Relating induction heads in Transformers to temporal context model in human free recall

Ji-An Li
University of California, San Diego

Abstract

Induction heads are crucial for the emergence of in-context learning abilities of large language models. In this study, I investigate the relationship between the intriguing properties of induction heads in Transformer models and the temporal context model (TCM) of human sequential memory retrieval during free recall. I establish theoretical connections by reformulating the TCM as a Q-composition-like induction head, followed by an empirical evaluation through the examination of conditional-attention logits in GPT-2 induction heads. My results reveal suprising similarities between Transformer models and cognitive processes in human memory, offering valuable insights into the capabilities of both Transformer models and human memory. Subsequent research will focus on investigating the impact of composition types on CRP-like features and their mechanistic emergence.

Keywords: mechanistic interpretability, induction head, temporal context model

1 Introduction

Mechanistic interpretability seeks to understand computations performed by Transformer generative models as they continue to scale. Previous studies have identified a critical type of self-attention heads known as induction heads [Elhage et al., 2021], which are specialized Transformer circuits that emerge during an early phase-change-like period in training and play a fundamental role in the in-context learning abilities of large language models [Olsson et al., 2022]. These induction heads look back over the previous tokens in the sequence for occurrences of the current token (called [A]), identify what followed (called [B]), and predict a repeat of this pattern (completing the sequence ...[C][A][B][D] ... [A] $\rightarrow$ [B]). Nevertheless, there is a limited understanding of how these circuits attend to tokens such as [C] and [D], which were in close proximity to previous occurrences of the token [A], extending beyond the immediate token [B] following [A].

Previous research has established connections between Transformer models and computational models of hippocampal formation [Whittington et al., 2021]. Through training on a spatial navigation task, it has been demonstrated that the learned position embeddings in Transformer models acquire neural representations resembling grid cells in the cortex, while the self-attention patterns exhibit similarities to the neural representations of place cells in the hippocampus. However, our understanding of the broader connections between Transformer models and hippocampal memory remains limited. The objective of this study is to investigate a potential link between induction heads and the temporal context model (TCM) of sequential memory retrieval during human free recall [Howard and Kahana, 2002, Gershman et al., 2012].

In a free recall task setting, participants are presented with a sequential list of words to study, and subsequently, they are asked to recall the studied words in any order. One of the most fundamental observations is that human recall often reflects the temporal structure of the preceding study list, even though the recall task does not explicitly require temporal ordering. The well-known recall contingency is measured by the conditional-response probability (CRP): if a word just recalled was

Research conducted at the Apart Research Alignment Jam #10 (Interpretability 3.0), 2023 (see https://alignmentjam.com/jam/interpretability)

studied at position j, what is the conditional probability that the next-recalled word was studied at position $j + lag$? For example, consider a participant who studied the sequence ...[C][A][B][D] ... and subsequently recalled word [A]. Consistent findings from various experiments reveal two robust effects (Fig. 1): (i) temporal contiguity effect, where words in close proximity to [A] (e.g., [C] and [B]) are more likely to be recalled compared to those further away; (ii) asymmetry effect, where subsequent words (e.g., [B]) are more likely to be recalled than preceding words (e.g., [C]).

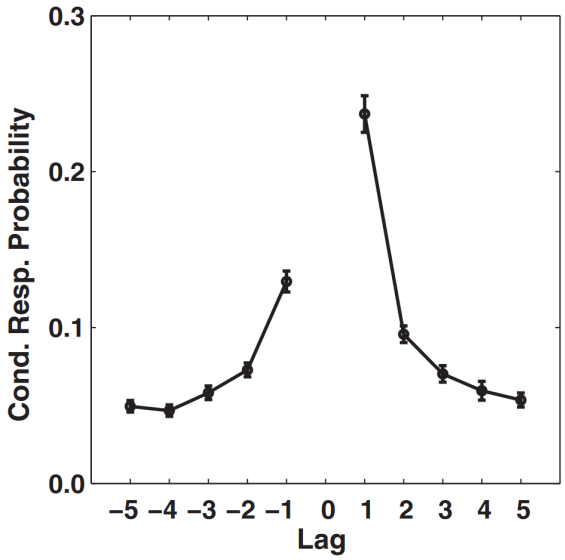

Figure 1: **The conditional-response probability (CRP), as a function of position lag, averaged over nine free recall studies. Figure from Sederberg et al..**

The high probability of attending to the token [B] following [A] in induction heads bears resemblance to the asymmetry effect observed in free recall. In this report, I establish the theoretical connections between induction heads and the TCM through the reformulation of the latter into a Q-composition-like induction head. Additionally, the empirical connections are investigated by analyzing the conditional-attention logits of induction heads in GPT-2, as a function of position lag. My report presents a novel connection between the emergent capabilities of Transformer models and human memory, serving as a bridge between the two domains.

2 Methods

2.1 Reformulating temporal context model as an induction head

In this section, I introduce the components of the TCM and illustrate the mapping of each component to the Q-composition of induction heads (see Fig. 2).

(1) The word (either studied in the list or recalled) at position i is represented as a vector f_i (e.g., one-hot), similar to the token embedding in Transformers.

(2) Each position i is associated with a context vector t_i, conceptually similar to the position embedding in Transformers.

(3) The pair [f_i, t_i] can be considered as the residual stream in Transformers, updated by the head outputs at each position i.

(4) TCM input equation: the proposed context vector update is $t_i^{IN} = M_i^{FT} f_i$, where M_i^{FT} is the item-to-context memory matrix at time i. If the participant studies the word f_i, t_i^{IN} is the f_i-associated context vector before the task experiment (i.e., pre-experimental). If the participant recalls the word f_i, t_i^{IN} is a linear combination of the pre-experimental context vector and the corresponding context vector at the time point when studying the word f_i, because M_i^{FT} has stored the association between the word f_i and the context vector at that time during study. The memory matrix M_i^{FT}

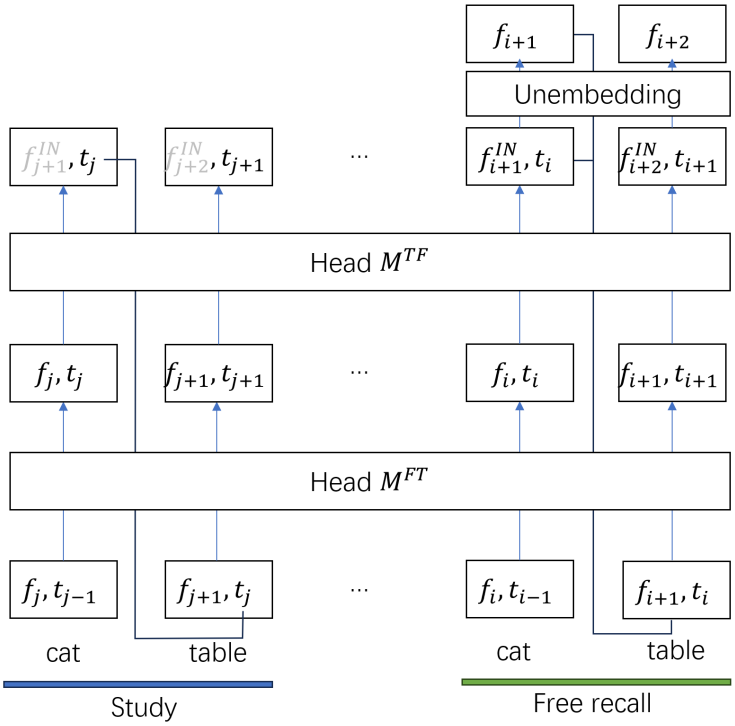

Figure 2: **Mapping TCM to the induction head. See the main text for details.**

(updated in (8)) acts as a first-layer duplicate token head: the current word f_i is the query, previously studied words f_j are keys, and the f_j-associated context vectors are values. This head effectively outputs "what is the position (associated context vector) I saw the same token (word f_i) in the past?"

(5) TCM evolution equation: the context vector is updated by $t_i = \rho_i t_{i-1} + \beta t_i^{IN}$, where ρ_i and β are coefficients to obtain a properly normalized t_i. Equivalently, the duplicate token head writes the "position" information into the residual stream, updating the pair from $[f_i, t_{i-1}]$ to $[f_i, t_i]$. The layer normalization in Transformers plays a similar role as ρ_i and β here.

(6) TCM memory retrieval: the retrieved memory is $f_{i+1}^{IN} = M_i^{TF} t_i$, a linear combination of past word memories, where M_i^{TF} is the context-to-item memory matrix at time i. Because t_i incorporates t_i^{IN}, the context vector associated with the time point studying f_i, f_{i+1}^{IN} will have larger inner products with the studied word vectors nearby the time point studying f_i (contiguity effect). The asymmetry effect comes from the fact that the part of the pre-experimental context vector is associated with words with $lag > 0$ but not $lag < 0$. The M_i^{TF} memory matrix thus acts as a second-layer induction head, taking t_i – affected by the output of the first-layer head – as the query (so-called Q-composition).

(7) TCM word recall: the final retrieved word probability is determined by the inner product between the retrieved memory f_{i+1}^{IN} and each studied word vector f_j, similar to the unembedding layer generating the output tokens from the residual stream.

(8) Updating the item-to-context matrix: $M_{i+1}^{FT} = M_i^{FT} - \frac{(1-A_i)}{|f_i|^2} M_i^{FT} f_i f_i^T + B_i t_i f_i^T$, where A_i and B_i are coefficients to properly normalize M_i^{FT}. Effectively, M_i^{FT} will unlearn the previous associations related to f_i and learn the new association between f_i and t_i. If $A_i = B_i = 1$, this update rule is equivalent to a causal attention head without softmax nonlinearity, where f_i is the key and t_i is the value.

(9) Updating the context-to-item matrix during study: $M_{j+1}^{TF} = M_j^{TF} + f_j t_j^T$ is equivalent to a causal attention head without softmax nonlinearity, where t_j is the key and f_j is the value.

(10) To summarize, the TCM structure is indeed a two-layer attention-only transformer with Q-composition-like induction heads.

A few differences also exist between the TCM model and the typical Transformers. For example, in TCM, t_i is recurrently updated at each time point i; in contrast, the position embeddings in Transformers are either pre-determined or learned without temporal dependence.

2.2 Conditional-attention logits of induction heads

Pretrained Transformer model. I studied the GPT2-small model available in the TransformerLens library. This model has 12 layers (indexed by L) and 12 heads (indexed by H) per layer.

Prompt design. From the unembedding layer of GPT2, I extracted the top 162 frequent tokens corresponding to the largest unembedding biases. Among these tokens, 100 tokens are English words with a leading space (to avoid unwanted tokenization behavior). These tokens are permuted and concatenated into a sequence. The prompt provided to GPT2 consists of two repeats of the same sequence, thus 201 tokens in total (one end-of-sequence token). The first repeat is similar to the study phase, and the second is similar to the recall phase.

Identification of previous token heads, duplicate token heads, and induction heads. Based on the designed prompt, I ran the $transformer\_lens.head\_detector$ function to calculate the previous-token-head matching score, duplicate-token-head matching score, and induction-head matching score of each head.

Copying score. I calculate the copying score of each head as a second measure for induction heads, defined based on the positivity of eigenvalues of the full OV circuit (the path from embedding to OV circuit to unembedding). A positive copying score means that the head will increase the output probability of the attended token.

Head ablation. To discriminate the K-composition type (relying on the previous token head in earlier layers) from the Q-composition type (relying on the duplicate token head in earlier layers) within induction heads, I ablated a subset of previous token heads (L2H2, L3H2, L3H3, L3H7, L4H11), by replacing their output values with zero values. This ablation should affect the K-composition heads more.

Conditional-attention logit. At the position of the second occurrence of token [A], I extract the attention scores for each head around the first occurrence of token [A], with a window width of 16 tokens ($-8 \leq lag \leq 7$). The maximum value is subtracted such that the largest logit in each window always equals 0.

3 Results

Several heads in GPT2 have a relatively large induction-head matching score (Fig. 3). I select the L5H1 head, which has the largest induction-head matching score (0.96), and visualize its conditional-attention logit (Fig. 1). I found that it shows both the temporal contiguity effect (e.g., the logit for $|lag| \leq 2$ is larger than $|lag| \geq 4$) and asymmetry effect (e.g., the logit for $lag > 0$ is larger than $lag < 0$), the two effects observed in human free recall (the $lag = 0$ point is always excluded in CRP analysis). Additionally, I observed a substantial difference between the logit at $lag = 0$ compared to $lag \neq 0$, where a logit difference of 7.5 corresponds to a probability ratio of 1808. To approximate the probability ratio in human experiments (where the range of probability ratio between 5 to 10 corresponds to the range of logit difference between 1.6 to 2.3), one possible approach is to introduce one temperature parameter for attention scores. However, the functional significance of such a substantial logit difference in transformers is not yet clearly understood.

I then searched for the CRP-like features in the conditional-attention logits of all heads and examined the composition type of all induction heads by ablating the previous-token heads (Fig. 5). I found that no head in the first five layers (from L0 to L4, Fig. 6, 7) shows induction head-like behavior (which should have non-zero matching scores and positive copying scores), nor CRP-like conditional-attention logits (although L4H8 is slightly CRP-like).

Starting from layer 5 (Fig. 8, 9, 10, 11), more heads show both induction head-like behavior and CRP-like conditional-attention logits (e.g., L5H1, L5H5, L5H0). I observed that there are non-

induction heads (e.g., L5H8 with a roughly zero matching index and negative copying score) showing CRP-like conditional-attention logits. In general, all induction-like heads (non-zero matching scores and positive copying scores) show CRP-like conditional-attention logits. Many of these CRP-like heads show a range of logit differences similar to that observed in human experiments. However, I found that these matching scores and conditional-attention logits of induction-like heads are generally affected by the ablation of previous-token heads. This suggests that the distinction between K- and Q-composition induction heads is not clear using the current method, and the relations between the composition types and the CRP-like features cannot be identified.

4 Discussion and Conclusion

I have established theoretical connections between the TCM and the Q-composition induction heads. In addition, their empirical similarities are reflected in the CRP-like behaviors of attention heads. One key remaining difference is that the context vector in the TCM recurrently depends on previous time points, different from the typically non-recurrent learned position embeddings in Transformers. Such recurrent position embeddings have been explored in the TEM-transformer model [Whittington et al., 2021].

Future work includes a better characterization of composition types of all the induction heads in the model, which will enable answering what are the effects of composition types on CRP-like features and how these effects emerge mechanistically.

References

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

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

James CR Whittington, Joseph Warren, and Timothy EJ Behrens. Relating transformers to models and neural representations of the hippocampal formation. *arXiv preprint arXiv:2112.04035*, 2021.

Marc W Howard and Michael J Kahana. A distributed representation of temporal context. *Journal of mathematical psychology*, 46(3):269–299, 2002.

Samuel J Gershman, Christopher D Moore, Michael T Todd, Kenneth A Norman, and Per B Sederberg. The successor representation and temporal context. *Neural Computation*, 24(6):1553–1568, 2012.

Per B Sederberg, Jonathan F Miller, MJ Kahana, and MW Howard. Temporal contiguity between recalls predicts episodic memory performance. *Memory & Cognition*.

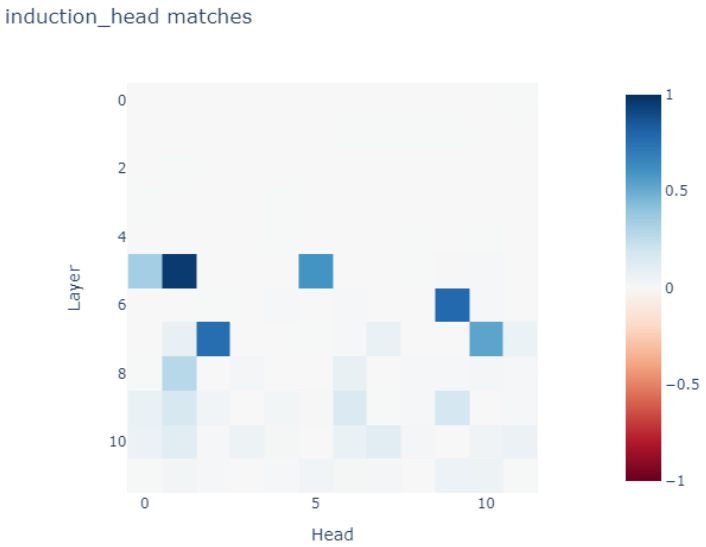

Figure 3: Several heads in GPT2 have a relatively large induction-head matching score (an ideal induction head has a score of one).

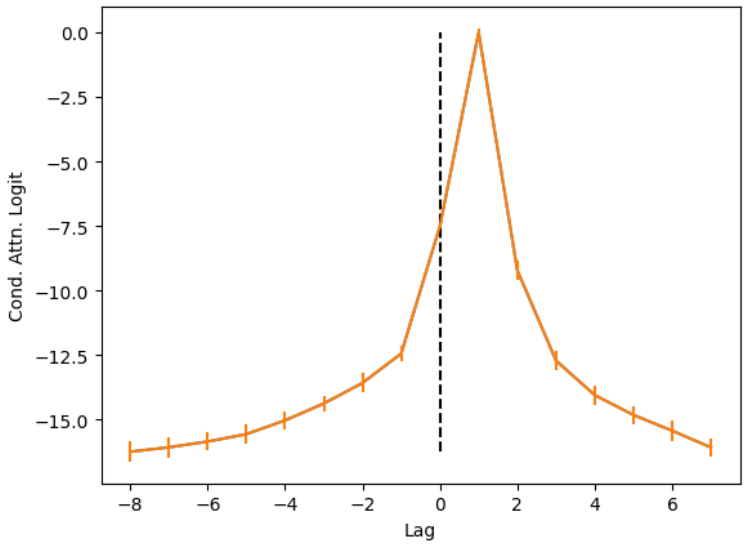

Figure 4: The conditional attention logit of the L5H1 head, as a function of position lag. The logit reaches the maximum at $lag = 1$, consistent with the sequence pattern completion behavior of induction heads. The point at $lag = 0$, which is typically omitted in CRPs, is calculated for comparison. Error bars show SEM across 100 tokens.

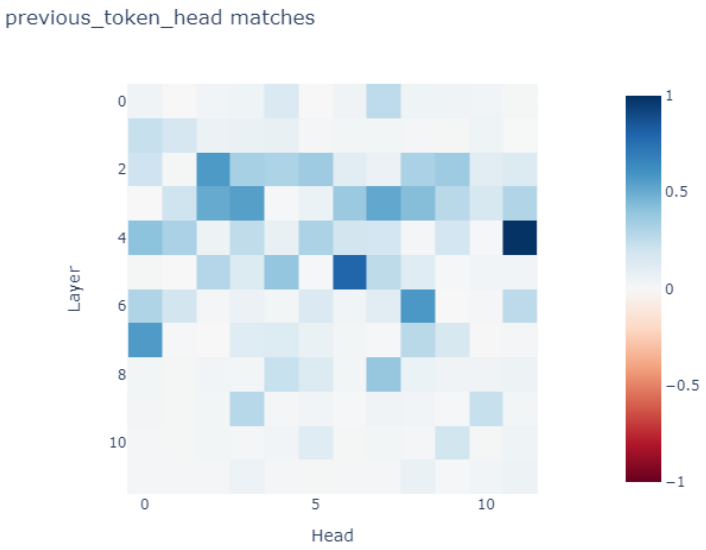

Figure 5: Many heads in GPT2 have a relatively large previous-token-head matching score (an ideal previous token head has a score of one).

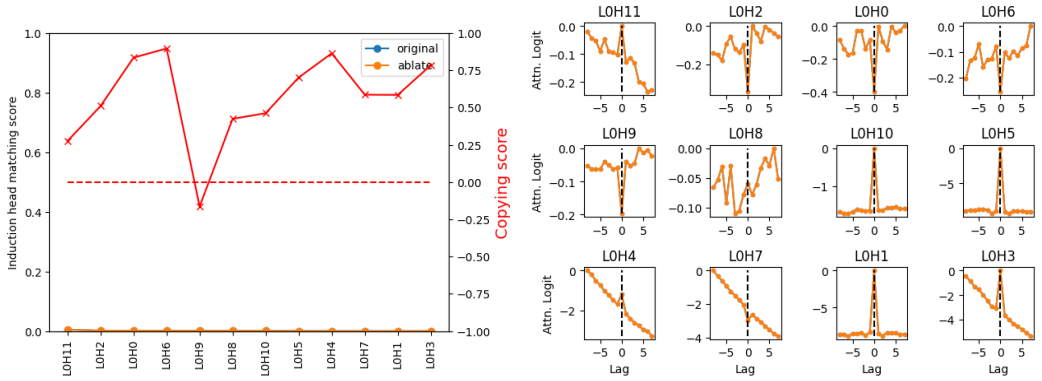

Figure 6: Behavior of all heads in layer 0. (Left) The induction head matching scores (from 0 to 1; left y-axis) of the original GPT2 model (blue), of the modified GPT2 model (orange) with previous token heads ablated, and the copying score (red; from -1 to 1; right y-axis) of the full OV circuit in GPT2 model. Heads are sorted by their induction head matching scores of the original model. (Right) The conditional attention logits of each head. No head shows induction behavior (near-zero matching scores) or CRP-like features.

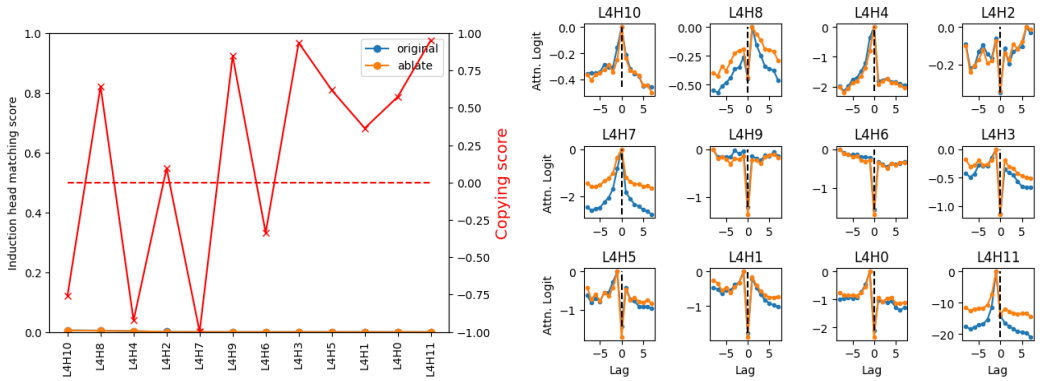

Figure 7: Behavior of all heads in layer 4. (Left) The induction head matching scores (from 0 to 1; left y-axis) of the original GPT2 model (blue), of the modified GPT2 model (orange) with previous token heads ablated, and the copying score (red; from -1 to 1; right y-axis) of the full OV circuit in GPT2 model. Heads are sorted by their induction head matching scores of the original model. (Right) The conditional attention logits of each head. No head shows induction behavior (near-zero matching scores). Only the L4H8 head shows weak CRP-like features.

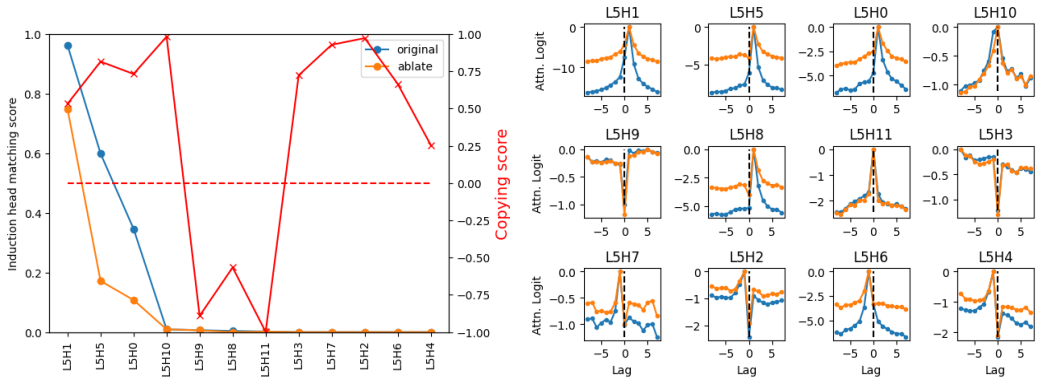

Figure 8: Behavior of all heads in layer 5. (Left) The induction head matching scores (from 0 to 1; left y-axis) of the original GPT2 model (blue), of the modified GPT2 model (orange) with previous token heads ablated, and the copying score (red; from -1 to 1; right y-axis) of the full OV circuit in GPT2 model. Heads are sorted by their induction head matching scores of the original model. (Right) The conditional attention logits of each head. Three heads (L5H1, L5H5, L5H0) show induction behavior (modest to strong matching scores and positive copying scores) and CRP-like features. L5H8 is not an induction head (negative copying score) but shows CRP-like features. These heads are generally affected by the ablation of previous token heads.

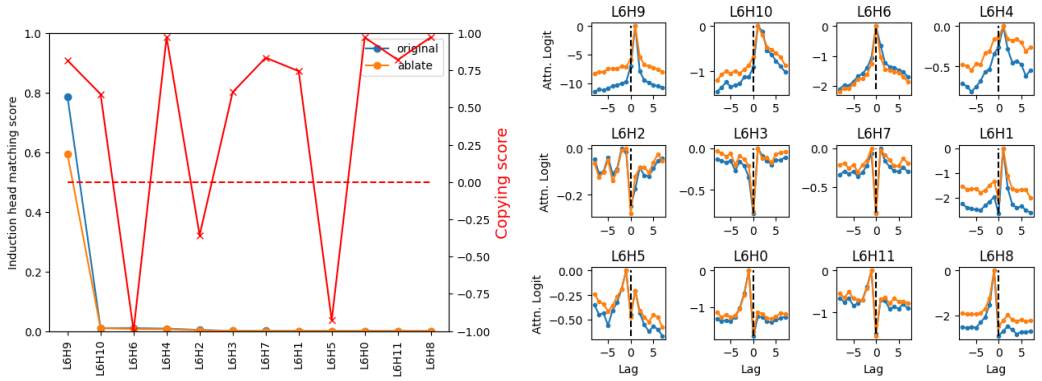

Figure 9: Behavior of all heads in layer 6. (Left) The induction head matching scores (from 0 to 1; left y-axis) of the original GPT2 model (blue), of the modified GPT2 model (orange) with previous token heads ablated, and the copying score (red; from -1 to 1; right y-axis) of the full OV circuit in GPT2 model. Heads are sorted by their induction head matching scores of the original model. (Right) The conditional attention logits of each head. The L6H9 head shows induction behavior (strong matching score and positive copying score) and CRP-like features. Three heads (L6H10, L6H4, L6H1) are not induction heads (due to the near-zero matching scores) but show CRP-like features. These heads are generally affected by the ablation of previous token heads.

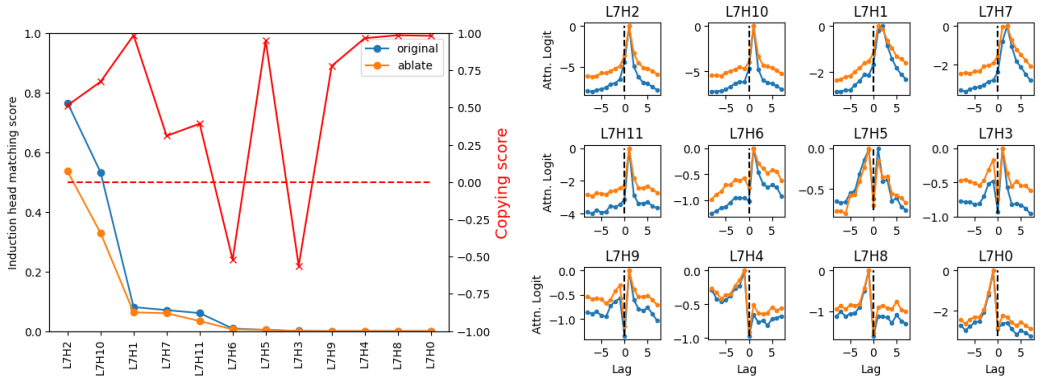

Figure 10: Behavior of all heads in layer 7. (Left) The induction head matching scores (from 0 to 1; left y-axis) of the original GPT2 model (blue), of the modified GPT2 model (orange) with previous token heads ablated, and the copying score (red; from -1 to 1; right y-axis) of the full OV circuit in GPT2 model. Heads are sorted by their induction head matching scores of the original model. (Right) The conditional attention logits of each head. Five heads (L7H2, L7H10, L7H1, L7H7, L7H11) show induction behavior (non-zero matching score and positive copying score) and CRP-like features. Two heads (L7H6, L7H3) are not induction heads (due to the near-zero matching scores and negative copying scores) but show weak CRP-like features. These heads are generally affected by the ablation of previous token heads.

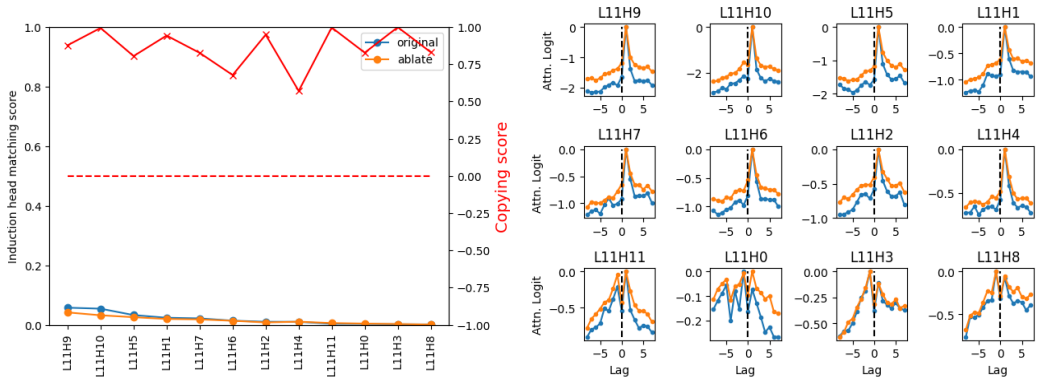

Figure 11: Behavior of all heads in layer 11. (Left) The induction head matching scores (from 0 to 1; left y-axis) of the original GPT2 model (blue), of the modified GPT2 model (orange) with previous token heads ablated, and the copying score (red; from -1 to 1; right y-axis) of the full OV circuit in GPT2 model. Heads are sorted by their induction head matching scores of the original model. (Right) The conditional attention logits of each head. Eight heads show induction-like behavior (non-zero although small matching score and positive copying score) and CRP-like features. These heads are generally affected by the ablation of previous token heads.