Emergence of Episodic Memory in Transformers: Characterizing Changes in Temporal Structure of Attention Scores During Training

Deven Mahesh Mistry, Anooshka Bajaj, Yash Aggarwal,
Sahaj Singh Maini, Zoran Tiganj
Department of Computer Science
Indiana University Bloomington

Abstract

We investigate in-context temporal biases in attention heads and transformer outputs. Using cognitive science methodologies, we analyze attention scores and outputs of the GPT-2 models of varying sizes. Across attention heads, we observe effects characteristic of human episodic memory, including temporal contiguity, primacy and recency. Transformer outputs demonstrate a tendency toward in-context serial recall. Importantly, this effect is eliminated after the ablation of the induction heads, which are the driving force behind the contiguity effect. Our findings offer insights into how transformers organize information temporally during in-context learning, shedding light on their similarities and differences with human memory and learning.

1 Introduction

Large language models (LLMs) have demonstrated a remarkable capacity for in-context learning. They are capable of adapting to new tasks using examples provided within the input prompt, without any parameter updates (Brown, 2020). The temporal position of tokens plays an important role in in-context learning. For instance, simply asking the model to repeat a sequence of words from the prompt requires the model to use the information about the temporal position of tokens. This resembles human learning, where the temporal organization of memory plays a critical role in recalling specific past episodes.

Previous work has demonstrated that some attention heads show a temporal induction property. These *induction heads* search the input prompt for the prior occurrence of the current token. If a match is found, they attend to the token that followed the previous presentation of the current token. This mechanism allows induction heads to effectively learn and reproduce sequences of tokens and it has been argued to contribute to the model's ability to perform tasks based on contextual information (Olsson et al., 2022; Elhage et al., 2021; Singh et al., 2024; Ji-An et al., 2024; Pink et al., 2024).

The induction property is related to human episodic memory. Numerous studies have demonstrated that episodic memory exhibits contiguity effect, where items or events that occur close together in time are more likely to be remembered together in memory recall (Kahana, 1996; Howard and Kahana, 2002; Lohnas and Healy, 2021; Polyn et al., 2009; Jenkins and Ranganath, 2010). For example, if a person experiences a sequence of events within a short time span, they are more likely to recall these events together than if they were spread out over a longer period. This effect was systematically studied through free recall, a task where participants are presented with a list of items (e.g., words, pictures) and then asked to recall them in any order they choose. Participants tend to recall items that were presented close together in the original list in clusters, indicating that the temporal proximity during encoding influences the retrieval process. This supports the idea that our memories are organized not just by the content of a stimulus but also by its temporal context (Howard and Kahana, 2002; Polyn et al., 2009). Neuroscience studies have demonstrated that neural activity in the brain during recall reinstates the neural activity observed during encoding, indicating the retrieval of temporal context or a mental "jump back in time" (Howard et al., 2012; Folkerts et al., 2018). The contiguity effect has been recently studied in deep neural networks, including recurrent neural networks (Li et al., 2024) and pretrained transformer models (Ji-An et al., 2024) where it was found that pretrained transformers contain attention heads that recover temporal context in a manner consistent with human episodic memory. Furthermore, human-like episodic memory is found to improve performance in tasks that require processing over extended temporal context (Fountas et al., 2024).

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 8893–8911
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

In addition to temporal induction, transformers exhibit *serial position effects*, specifically *recency* and *primacy* (Janik, 2023; Peysakhovich and Lerer, 2023; Wang et al., 2023; Guo and Vosoughi, 2024; Angne et al., 2023). Recency implies that tokens that are more recent in the prompt (closer to the present token) are going to be more important in generating the subsequent token. Conversely, primacy implies the same property but for the tokens that are presented at the beginning of the prompt. The emergence of these effects in transformers can be shaped by positional encoding (Janik, 2023; Peysakhovich and Lerer, 2023). Positional encoding embeddings can be learned (Devlin et al., 2019) or fixed to vectors that convey positional information, such as Rotformer, which uses rotary positional embeddings (Su et al., 2024) or early transformer models that used sinusoidal embeddings (Vaswani, 2017). These positional embeddings capture translation invariance, monotonicity and symmetry through distance metrics in their vector embeddings (Wang et al., 2021).

Recency and primacy are also characteristic of human memory, where we often exhibit better recall for items presented at the beginning (primacy effect) and end (recency effect) of a list (Atkinson and Shiffrin, 1968; Glanzer and Cunitz, 1966; Murdock, 1962). This parallel suggests that LLMs, like humans, may be sensitive to the temporal context of information, with recent and initial tokens playing a more significant role in shaping the model's internal representations and influencing its output.

Here we investigate the temporal aspects of attention heads and transfromer outputs during learning. We train two transformer models of different sizes (GPT-2 small and GPT-2 medium) on three datasets: Wikitext-103 (Merity et al., 2016) and two sampled datasets of FineWeb (HuggingFaceFW, 2024) (with 1B tokens and 10B tokens). We adopt tools used in cognitive science to characterize temporal aspects of human memory and use them to examine attention heads and outputs of transformer models. Specifically, we use Lag-Conditional Recall Probability (Lag-CRP) analysis which measures the probability of recalling an item a certain number of positions away (called *lag*) from the previously recalled item. We apply this analysis to transformer outputs and attention heads to understand how temporal relationships shape attention scores and token predictions. We characterize the impact of trainable positional encoding, model size, and number of training interactions on the emergence of contiguity effect and serial position effects. We also ablate induction heads to examine their role in shaping temporal properties of transformer outputs. Our findings provide novel insights into similarities and differences between human memory and transformers, improving our understanding of in-context learning.

2 Methods

2.1 Models and training

We used models based on the GPT-2 small and GPT-2 medium architectures (Radford et al., 2019). GPT-2 small has approximately 124 million parameters, consisting of 12 attention heads, 12 transformer layers, an embedding size of 768 dimensions, and an MLP with 3,072 neurons. GPT-2 medium has approximately 353 million parameters and consists of 16 attention heads, 24 transformer layers, an embedding size of 1024 dimensions and an MLP with 4096 neurons. The vocabulary consists of 50,257 tokens and we used the GPT-2 byte-pair encoder.

We trained GPT-2 small and medium on Wikitext-103 (Merity et al., 2016) for 4000 iterations. The dataset contains 117M tokens in the training set, 0.24M tokens in the validation set and 0.28M tokens in the test set. We also trained GPT-2 small on two larger datasets sampled from FineWeb (HuggingFaceFW, 2024) that included 1B tokens and 10B tokens. 1B dataset had 996M tokens in the training set and 4M in the validation set and 10B dataset had 10.3B tokens in the training set and 51.4M in the validation set. In our experiments, we used the *nanoGPT* codebase (Karpathy, 2022). We trained all the configurations of GPT-2 small and GPT-2 medium with the same hyperparameters, including learning rate, number of warm-up iterations, and weight decay. During training, all models had a maximum learning rate of 10^{-4} and a learning rate warm-up period of 450 iterations. The training was done on four 40GB A100 GPUs.

2.2 Calculation of lag-CRP curves for attention heads

To compute lag-CRP curves for attention heads, we prompt the models with a 1000 tokens long prompt composed of source and destination sequences (the maximum length of the prompt for both GPT-2 small and GPT-2 medium is 1024 tokens). The source sequence consisted of 500 most frequent tokens in a given dataset. The tokens were presented in a random order. The source sequence was followed by 500 tokens long destination sequence. The tokens in the destination sequence were the exact repetition of the source sequence. We computed the attention scores between each token in the destination sequence and the source sequence to calculate the lag-CRP curve. Given a single sequence of tokens, the lag-CRP score for lag at position zero was calculated as the average attention score between the same tokens in the source and destination sequences. Similarly, the lag-CRP score for lag l was calculated as the average attention score between the tokens in the destination sequence, and a different token placed l positions away from the corresponding token in the source sequence. When l is positive, the lag-CRP score is calculated for the tokens following the corresponding token in the source sequence, and when l is negative, the score is calculated for the tokens that occur before the corresponding token. To reduce the impact of semantic similarity, we averaged the lag-CRP scores for a given lag across ten randomly permuted sequences. We produce the lag-CRP curves for all heads. Mathematically, the lag-CRP score for an individual head is computed as follows (Ji-An et al., 2024):

$$S_l = \sum_{i \in M} \frac{1}{N - |l| * 2} \sum_{|l| < s \leq N - |l|} a_{s+N, s+l}. \quad (1)$$

Here S_l refers to the score for lag l, N is the size of the source (or) destination sequence, M is the set of example sequences generated by permuting the source tokens along with similarly permuted destination tokens. $a_{i,j}$ is the attention score calculated between the token at position i in the destination sequence and position j in source sequence. In all our experiments, M contains 10 sequences. We note that since we used attention scores to compute the lag-CRP curve, it is no longer restricted to response probabilities, therefore it can take any range of values.

2.3 Calculation of induction matching score

Following previous work on in-context learning (Olsson et al., 2022; Elhage et al., 2021) we computed induction scores for each attention head. The induction score expresses the degree to which the head attends to the token following the previous occurrence of the current token in the sequence.

Given a sequence of tokens to compute the induction matching score for an attention head, we first extract the attention pattern from the model for the corresponding layer and head. This attention pattern provides the weights indicating how much each token in the sequence attends to every other token. We then construct a target matrix that records matches based on the induction rule: if the token at a destination position matches the token before a particular source position, the corresponding entry in the target matrix is set to 1. Next, we compute the element-wise product of the attention pattern and the target matrix to isolate attention values corresponding to induction matches. The numerator is the sum of these matched attention values, while the denominator is the sum of all attention values between the source and destination positions. The induction matching score is obtained by dividing the numerator by the denominator.

More formally, let N be the sequence length, $a_{i,j}$ denote the attention value from token i (destination position) to token j (source position), and $t_{i,j}$ be the target matrix entry, where $t_{i,j} = 1$ if the token at position i matches the token before position j, and $t_{i,j} = 0$ otherwise. The induction matching score is given by:

$$\mathrm{I} = \frac{\sum_{(i,j)} a_{i,j} \cdot t_{i,j}}{\sum_{(i,j)} a_{i,j}}, \quad (2)$$

where the summation $\sum_{(i,j)}$ is taken over all valid pairs of (i, j) within the sequence.

2.4 Computing the temporal extent of the contiguity effect and strength of recency effect

While the induction matching score quantifies the tendency to attend to the token that follows the previous occurrence of the current token, human episodic memory is characterized by a lag-CRP curve that has a strong contiguity effect. This implies a gradual falloff of the lag-CRP curve as a function of positive and negative lags. In our experiments, we choose a subset of heads that have the highest lag-CRP score at $l = 1$ when the lag-CRP curve is computed between $l = -10$ to $l = 10$. This enabled the identification of temporal contiguity even in the presence of high recency or primacy (i.e., high values of the lag-CRP curve for large values of l).

To quantify the recency effect, we computed the lag-CRP curves for the chosen heads. To isolate the recency effect and to ensure that the contiguity effect does not superimpose with the recency effect, we remove the lag-CRP scores between the lags -50 and +50. We then fit a linear regression model to the rest of the lag-CRP scores and compute the average slope of the linear fits across selected attention heads. This expresses the recency bias in the attention heads.

In our experiments, while exploring the contiguity effect in the selected attention heads, we observed a gradual falloff of lag-CRP scores as a function of positive but not negative lags. To quantify this falloff, we first subtract the linear fit from the lag-CRP scores in order to remove the recency effect. We then fit an exponential function to the positive lags of the lag-CRP curve (exponential fit is commonly applied in human episodic memory models (Howard and Kahana, 2002; Polyn et al., 2009)). We used Levenberg-Marquardt algorithm (Marquardt, 1963; Levenberg, 1944) and optimized the following function: $ae^{-t/\tau}$ where a and τ (time constant) are parameters and t is the lag (Tab. 2).

2.5 Positional encoding with different magnitudes

We trained five variants of the GPT-2 small model with five different magnitudes of positional encoding: 0, 0.5, 1, 1.5, and 2. These factors multiplied positional embeddings before positional embeddings were added to word embeddings. Positional embeddings were learnable, and they used the same weight initialization.

3 Results

We trained GPT-2 small and GPT-2 medium transformer models. On the Wikitext-103 dataset GPT-2 small converged after around 4000 iterations, and the larger GPT-2 medium model converged after around 2000 iterations. The convergence was determined by monitoring the validation loss. The models converged to perplexities similar to those of models with comparable size, including Transformer-XL Standard (Dai et al., 2019) LaMemo (Ji et al., 2022), Hybrid H3 (Fu et al., 2023) and TrimeLM Long (Zhong et al., 2022), all of which have perplexity above 20 (Tab. 1). Importantly, regardless of the magnitude of the positional encoding, the models converged to similar values of perplexity, consistent with results in Haviv et al. (2022). We also trained GPT-2 small on FineWeb-1B and 10B. Both models converged after around 10000 iterations.

After training, most attention heads exhibited some form of structured temporal modulation–recency, primacy, contiguity, or a combination of these effects. For Wikitext-103, heads in layers closer to the input were characterized by recency and primacy effects, while heads in layers closer to the output had strong recency and contiguity effects. Attention scores of GPT-2 small as a function of lag across all heads before and after training are shown in Fig. A3 and Fig. 1 respectively. Attention scores across all heads after training of GPT-2 medium model are shown in Fig. A4. Fig. 2 shows attention scores as a function of lags for two representative attention heads from layers closer to the output. The top row (Fig. 2A-B) illustrates the raw values, while the middle and bottom rows (Fig. 2C-F) show results after the recency effect has been removed, highlighting the contiguity effect. For FineWeb-1B and 10B attention scores across all heads after training are shown in Fig. A8 and Fig. A9 respectively. While similar to the results on Wikitext-103, they often showed more complex, non-linear patterns across a wide range of lags.

3.1 Temporal properties of attention heads emerge gradually throughout training

We observed the increase in the average induction score (averaged across all heads identified as induction heads as defined in the Methods section) and the number of induction heads throughout training (Tab. 2 for Wikitext-103, Tab. 3 for FineWeb-1B and Tab. 4 for FineWeb-10B). For Wikitext-103 we also computed the average recency slope the average time constant fitted to the lag-CRP curve. (These were not computed for FineWeb-1B and FineWeb-10B since their lag-CRP curves for induction heads commonly had non-linear temporal profiles across a wide range of tested lags, making it more difficult to isolate and characterize the recency effect.) For Wikitext-103, the average recency slope increased and plateaued around iteration 3000. Prior to training (iteration 0), the average recency slope was also 0. This is because we used trained positional embeddings so prior to training the positional embedding vectors were random. The average time constant fitted to the lag-CRP curve (attention scores as a function of lag) did not systematically change as training progressed but rather converged to a relatively low value around three tokens (note that *lag* is in the units of tokens) after around 3000 iterations.

To better understand the emergence of induction heads, we visualized the induction scores at

Table 1: Lowest perplexity values for GPT-2 small and GPT-2 medium for different positional encoding magnitudes after training on Wikitext-103 dataset.

	GPT-2 small					GPT-2 medium
Positional encoding magnitude	0	0.5	1	1.5	2	1
Perplexity	19.8	19.8	19.8	19.7	19.8	19.5

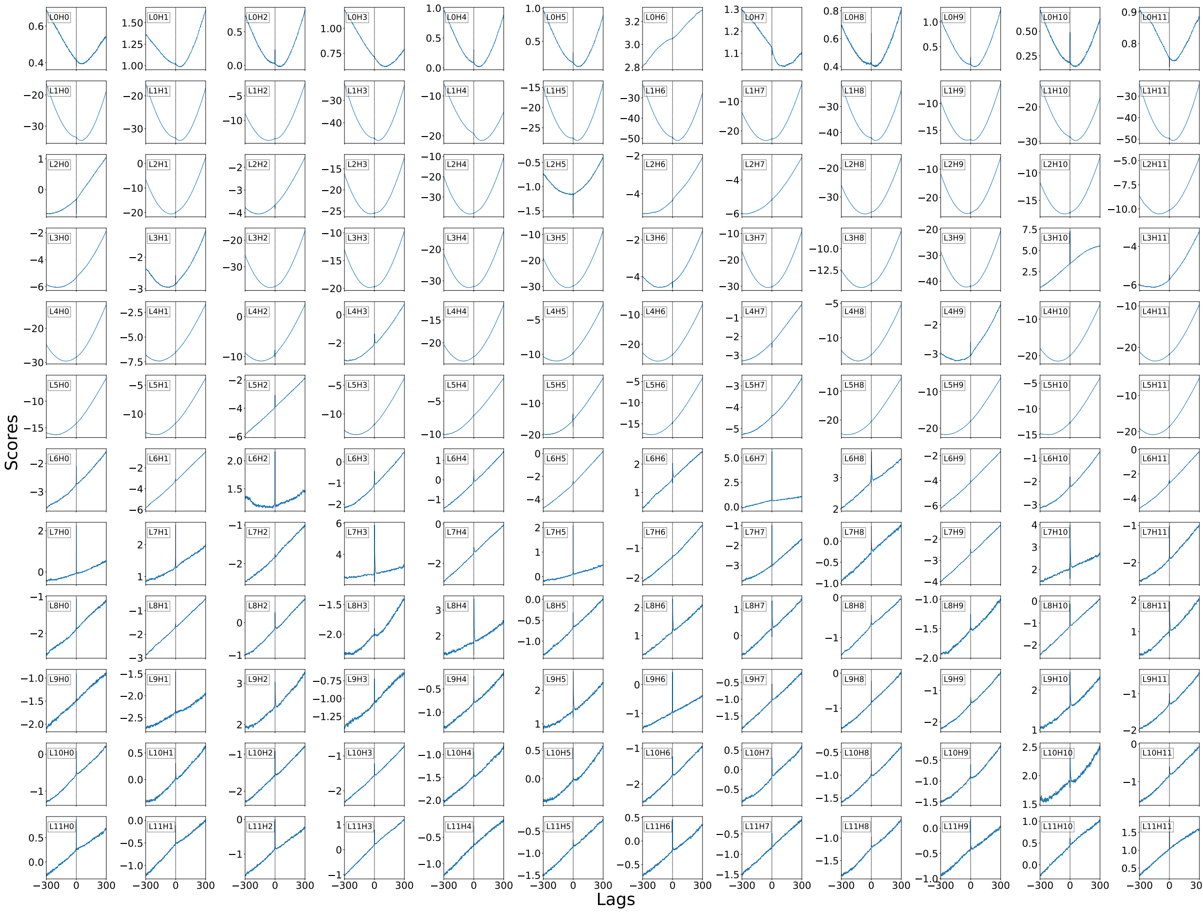

Figure 1: Attention scores as a function of lag for all attention heads in GPT-2 small after 4000 iterations on the Wikitext-103 dataset (baseline positional encoding).

Table 2: Induction head properties as a function of training iteration for the Wikitext-103 dataset.

	Training Iteration						
Metric	0	100	500	1000	2000	3000	4000
Average Induction Score	0.0007	0.0007	0.0006	0.0012	0.0019	0.0015	0.0024
Average Time Constant	0.7	0.5	6.43	5	6.4	3.2	3.1
Average Recency Slope	0	0	0.0001	0.0015	0.0018	0.0032	0.0029
Number of Induction Heads	4	4	3	19	16	19	20

all heads at five different training steps on GPT-2 small for Wikitext-103 (Fig. 4). Early in training, induction scores are very small in magnitude, and they increase gradually, showing presence mainly in layers six to nine. The locations and number of induction heads did not change during training, showing gradual shaping of the temporal properties. In Fig. 3, we showed attention scores as a function of lag for the same head at two different stages of training – note the order of magnitude change on

Table 3: Induction head properties as a function of training iteration for FineWeb-1B dataset.

	Training Iteration						
Metric	0	50	100	500	1000	5000	10000
Average Induction Score	0.0007	0.0007	0.0007	0.0002	0.008	0.011	0.016
Number of Induction Heads	4	4	1	3	25	36	25

Table 4: Induction head properties as a function of training iteration for FineWeb-10B dataset.

	Training Iteration						
Metric	0	50	100	500	1000	5000	10000
Average Induction Score	0.0007	0.0007	0.0007	0.0003	0.03	0.05	0.05
Number of Induction Heads	4	7	3	4	24	29	39

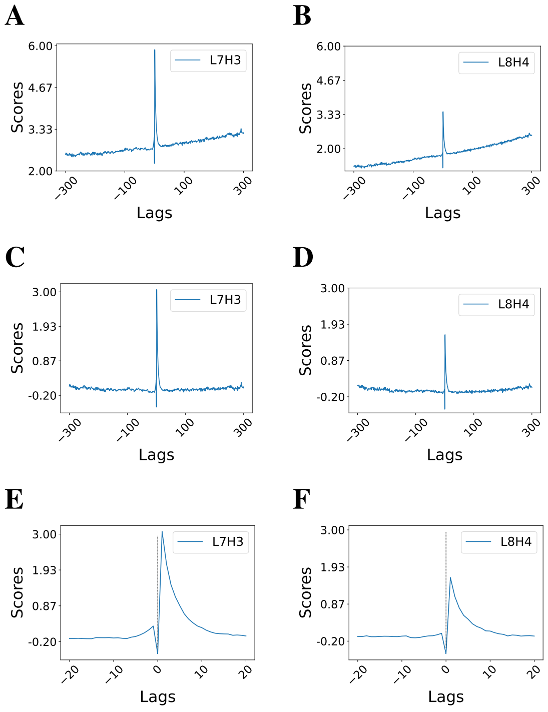

Figure 2: Two induction heads before (top row) and after (middle row) adjusting for recency effect. The bottom row shows zoomed-in version of the middle row.

the y-axis from iteration 300 (Fig. 3A) to iteration 4000 (Fig. 3B). Fig A3 shows attention scores of all heads before training, Fig A7 after 1000 iterations and Fig 1 after 4000 iterations. These figures further illustrate the gradual emergence of temporal profiles, including both induction and recency.

3.2 Positional encoding magnitude shapes recency and contiguity effects

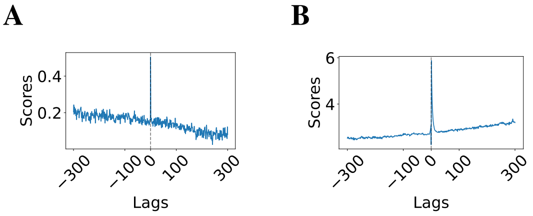

Figure 3: Example of the same induction head L7H3 during different stages of training. **A.** After 300 iterations. **B.** After 4000 iterations.

Increasing the magnitude of positional encoding increased the average induction score (Tab. 5). This is consistent with the hypothesis that vector similarity induced by positional encoding creates a temporal link needed for the induction heads to identify the token that followed the current token in the previous sequence. To further test this hypothesis, we plotted the Pearson correlation coefficient between the positional embeddings for different magnitudes of positional encoding and different training iterations (Fig. 5). The plot reveals an interesting trade-off in the magnitude of positional encoding and training iteration: for example, the correlation profile after 1000 iterations and a magnitude of 0.5 is similar to the correlation profile after 2000 iterations and a magnitude of 1. This relationship is visible across the diagonals of Fig. 5, except for the magnitude of positional encoding equal to 2. For a magnitude of 2, we see more complex temporal patterns that include oscillatory dynamics in the amount of temporal correlations. Overall, the profiles indicate an increase in temporal correlations with training and magnitude of positional encoding, supporting the observed increase in the average induction score shown in Tab. 5. The number of induction heads decreased with the mag-

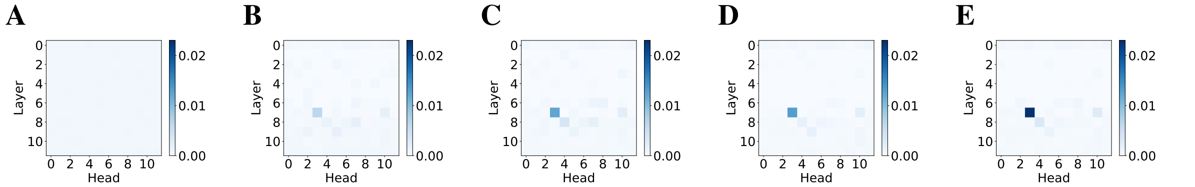

Figure 4: Induction scores for five checkpoints throughout the training of GPT-2 small on Wikitext-103 dataset. **A.** Random initialization. **B.** 1000 iterations. **C.** 2000 iterations. **D.** 3000 iterations. **E.** 4000 iterations.

Table 5: Impact of positional encoding on temporal properties of attention heads (GPT-2 small trained on Wikitext-103 dataset).

	Positional encoding magnitude				
Metric	0	0.5	1	1.5	2
Average Induction Score	0.0010	0.0006	0.0024	0.0038	0.0055
Average Time Constant	29.1	2.0	3.1	3.8	34.1
Average Recency Slope	0	0.0034	0.0029	0.0045	0.0053
Number of induction heads	6	25	20	15	13

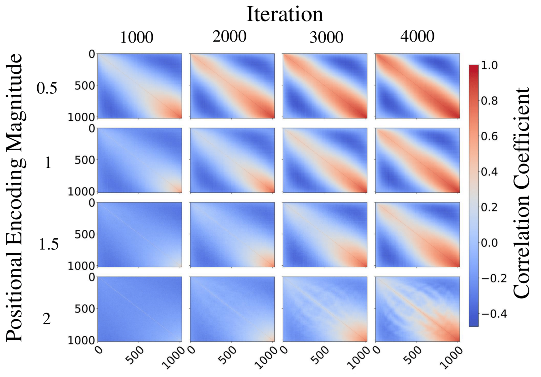

Figure 5: Correlation in positional encoding vectors scales with training iterations and positional encoding magnitude during training of GPT-2 small on Wikitext-103 dataset.

nitude of positional encoding (Tab. 5). This is also visible in Fig. 7, where the scores of some induction heads increase, while for others, the scores decrease, so they no longer fit the criteria set for induction heads.

The impact of the magnitude of the positional encoding on the average time constant was mixed. Some heads showed long time constants, including heads in models without positional encoding and in models with double the amount of positional encoding. However, heads with a magnitude of positional encoding equal to 0.5, 1 and 1.5 all had relatively short time constants in the range of 2 to 4 lags. Thus models with these balanced magnitudes of positional encoding did not retrieve extended temporal context. Fig 6 shows scores of a single attention head for different magnitudes of positional encoding illustrating relatively short time constants. The average recency slope increased with the magnitude of the positional encoding, as expected, due to increased temporal similarity induced by the positional encoding.

Tab. 5 indicates that with no positional encoding, the slope at the six induction heads was 0. A closer look at all of the attention heads with no positional encoding (Fig. A5) reveals several heads that exhibited weak recency effect (note that the range of magnitudes of the heads that show the recency effect was typically small without positional encoding). Previous work has argued that causal masking could have similar effects as positional encoding because it introduces sequential dependencies (Haviv et al., 2022). These dependencies can result in the encoding of the input order and could explain the weak recency effect.

3.3 Temporal effects are consistent across models of different sizes

All previous results were discussed for GPT-2 small model. We also trained GPT-2 medium but only for the baseline magnitude of positional encoding and on Wikitext-103. Overall, we observed similar attention score profiles across the two models (compare Fig. 1 and Fig. A4). We quantified these observations in Tab. 6. Fig 8 shows induction scores in the two models, indicating that GPT-2 medium

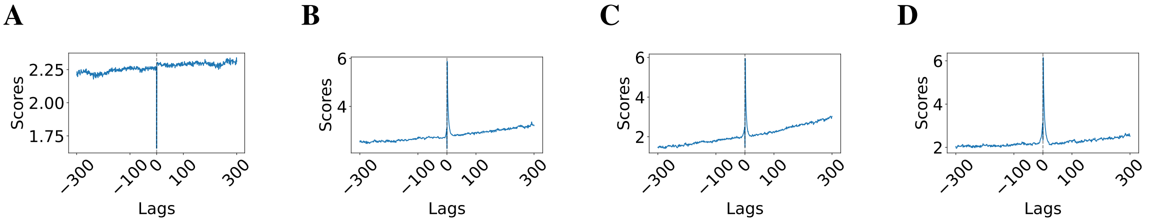

Figure 6: Example of the same induction head L7H3 for different magnitudes of positional encoding. **A.** No positional encoding. **B.** Positional encoding with magnitude 1 (baseline model). **C.** Positional encoding with magnitude 1.5. **D.** Positional encoding with magnitude 2.

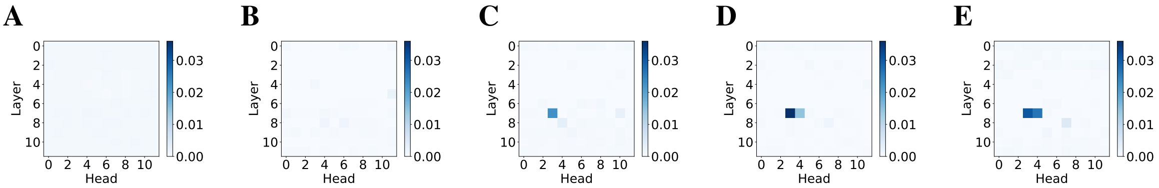

Figure 7: Induction scores for five different magnitudes of positional encoding (GPT-2 small trained on Wikitext-103 dataset). **A.** No positional encoding, **B.** 0.5, **C.** 1, **D.** 1.5, **E.** 2.

Table 6: Impact of model size on temporal properties of attention heads.

	Model type	
Metric	GPT-2 small	GPT-2 medium
Average Induction Score	0.0024	0.0012
Average Time Constant	3.1	12.1
Average Recency Slope	0.003	0.007
Number of Induction Heads	20 (out of 144, 14%)	45 (out of 384, 12%)

had larger scores concentrated closer to the input layer than GPT-2 small.

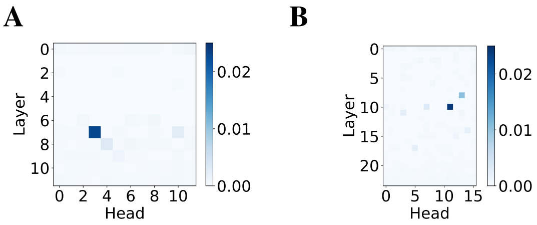

Figure 8: Induction scores for two different models. **A.** GPT-2 small, **B.** GPT-2 medium.

3.4 Characterizing the contiguity effect across the attention heads

To better understand the span of temporal context retrieval in transformers, we investigated the distribution of the time constants from fitting the attention scores as a function of lag. We found that the time constants are mainly concentrated in the narrow range of 2-4 lags, with only a few heads covering larger lags. This result holds for models with different magnitudes of positional encoding and for both model sizes. This suggests that when retrieving in-context information given a repeated token,

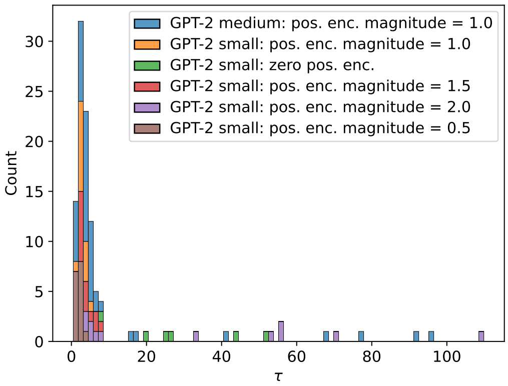

Figure 9: Distributions of fitted time constants of induction heads for different models and magnitudes of positional encoding.

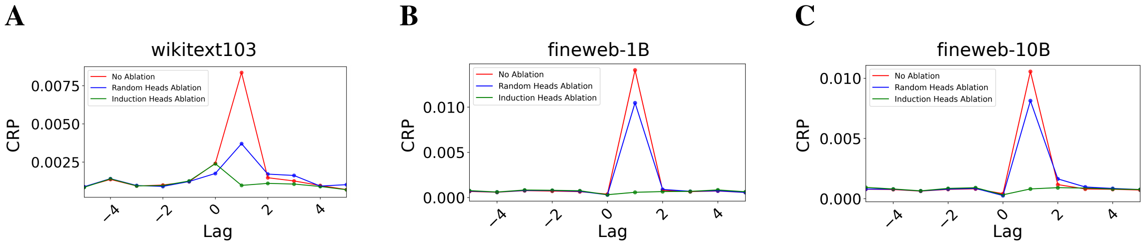

Figure 10: CRP during downstream evaluation showing impact of induction head ablation.

aside from primacy and recency effects, transformers will primarily focus attention on the very local (2-4 lags) neighborhood of that token.

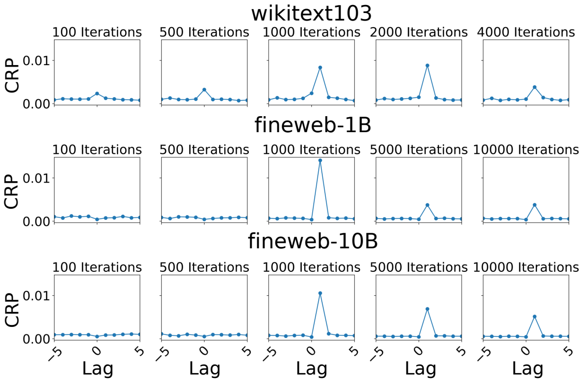

Figure 11: CRP as a function of training iteration in downstream evaluation.

3.5 Downstream evaluation and impact of induction heads ablation

To better understand the impact of temporal context retrieval in attention heads on transformer outputs, we conducted a downstream evaluation inspired by the free recall memory task. After training the models, we probed them with a sequence of 500 randomly ordered tokens (we selected 500 tokens that were most frequently occurring in each dataset) followed by a middle token (e.g., *GRDBTHMB*, where each character corresponds to an individual token). We then quantified the probability of the next token as a function of lag (distance from the middle token). Temporal contiguity predicts a larger probability for tokens that are temporally adjacent to the middle token, while recency and primacy effects predict a higher probability for tokens from the beginning (large negative lag) and end (large positive lag) of the list.

During the training (after about 1000 iterations), we observed a strong (around 10 times) increase in the probability of recall for items at lag +1 (Fig. 11, Fig. A1), indicating strong preference for serial recall. To investigate the relationship between induction heads and this effect, we ablated the heads that had induction scores above 0.01 (the ablation was done similarly to Crosbie and Shutova (2024) by setting the attention scores for ablated heads to $-\infty$). Even though the number of ablated heads was around 5% of the total number of heads, the ablation of induction heads eliminated the contiguity effect (Fig. 10, Fig. A2). Ablating the same number of non-induction heads in a layer-matched fashion made much smaller impact on the output probabilities, especially for FineWeb-1B and 10B.

4 Discussion

We quantified temporal properties of attention patterns in transformer outputs using lag-CRP analysis, commonly used for studying episodic memory and serial position effects in human memory experiments. By using multiple permutations of the input sequences, we were able to reduce the semantic effects of token similarity and isolate the temporal effects making it possible to observe primacy, recency and contiguity effects in the attention heads.

Unlike human memory experiments, where the contiguity effect is robust across a wide range of scales (Howard et al., 2008), supporting power-law decay of memory (Wixted and Ebbesen, 1991; Rubin and Wenzel, 1996; Donkin and Nosofsky, 2012), we did not find evidence for retrieval of a broad temporal context in transformers. In fact, training typically had an impact of reducing the time constants of lag-CRP to small values in the range of 2 to 4 lags. Downstream analysis demonstrated a strong preference towards a serial recall that was eliminated after ablation of the induction heads. Overall, we showed that tools from cognitive science can be used to better understand learning in transformers, providing valuable insights into the emergence of temporal structure during in-context learning.

5 Limitations

Our approach used relatively small models. Training larger, instruction fine-tuned language models could provide additional insights into the temporal properties of in-context learning. While our approach was inspired by studies of episodic memory in humans, a number of methodological differences between the present analysis and human experiments (such as the fact that humans receive task instructions) prevent us from making direct parallels with human memory and learning.

Acknowledgment

This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

References

Hemali Angne, Charlotte Cornell, and Qiong Zhang. 2023. Why two heads together are worse than apart: A context-based account of collaborative inhibition in memory search. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Richard C. Atkinson and Richard M. Shiffrin. 1968. Human memory: A proposed system and its control processes. 2:89–195.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Joy Crosbie and Ekaterina Shutova. 2024. Induction heads as an essential mechanism for pattern matching in in-context learning. *arXiv preprint arXiv:2407.07011*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.

Chris Donkin and Robert M Nosofsky. 2012. A power-law model of psychological memory strength in short- and long-term recognition. *Psychological science*, 23(6):625–634.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.

Sarah Folkerts, Ueli Rutishauser, and Marc W Howard. 2018. Human episodic memory retrieval is accompanied by a neural contiguity effect. *Journal of Neuroscience*, 38(17):4200–4211.

Zafeirios Fountas, Martin A Benfeghoul, Adnan Oomerjee, Fenia Christopoulou, Gerasimos Lampouras, Haitham Bou-Ammar, and Jun Wang. 2024. Human-like episodic memory for infinite context llms. *arXiv preprint arXiv:2407.09450*.

Daniel Y Fu, Tri Dao, Khaled Kamal Saab, Armin W Thomas, Atri Rudra, and Christopher Re. 2023. Hungry hungry hippos: Towards language modeling with state space models. In *The Eleventh International Conference on Learning Representations*.

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

Xiaobo Guo and Soroush Vosoughi. 2024. Serial position effects of large language models. *arXiv preprint arXiv:2406.15981*.

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. 2022. Transformer language models without positional encodings still learn positional information. *arXiv preprint arXiv:2203.16634*.

Marc W Howard and Michael J Kahana. 2002. Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(4):923.

Marc W Howard, Indre V Viskontas, Karthik H Shankar, and Itzhak Fried. 2012. Ensembles of human mtl neurons "jump back in time" in response to a repeated stimulus. *Hippocampus*, 22(9):1833–1847.

Marc W Howard, Tess E Youker, and Vijay S Venkatadass. 2008. The persistence of memory: Contiguity effects across hundreds of seconds. *Psychonomic Bulletin & Review*, 15:58–63.

HuggingFaceFW. 2024. fineweb (revision af075be).

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

Lila J Jenkins and Charan Ranganath. 2010. Prefrontal and medial temporal lobe activity at encoding predicts temporal context memory. *Journal of Neuroscience*, 30(45):15596–15603.

Haozhe Ji, Rongsheng Zhang, Zhenyu Yang, Zhipeng Hu, and Minlie Huang. 2022. LaMemo: Language modeling with look-ahead memory. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, pages 5747–5762, Seattle, United States. Association for Computational Linguistics.

Li Ji-An, Corey Y Zhou, Marcus K Benna, and Marcelo G Mattar. 2024. Linking in-context learning in transformers to human episodic memory. *arXiv preprint arXiv:2405.14992*.

Michael J Kahana. 1996. Associative retrieval processes in free recall. *Memory & Cognition*, 24(1):103–109.

Andrej Karpathy. 2022. https://github.com/karpathy/nanoGPT.

Kenneth Levenberg. 1944. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2(2):164–168.

Moufan Li, Kristopher T Jensen, Qihong Lu, Qiong Zhang, and Marcelo G Mattar. 2024. Modeling multiplicity of strategies in free recall with neural networks.

Laura J Lohnas and Alice F Healy. 2021. The role of context in episodic memory: Behavior and neurophysiology. *Psychology of Learning and Motivation*, 75:157–203.

Donald W Marquardt. 1963. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2):431–441.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models.

Bennet B. Jr. Murdock. 1962. The serial position effect of free recall. *Journal of Experimental Psychology*, 64(5):482–488.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Alexander Peysakhovich and Adam Lerer. 2023. Attention sorting combats recency bias in long context language models. *arXiv preprint arXiv:2310.01427*.

Mathis Pink, Vy A. Vo, Qinyuan Wu, Jianing Mu, Javier S. Turek, Uri Hasson, Kenneth A. Norman, Sebastian Michelmann, Alexander Huth, and Mariya Toneva. 2024. Assessing episodic memory in llms with sequence order recall tasks. *Preprint*, arXiv:2410.08133.

Sean M Polyn, Kenneth A Norman, and Michael J Kahana. 2009. A context maintenance and retrieval model of organizational processes in free recall. *Psychological review*, 116(1):129.

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

David C Rubin and Amy E Wenzel. 1996. One hundred years of forgetting: A quantitative description of retention. *Psychological review*, 103(4):734.

Aaditya K Singh, Ted Moskovitz, Felix Hill, Stephanie CY Chan, and Andrew M Saxe. 2024. What needs to go right for an induction head? a mechanistic study of in-context learning circuits and their formation. *arXiv preprint arXiv:2404.07129*.

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

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Benyou Wang, Lifeng Shang, Christina Lioma, Xin Jiang, Hao Yang, Qun Liu, and Jakob Grue Simonsen. 2021. On position embeddings in {bert}. In *International Conference on Learning Representations*.

Yiwei Wang, Yujun Cai, Muhao Chen, Yuxuan Liang, and Bryan Hooi. 2023. Primacy effect of chatgpt. *arXiv preprint arXiv:2310.13206*.

John T Wixted and Ebbe B Ebbesen. 1991. On the form of forgetting. *Psychological science*, 2(6):409–415.

Zexuan Zhong, Tao Lei, and Danqi Chen. 2022. Training language models with memory augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5657–5673, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Appendix

Below we provide plots showing CRP for downstream evaluation, attention scores as a function of lag for all transformer heads for different models (positional encoding magnitude and model size) and different training stages.

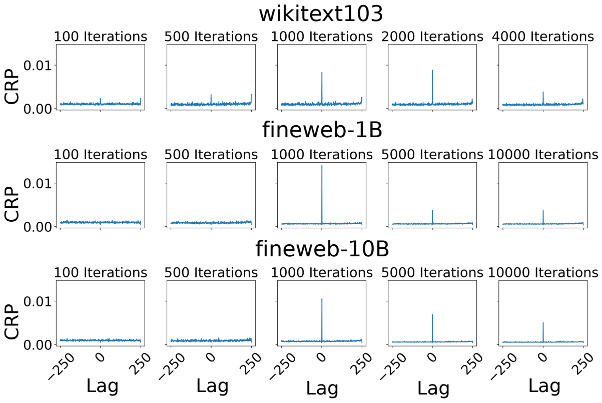

Figure A1: CRP as a function of training iteration in downstream evaluation (same as Fig. 11 but for more lags).

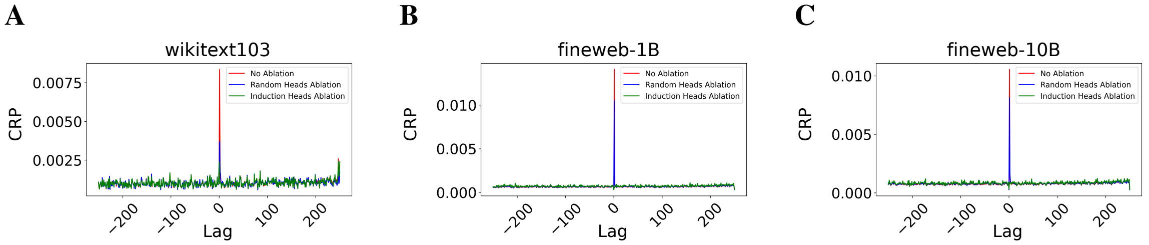

Figure A2: CRP during downstream evaluation showing impact of induction head ablation (same as Fig. 10 but for more lags).

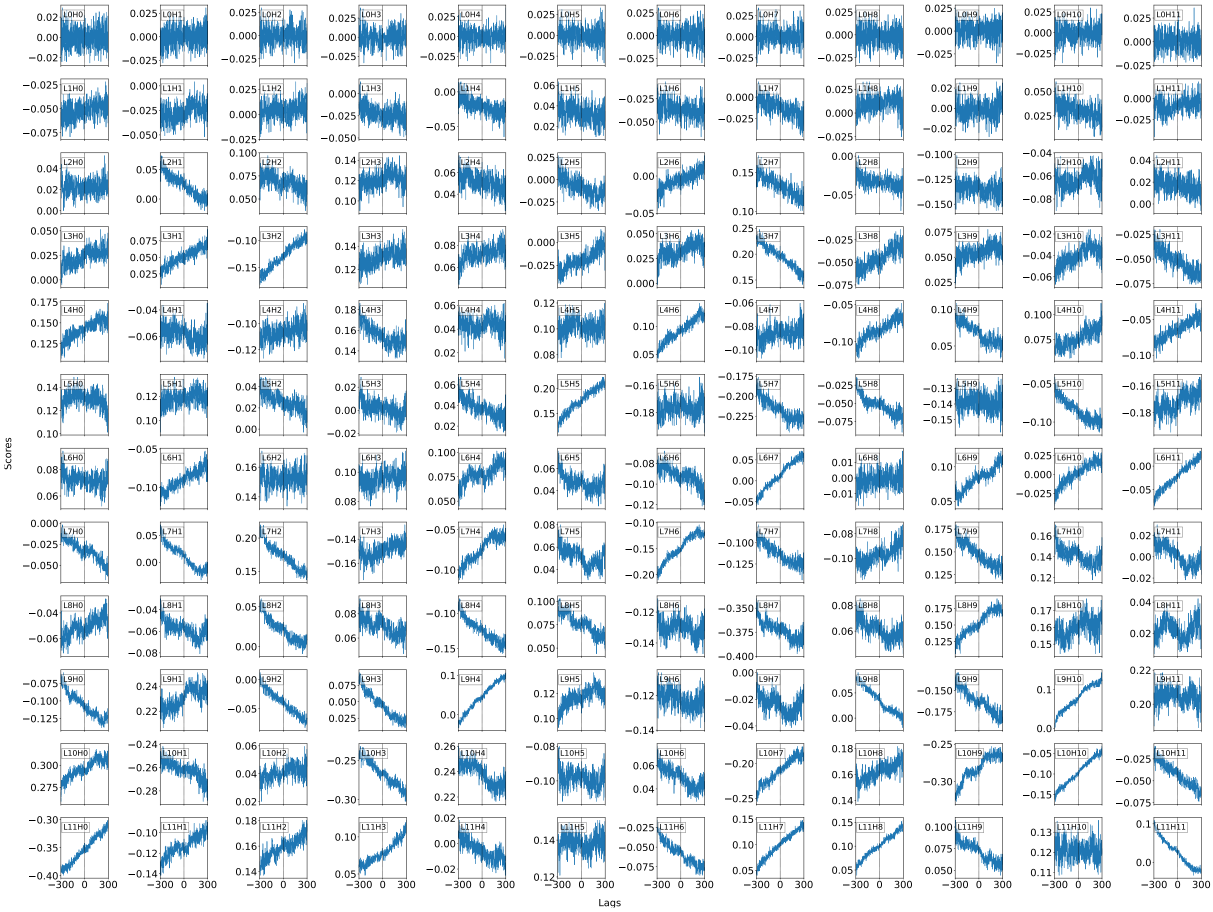

Figure A3: Attention scores as a function of lag for all heads of GPT-2 small before training.

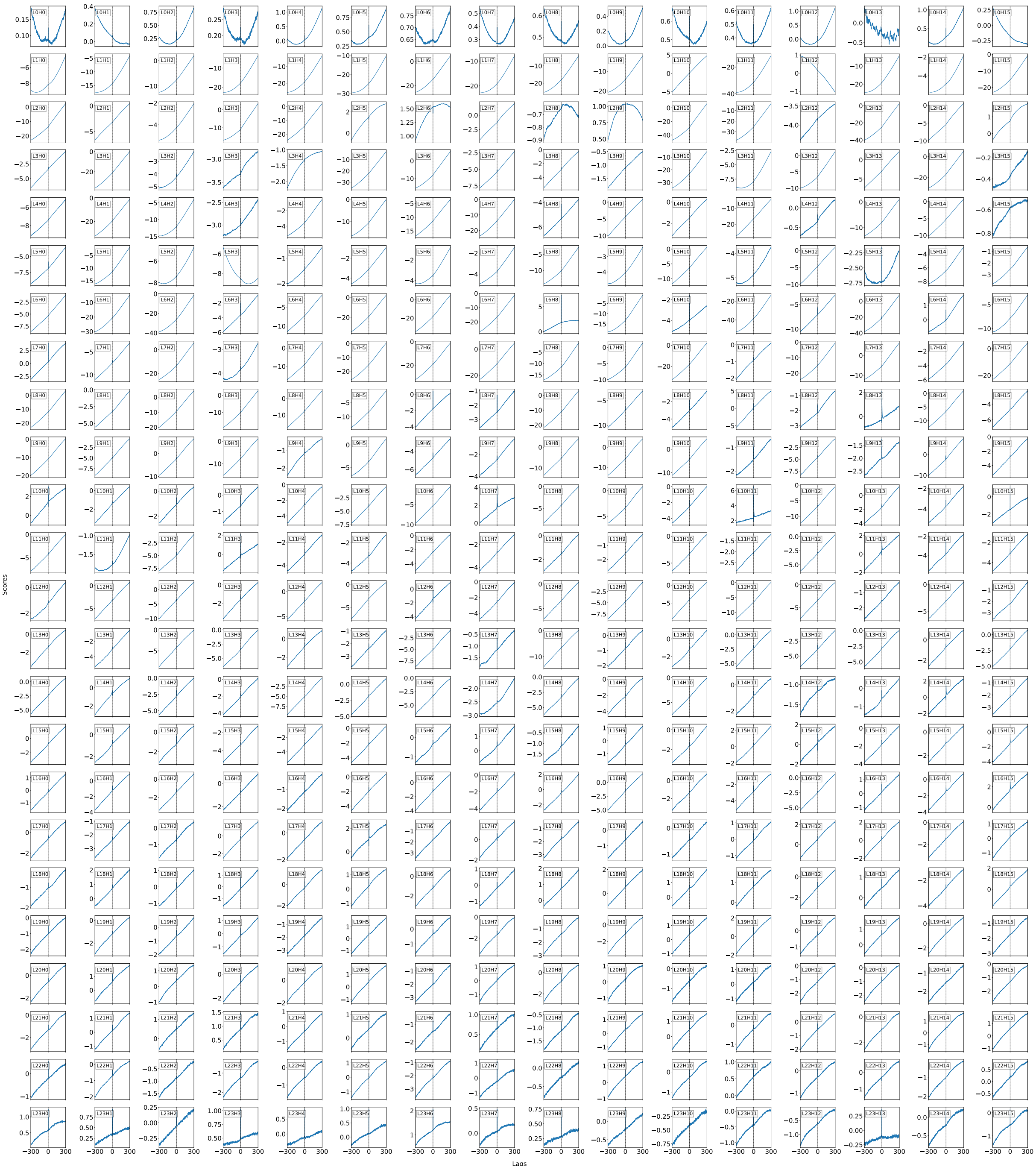

Figure A4: Attention scores as a function of lag for all heads of GPT-2 medium after 2000 iterations for WikiText-103 dataset with a standard amount of positional encoding.

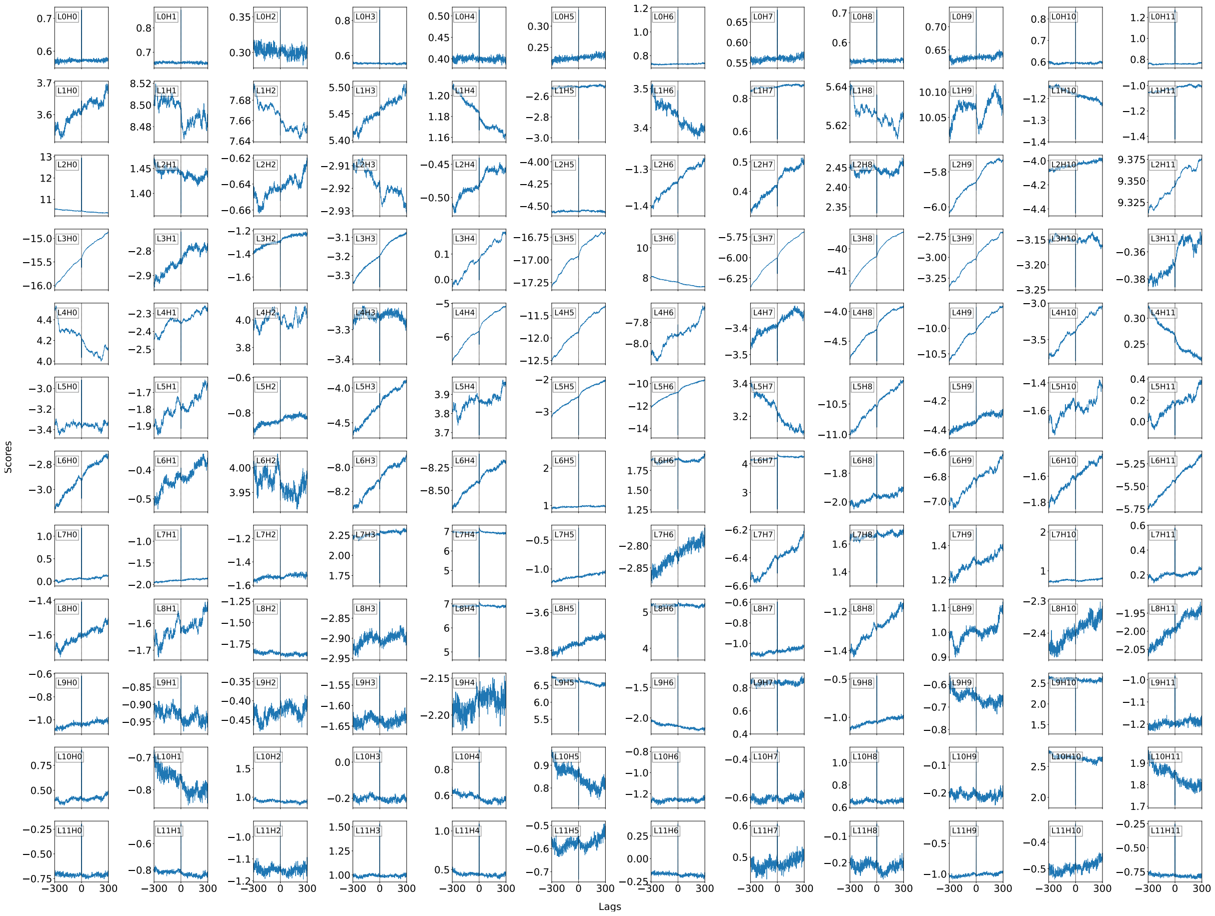

Figure A5: Attention scores as a function of lag for all heads of GPT-2 small after 4000 iterations for WikiText-103 dataset with no positional encoding.

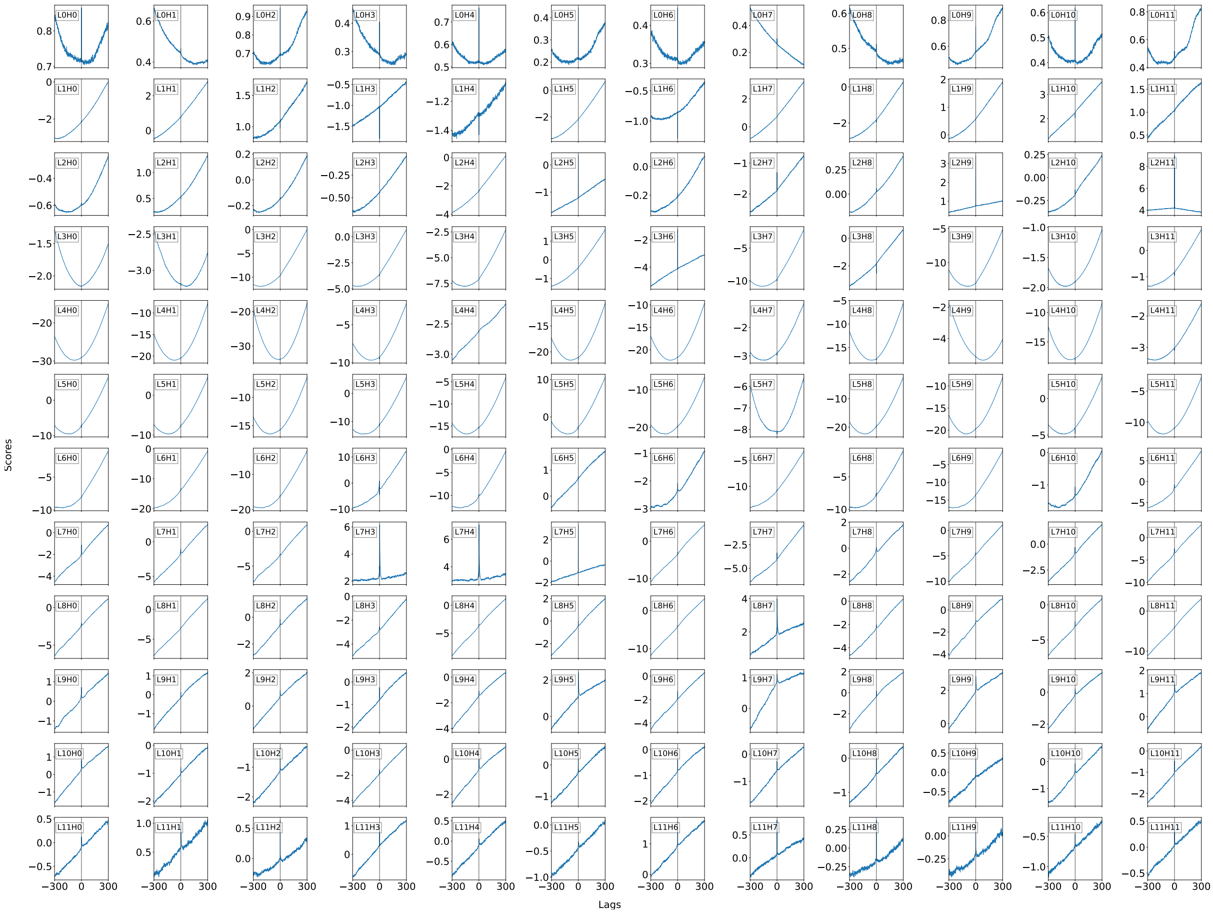

Figure A6: Attention scores as a function of lag for all heads of GPT-2 small after 4000 iterations for Wikitext-103 dataset with double amount of baseline positional encoding.

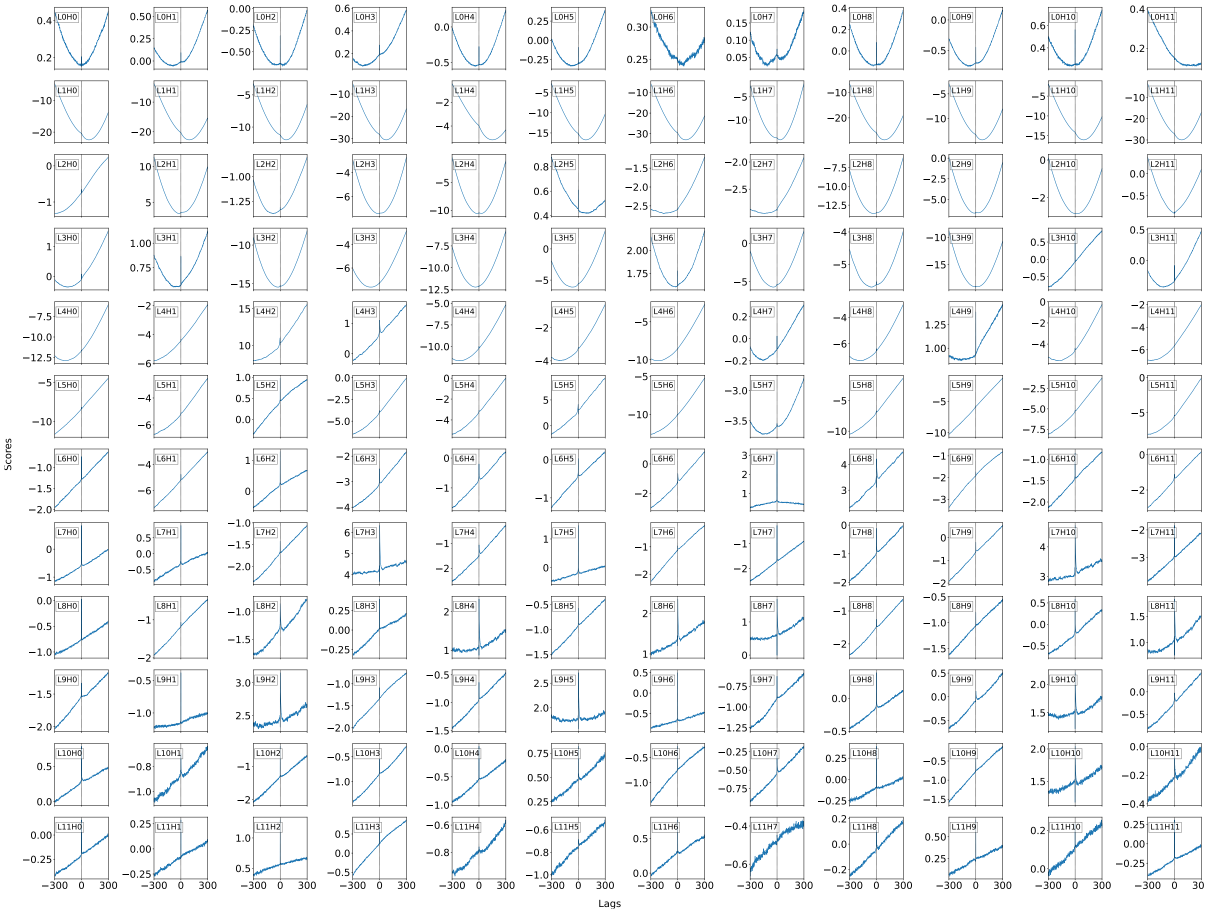

Figure A7: Attention scores as a function of lag for all heads of GPT-2 small after 1000 iterations for Wikitext-103 dataset with baseline amount of positional encoding.

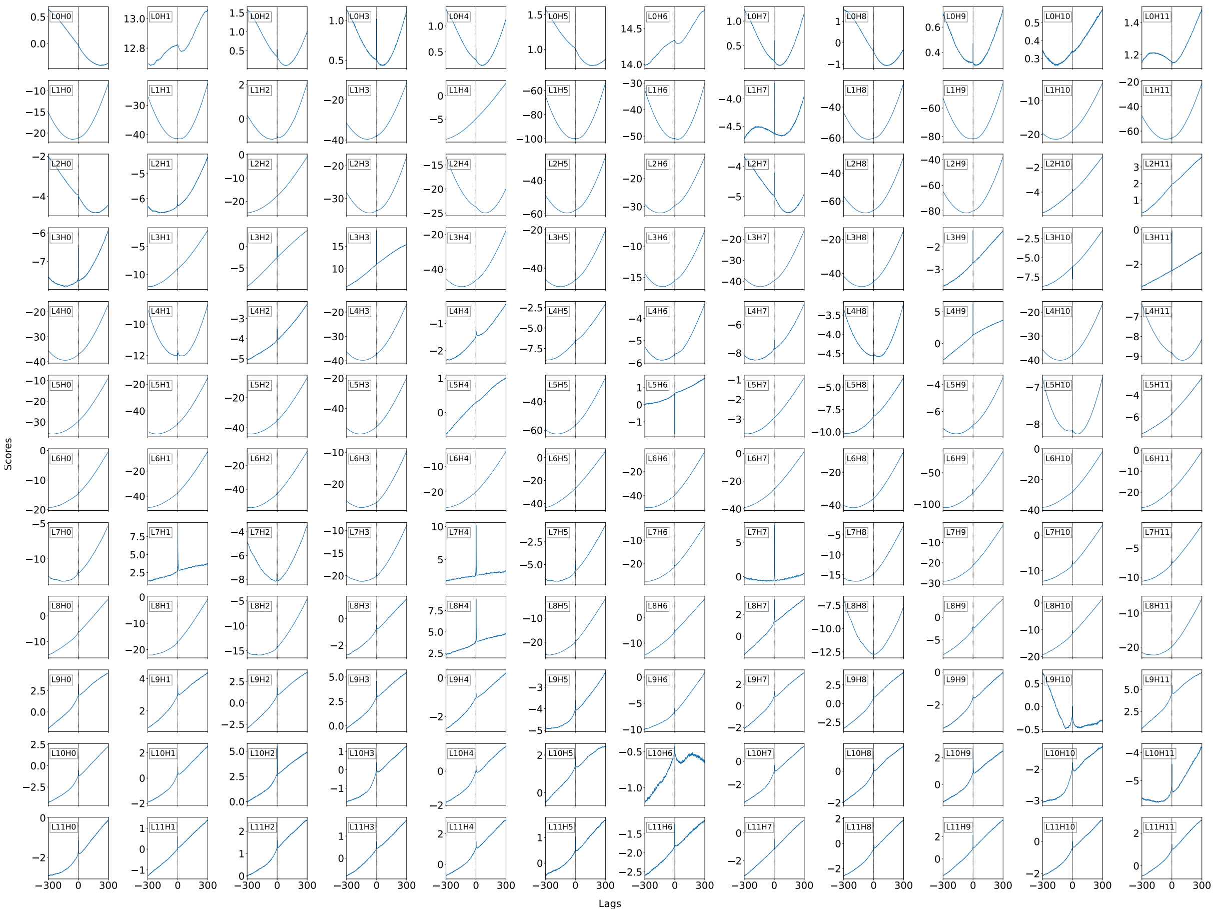

Figure A8: Attention scores as a function of lag for all heads of GPT-2 small after 10000 iterations for FineWeb-1B dataset with baseline amount of positional encoding.

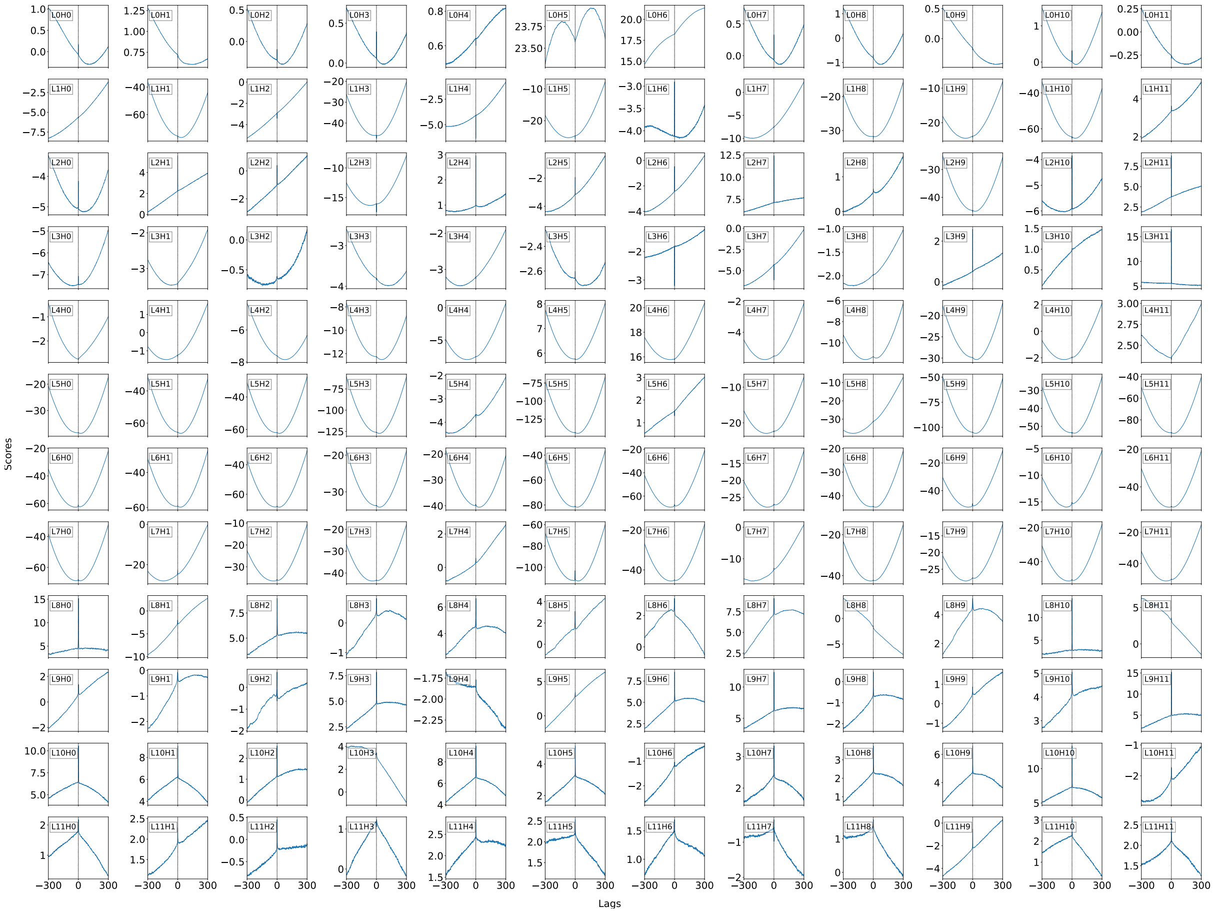

Figure A9: Attention scores as a function of lag for all heads of GPT-2 small after 10000 iterations for FineWeb-10B dataset with baseline amount of positional encoding.