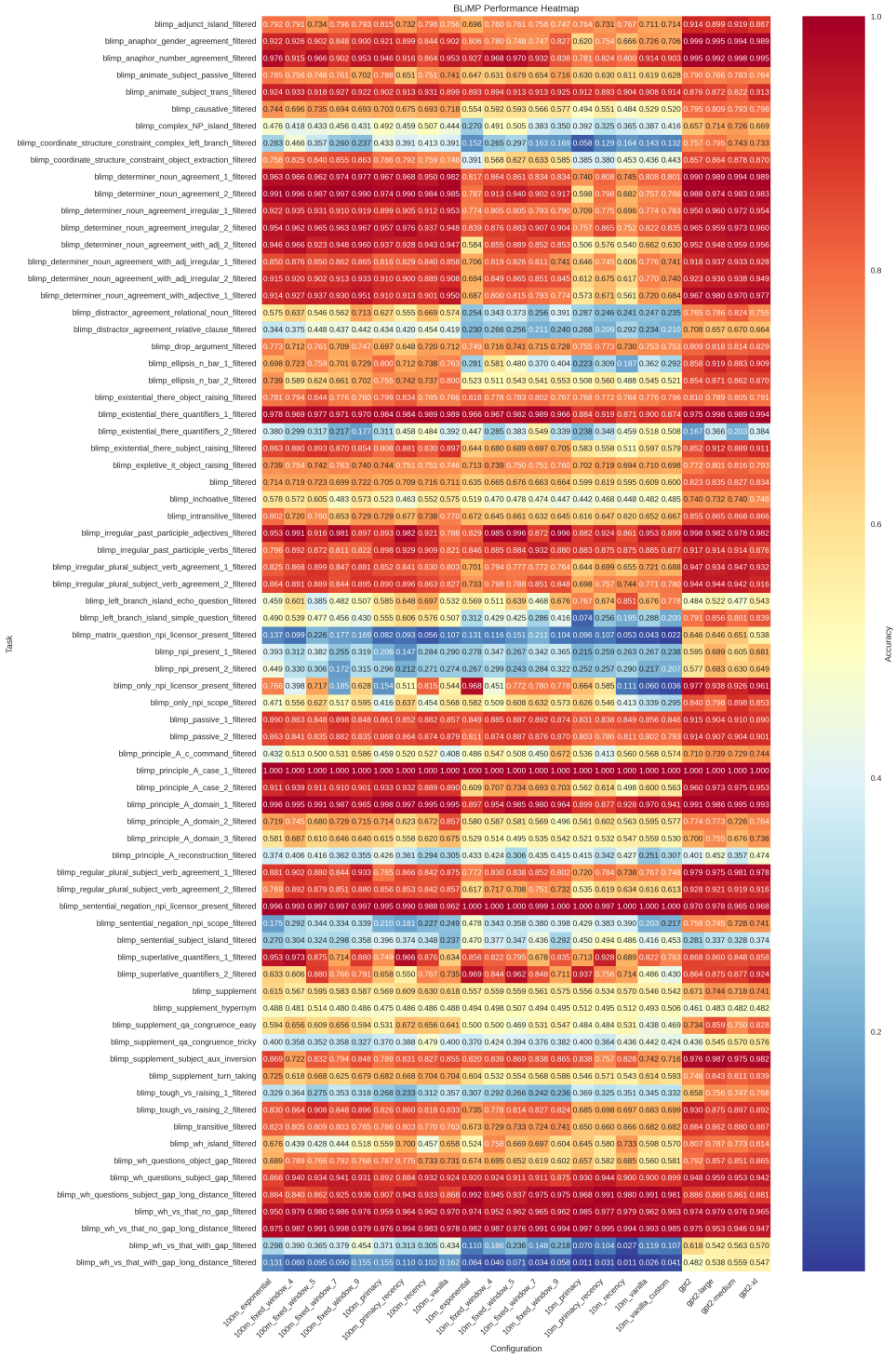


Figure 6: Model performance on BLiMP across individual tasks and models.

Binding phenomena show a parallel pattern. Sub-tasks such as `principle_A_c_command` and `principle_A_case` test sensitivity to the structural constraints governing the relationship between reflexive pronouns and their antecedents. Constrained models, particularly fixed windows of size 5, perform competitively on these tasks despite the fact that binding relations are not always strictly local. This is theoretically interesting: the principle of binding is a structural constraint over c-command relations rather than over linear distance, and one might expect that models with restricted linear access would struggle. The empirical pattern suggests either that the relevant binding relations in BLiMP minimal pairs typically fall within the model's window, or that the representational struc-

ture encouraged by the locality constraint generalises usefully to capturing structural rather than purely linear dependencies. Distinguishing these two possibilities empirically would require a controlled analysis on minimal pairs stratified by the linear distance between binder and bindee, which we leave to future work.

Two categories show consistently weak performance across all models, including the constrained variants: Island Effects and NPI Licensing. The performance ceiling on these categories is substantially below the model's overall accuracy, and the constrained models offer no clear advantage over the baseline. We take this as informative rather than disappointing.

Island Effects test the model's sensitivity to constraints on syntactic movement, such as the wh-island and complex NP-island constraints. These constraints are not reducible to linear distance or to local agreement; they require sensitivity to abstract syntactic configurations that hold over arbitrarily extensive structural domains. A model whose architectural constraint is defined in terms of linear distance, as ours are, has no principled mechanism for representing such configurations. The persistence of poor performance across all our locality-based mechanisms therefore aligns with theoretical expectation: linear locality is the wrong abstraction for capturing island constraints. A more interesting open question, raised in the Limitations section, is whether cognitive constraints have natural analogues that operate over abstract structural distance, which might capture island phenomena where linear constraints cannot.

NPI Licensing tests the model's sensitivity to the requirement that negative polarity items such as *ever* or *any* appear in the scope of a licensing operator such as negation or a question. The constraint is logical and scope-based rather than structural in the syntactic sense, and the relevant licensing operator may be arbitrarily distant from the NPI in linear terms. As with island constraints, a locality-based constraint is unlikely to offer a principled solution here, and the empirical results bear this out. Performance on `npi_present_1` and related sub-tasks remains close to chance for most models at both training scales.

The heatmap also reveals systematic differences among the fixed-window variants themselves. Window size 5 emerges as the most consistent performer across phenomena, achieving the highest overall accuracy at 10M and remaining competitive at 100M. Windows of size 7 and 9 perform comparably to size 5 on most sub-tasks but show occasional degradation on phenomena requiring particularly local sensitivity, such as some morphological agreement tasks. Window size 4 exhibits the opposite pattern, performing well on local phenomena but showing larger degradations on sub-tasks involving moderately non-local dependencies. The convergence on $W = 5$ as the best-performing choice is itself empirically interesting and may reflect something about the typical phrasal length over which English syntactic dependencies operate, although we emphasise that this observation is suggestive rather than conclusive without controlled cross-linguistic replication.

The exponential and logistic decay variants show a different profile. They perform competitively with the fixed-window models on phenomena requiring strict locality but underperform on tasks where information from slightly more distant tokens is informative. We attribute this to the soft nature of the decay constraint: while the bias favours nearby tokens, attention to more distant tokens is not strictly prevented, and the model appears to retain some capacity to attend non-locally when the training signal calls for it. Whether this represents a genuine weakness of soft decay relative to hard windowing, or whether it reflects inadequate hyperparameter tuning of the decay mechanisms, is a question we cannot fully resolve here.

D AI Assistant Use

We made use of generative AI tools to assist in the drafting of the manuscript and the refinement of the research code. These tools were used only for linguistic clarity and code scaffolding (VS Code). All technical content and final interpretations were rigorously reviewed and verified by us.