Temporal Token Matters: Investigating and Interpreting the Consistency of Temporal Ordering in Large Language Models

Zhen Yang[1†], Xinyue Zhang[1†], Ping Jian[1*], Chengzhi Li[1], Zhongbin Guo[1], Jiaping Feng[1], Wenpeng Lu[2]

[1]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
[2]Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
{bityangzhen, zhangxinyue, pjian}@bit.edu.cn

Abstract

Despite the remarkable performance across numerous tasks, Large Language Models (LLMs) still exhibit notable deficiencies in temporal reasoning, even in simple event ordering tasks. For instance, a slight alteration in the temporal phrasing of the question (e.g., changing "*Is event A before B?*" to "*Is event A after B?*") can lead LLMs to hallucinate and produce inconsistent answers, reflecting a lack of robust temporal reasoning. Although many prior studies have focused on benchmarking and improving the temporal reasoning ability of LLMs, little is known about the intrinsic mechanisms within LLMs when performing temporal reasoning. In this work, we investigate the mechanistic interpretability of temporal ordering within event temporal reasoning through a structured "Identify-Interpret-Verify" pipeline. We first employ path patching to identify a sparse subset of attention heads that are causally responsible for reasoning outcomes. Detailed pattern analysis reveals that these key heads specialize in attending to either *temporal keywords* (semantic cues) or *structural delimiters* (syntactic cues). Furthermore, we rigorously validate the observed mechanism through comprehensive intervention-based experiments, ranging from head ablation to targeted attention modulation. We demonstrate that dynamically modulating the attention of these specific heads can robustly enhance model performance, which serves as strong empirical evidence that our identified mechanism faithfully captures the internal logic of temporal ordering in LLMs.

1 Introduction

Large Language Models (LLMs) have achieved human-comparable performance across diverse natural language processing (NLP) tasks, including complex reasoning domains such as mathematical computation (Jie et al., 2022) and chain-of-thought reasoning (Kojima et al., 2022). However, it has been observed that LLMs often struggle with **temporal reasoning** (Su et al., 2024; Fatemi et al., 2025), which requires models to not only extract factual content but also resolve the relative chronological relationships between events. Even in **basic temporal ordering tasks** such as determining whether "Event A occurs before/after Event B", LLMs frequently exhibit inconsistent outputs when the phrasing is simply flipped from "before" to "after". This lack of robustness poses significant challenges for the reliability of LLMs in time-sensitive applications. Despite considerable efforts to enhance the performance of LLMs in temporal reasoning (Fang et al., 2024a; Ning et al., 2024), the intricate mechanisms governing the model's ability to accomplish these tasks still remain unexplored.

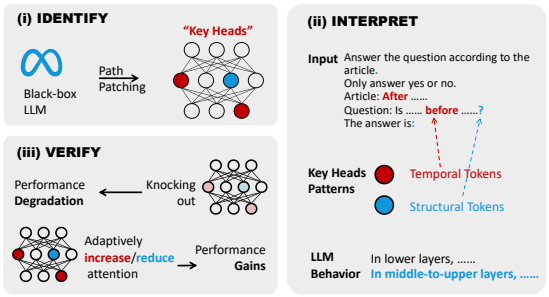


Figure 1: The overall pipeline of our work. (i) **Identifying** key components attributed to event temporal ordering. (ii) **Interpreting** each key attention head's pattern and function. (iii) **Verifying** the mechanism through multi-level causal interventions (knock-out, head-wise, and self-adaptive intervention).

Understanding these mechanisms is crucial for bridging the gap between LLMs' general reasoning capability and their specific failures in the temporal domain. In this work, we aim to uncover and interpret the inner mechanisms of the temporal ordering task within event temporal reasoning (Fang et al., 2024b) in a human-comprehensible manner. Rather than treating the model as a black box for surface-level enhancement, we propose a mech-

† Equal contribution. ∗ Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 2577–2596
July 2-7, 2026

anistic interpretability perspective to identify the specific internal components driving temporal reasoning and validate their causal functions through targeted interventions.

Specifically, to investigate which components play a significant role in the model's temporal ordering process, we implement a classic causal intervention interpretability tool, path patching (Wang et al., 2023b), which quantifies the causal effect of each attention head by selectively perturbing certain activations and observing the changes of final logits. By measuring such logits changes (i.e., causal effect), we identify a sparse set of attention heads, referred to as "key heads". Furthermore, a detailed attention pattern analysis reveals **distinct attention preferences**: while some key heads specifically attend to *temporal tokens* (e.g., "before"/"after"), the others prioritize *structural tokens* (e.g., "<s>" and "?"). This suggests a hierarchical processing mechanism where distinct heads are responsible for parsing semantic temporal cues versus syntactic structural frames.

To rigorously validate this mechanism, we first perform an *ablation study* (knock-out) to confirm the faithfulness of the identified heads, where the attention on specific token types are manually suppressed or enhanced to verify each head's functional role. Furthermore, we propose a *self-adaptive attention intervention* method that dynamically adjusts attention scores based on a head's causal effect and its intrinsic attention pattern. Notably, experiments across distinct models demonstrate that by simply intervening these identified heads on specific tokens, we achieve improvements in both accuracy and consistency. The performance gains provide strong empirical evidence for the faithfulness of uncovered mechanism, which also shows potential for further improving the temporal reasoning ability of LLMs.

In summary, the main contributions of this work are as follows:

- **Identification:** We identify a sparse set of key attention heads within LLMs that are causally responsible for temporal ordering within event temporal reasoning. (Section 3.1)
- **Interpretation:** We reveal a hierarchical mechanism where key heads specialize in attending to either *temporal tokens* or *structural tokens*, providing a fine-grained human-comprehensible view of how LLMs process temporal ordering tasks. (Section 3.2)
- **Causal Verification:** We validate these findings through multi-level interventions and propose a self-adaptive intervention method that improves the temporal ordering performance without fine-tuning. *The resulting performance gains confirm the faithfulness of our identified mechanism.* (Section 4)

2 Task Definition and Dataset

2.1 Task Definition

Event temporal reasoning is defined as identifying temporal relations between events in narratives (Fang et al., 2024b). To facilitate mechanistic analysis, we simplify the task to **ordering two events** based on binary temporal relations (*"before"* or *"after"*). Specifically, given an article with events A and B, the model answers "*yes*" or "*no*" to the question: "*Is event A before/after event B?*" based on the article context.

Formally, the input format of the task is defined as follows:

> **Answer the question according to the article.**
> **Only answer yes or no.**
> **Article:** {An article that contains two events, denoted as event A and event B}
> **Question:** Is event A {before/after} event B?
> **The answer is:**

2.2 Evaluation Metrics

We employ two metrics to evaluate the model performance and robustness of LLMs in event temporal reasoning: accuracy and consistency.

Accuracy measures the proportion of correct predictions:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i), \qquad (1)$$

where N is the dataset size, y_i is the gold label, and $\hat{y}_i$ is the predicted label.

Consistency quantifies the model's robustness to phrasing shifts. For each instance, we generate a reversed counterpart by swapping "before" and "after" in the question. Consistency is defined as the proportion of instances where the model provides logically coherent answers across both versions:

$$\text{Consistency} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i \neq \hat{y}_i^{\text{rev}}), \qquad (2)$$

where $\hat{y}_i$ and $\hat{y}_i^{\text{rev}}$ are predictions for original and reversed questions.

2.3 Dataset

Motivation While existing benchmarks for event temporal ordering, such as TempEvalQA-Bi (Qiu et al., 2023) and TempReason (Tan et al., 2023), provide valuable foundations, they are not suitable for mechanistic interpretability analysis. Specifically, the narratives in these datasets frequently contain redundant or ambiguous contexts where the target events are confounded by irrelevant information, making it difficult to isolate the model's internal temporal reasoning chain. Such linguistic noise obscures the model's internal decision-making process, necessitating a cleaner and simplified dataset to facilitate fine-grained interpretable analysis.

Construction Method We construct a new dataset tailored for interpretability research using GPT-4o (Hurst et al., 2024). For each instance, the model is prompted to generate a daily-life scenario featuring two distinct events with a clear, context-dependent temporal relation. We strictly enforce constraints to avoid meta-references (e.g., "this event") and ensure diverse scenarios. The generation prompt is detailed in Appendix A.1.

From an initial pool of 1,200 instances, we conduct multiple rounds of manual verification. We filter out entries that failed to meet the formatting standards or those where the temporal relation could be inferred via prior knowledge (e.g., "security check" vs. "boarding"), ensuring LLMs to reason based on the provided context.

The final dataset comprises 1,109 entries (539 "before" and 570 "after" instances). The accuracy and consistency of the dataset are evaluated across multiple LLMs, with results presented in Table 1. It can be observed that although the temporal relations are constructed to be simple and explicit, the models still exhibit substantial room for improvement. For both LLaMA models, the consistency is considerably lower than the accuracy, suggesting that their internal reasoning paths for temporal relations are sensitive to phrasing and warrant a deeper mechanistic investigation. Statistical metrics of the constructed datasets are provided in Appendix A.3.

Model Name	Acc(%)	Cons(%)
LLaMA2-7B	60.69	48.69
LLaMA3-8B	90.08	67.72
Qwen2-7B	89.18	88.73

Table 1: Accuracy and consistency of the constructed dataset on LLaMA2-7B(Touvron et al., 2023), LLaMA3-8B(Grattafiori et al., 2024) and Qwen2-7B(Yang et al., 2024a).

3 Mechanistic Analysis

3.1 Identification of Key Attention Heads

3.1.1 Method

The architecture of LLMs can be conceptualized as a directed acyclic graph (DAG), as established in Elhage et al. (2021). Building on this perspective, we employ **path patching** (Wang et al., 2023b), a causal intervention technique that enables fine-grained analysis of specific information pathways by selectively perturbing individual components during the model's forward pass.

To perform path patching on event temporal reasoning, we construct a set of corrupted data, X_c, corresponding to the reference data X_r. While X_r contains the original temporal relations, X_c is generated by altering the relations of questions to induce a distinct model prediction. The forward pass is then computed (i) for both X_r and X_c with all activations, and (ii) by injecting the corrupted activation of a specific attention head h into the forward computation of X_c, while keeping the activations of all other components fixed on X_r.

Formally, the causal effect of head h for the i-th data pair is calculated by,

$$e_h^{(i)} = \frac{logit_p - logit_r}{logit_r - logit_c}, \tag{3}$$

where $logit_r$, $logit_c$ and $logit_p$ denote the output logits from the residual stream from reference data, corrupted data and the patched input, respectively. Aggregating over a dataset with N pairs of data instances, the overall *causal effect* $\overline{e_h}$ of head h is,

$$\overline{e_h} = \frac{1}{N}\sum_{i=1}^{N} e_h^{(i)}. \tag{4}$$

We utilize $\overline{e_h}$ to assess the contribution of each head h to the model's temporal reasoning abilities.

3.1.2 Results

Figure 2 visualizes the causal effects $\overline{e_h}$ across all attention heads and layers. In the heatmap, shades of red signify a negative impact on the target logits when the temporal relation is flipped, while shades of blue indicate a positive contribution, with color intensity reflecting the effect magnitude. Equivalent analyses for LLaMA3-8B and Qwen2-7B are provided in Appendix C.1.

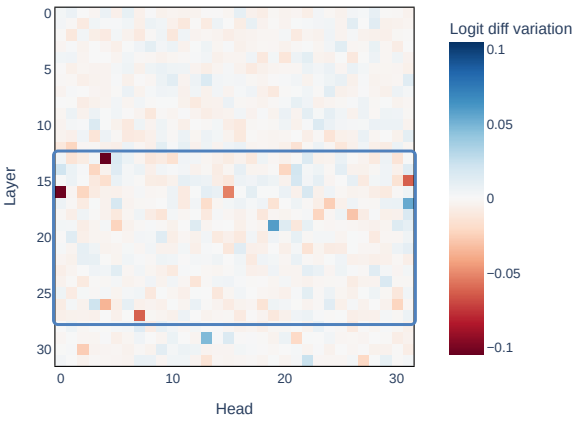


Figure 2: Path patching results on LLaMA2-7B. For each head, a darker color indicates a larger logit difference from the model before patching. The blue box shows the layers where the key attention heads are located.

In Figure 2, it can be observed that, (i) **Only a small portion of attention heads have a noteworthy influence on the output**, highlighting the sparsity of key components in the temporal reasoning task. Following the threshold established in (Zhang et al., 2024), we identify "key heads" as those inducing a logit shift exceeding 5%, with detailed statistics listed in Appendix B.1. (ii) **The discovered "key heads" are mainly located in the middle-to-upper layers**, specifically between layers 13 and 27 for LLaMA2-7B. The above observations align with previous research on other NLP tasks (Zhang et al., 2024; Yang et al., 2025), reinforcing that the lower and upper layers of transformer-based LLMs undertake distinct functions during natural language inference.

3.2 Attention Pattern Analysis for Key Heads

3.2.1 Method

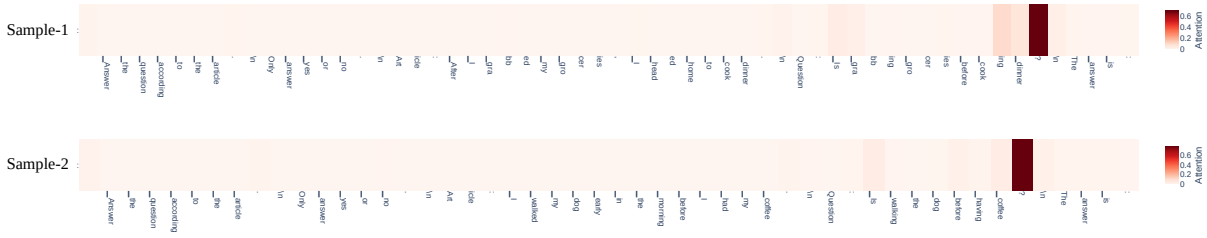
We investigate the functional roles of the identified key heads by examining their attention patterns. Given an input sequence of length s and its attention matrix $A_{i,j} \in \mathbb{R}^{s \times s}$ for the j-th head of layer i, we focus on the attention weights associated with the final token (the query position). Specifically, we extract the last row, denoted as $A_{i,j}^{\text{END}} \in \mathbb{R}^{1 \times s}$, to visualize the distribution of attention scores over preceding tokens. By observing which tokens receive the highest attention scores across distinct samples, we can infer the specific type of information encoded by these heads during the temporal reasoning process.

3.2.2 Results and Discussion

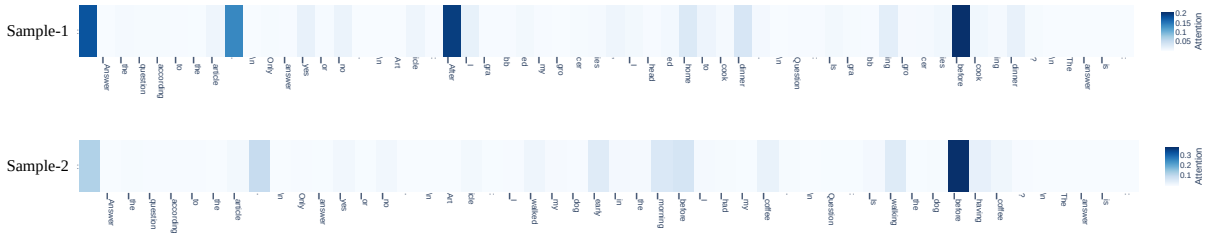
To systematically interpret the behaviors of the heads, we categorize input tokens into four types, following the taxonomy in Zhu et al. (2025): (i) *structural tokens* (e.g., "<s>", "?", "\n"), which provide syntactic and delimiting cues; (ii) *functional tokens* (e.g., "what", "is", "the"), which fulfill grammatical roles; (iii) *temporal tokens* (e.g., "before", "after", "then"), which explicitly indicate temporal relations; and (iv) *event-related tokens* (e.g., "finishing", "checking"), which carry the semantic content of the events.

Figure 3 visualizes the attention patterns of two representative key heads: (13, 4) (red, negative causal effect) and (19, 19) (blue, positive causal effect). Additional visualizations of other key heads on more data samples are provided in Appendix B.3. It can be observed that: (i) **Key heads with a positive causal effect (e.g., (13, 19)) primarily attend to temporal tokens such as "before" and "after", whereas heads with a negative effect (e.g., (13, 4)) focus on structural tokens like "?".** This distinct functional division generalizes robustly to other temporal synonyms (e.g., "prior to" and "subsequent to"), demonstrating that these heads capture semantic temporal information and syntactic structures rather than merely memorizing specific token IDs (see Appendix B.4). It is also worth noting that Head (19, 19) allocates excessive attention to the initial token, a behavior consistent with the "attention sink" phenomenon described in (Xiao et al., 2024). (ii) **None of the key heads exhibiting high causal effects demonstrate a strong emphasis on event-related tokens.** Further analysis reveals that heads with lower causal effects are responsible for attending to event-related tokens, as illustrated in Appendix B.2. These patterns are consistent across LLaMA3-8B and Qwen2-7B (see Appendix C.2).

These observations suggest **a hierarchical processing mechanism for event temporal reasoning in LLMs**: Lower layers (typically associated with low causal effects) employ distributed attention heads to extract diverse and fine-grained information, such as event-related semantics and functional syntax. This information is subsequently integrated and aggregated into the sparse "key" heads located in the middle-to-upper layers. These high-effect key heads specialize in processing higher-level abstractions of inputs, specifically temporal relations and structural delimiters, thereby exerting a deci-



(a) Attention pattern of head (13,4), prioritizing structural tokens.



(b) Attention pattern of head (19,19), prioritizing temporal tokens.

Figure 3: Visualization of attention patterns for two representative key heads. (a) Head (13, 4) (negative impact) focuses predominantly on structural tokens. (b) Head (19, 19) (positive impact) especially attend to temporal tokens across distinct samples.

sive influence on the final prediction.

While these attention patterns offer qualitative insights, a critical question remains: *How to validate that key heads prioritize structural and temporal information, and to what extent do these patterns affect the model's overall performance?*

4 Causal Significance Validation

To validate the faithfulness of the identified key heads, we further conduct ablation and intervention-based experiments to ensure that these heads indeed exhibit causal significance for temporal reasoning.

4.1 Faithfulness Validation via Knocking-out

We first assess the *faithfulness* of the key heads through a "hard knockout" ablation study. Specifically, we sequentially zero out the parameters of the top-k heads ($k \in [1, 7]$, for seven identified key heads) to effectively removing them from the computational graph, ranked in descending order of their causal effect magnitude. The resulting performance degradation is compared with a baseline of randomly ablating the same number of heads (averaged over 10 seeds).

As illustrated in Figure 4, ablating the key heads results in a sharp decline in both accuracy and consistency, while the model demonstrates robustness under the random ablation, with performance levels comparable to the original model. This performance disparity empirically validates that **the identified key heads are essential and necessary for effective event temporal reasoning**.

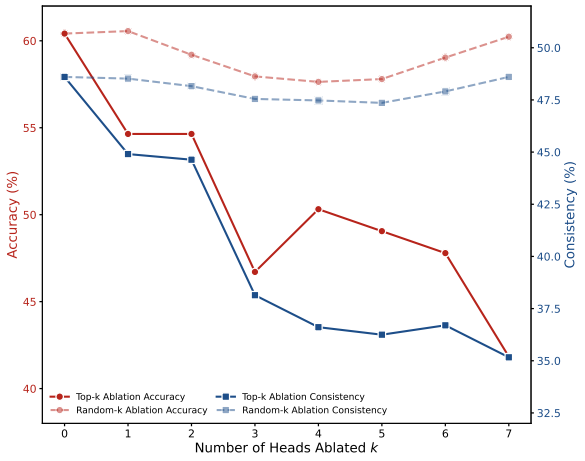


Figure 4: Performance degradation under attention head ablation. Solid lines indicate the impact of removing key heads ranked by causal effect, while dashed lines represent the random ablation baseline.

4.2 Head-wise Intervention

4.2.1 Method

To validate aforementioned mechanism that specific key heads specialize in processing distinct information types (temporal vs. structural), we perform direct attention interventions (Yu et al., 2024). For each target head, we identify a set of specific tokens $\mathcal{T}$ and manipulate their attentions weights during inference to observe the subsequent impact

on model performance.

Formally, given the original attention matrix $A \in \mathbb{R}^{s \times s}$ (indices omitted for brevity), we apply a scaling factor β to the weights of tokens in $\mathcal{T}$ and re-normalize the remaining weights to maintain a valid probability distribution. The intervened attention weights $\hat{A}_{k,t}$ are computed as:

$$\tilde{A}_{k,t} = \begin{cases} \beta \cdot A_{k,t}, & t \in \mathcal{T} \\ A_{k,t}, & t \notin \mathcal{T} \end{cases}$$
$$\hat{A}_{k,t} = \frac{\tilde{A}_{k,t}}{\sum_{j=1}^{s} \tilde{A}_{k,j}}. \quad (5)$$

We design two intervention settings based on Eq. 5. (i) **Structural Attention Reduction**. Motivated by observations that LLMs often over-attend to semantically vacuous structural tokens (Xiao et al., 2025; Sharma et al., 2022), we suppress these tokens (defined as $\mathcal{T}$) to force the redistribution of attention, excluding the attention sink "<s>" (Xiao et al., 2024). (ii) **Temporal Attention Enhancement**. To verify the head's sensitivity to temporal logic, we define $\mathcal{T}$ as temporal keywords (e.g., "before", "after") and amplify their weights, thereby reinforcing the signal for temporal reasoning.

4.2.2 Results and Discussion

Effect of Structural Reduction. Figure 5 reveals divergent outcomes when attention to structural tokens is suppressed: while accuracy improves for certain heads, it notably declines for others. To elucidate the mechanism behind this variation, we examine the attention redistribution in Figure 6. It can be observed that performance improves specifically when the suppression effectively forces the head to redirect its focus toward temporal keywords (e.g., Head (15, 31), Fig. 6(b)). Conversely, performance degrades when the head fails to reallocate attention to temporal cues, instead dispersing it to irrelevant contexts (e.g., Head (27, 7), Fig. 6(a)). This contrast strongly supports our conclusion that **key heads in event temporal reasoning are responsible for identifying either temporal or structural patterns from inputs**, thus **suppressing structural attention proves beneficial only when it triggers a compensatory shift toward temporal logic**.

Effect of Temporal Enhancement. As shown in Figure 7, explicitly enhancing attention on temporal keywords consistently maintains or improves

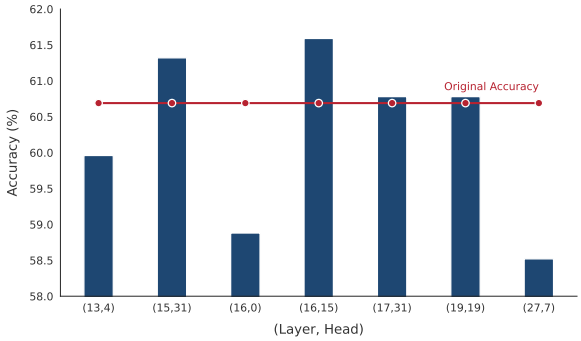


Figure 5: Model accuracy shifts under structural attention reduction. Red line denotes the baseline performance.

accuracy across the identified heads. A minor fluctuation in head (17, 31) remains within a negligible margin. This result confirms that **these heads are functionally dedicated to processing temporal relations, as reinforcing the relevant signal proves generally beneficial**.

4.3 Self-Adaptive Intervention

4.3.1 Method

As aforementioned, we have demonstrated that targeted intervention (suppressing structural attention or enhancing temporal attention) on individual key heads yields performance gains. Therefore, we further explore the potential of self-adaptive attention intervention method. We propose an adaptive intervention method that simultaneously modulating attention patterns across all identified key heads based on their causal effect to improve the temporal reasoning performance.

Specifically, for each key head h, we apply context-aware adjustments derived from the functional verification results (Section 4.2): (i) Attenuating attention scores at structural token positions when suppression improved accuracy, and (ii) Amplifying attention at temporal tokens when suppression degraded performance. The intervention magnitude β_h adapts to each head's average causal effect $\overline{e_h}$:

$$\beta_h = \beta \times \omega_h, \quad (6)$$

where for each head h, ω_h is calculated by,

$$\omega_h = \frac{\overline{e_h}}{\max_{j \in 0,\cdots,N} \overline{e_j}}, \quad (7)$$

where β denotes the global intervention strength, and $\overline{e_h}$ quantifies the causal effect of head h to temporal reasoning, derived from the path patching results. The self-adaptive scaling ensures that

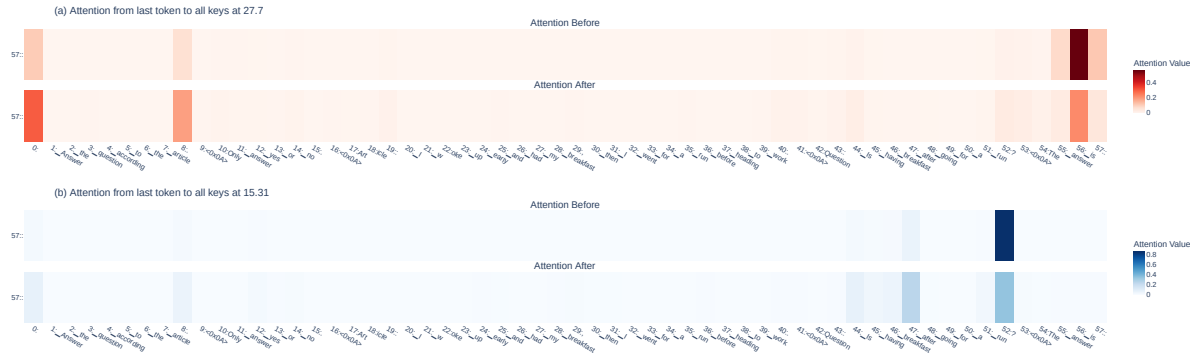


Figure 6: Attention redistribution dynamics following reduction intervention. (a) Head (27, 7) fails to reallocate attention to semantic tokens, degrading performance. (b) Head (15, 31) successfully shifts focus from structural to temporal tokens, resulting in accuracy improvement.

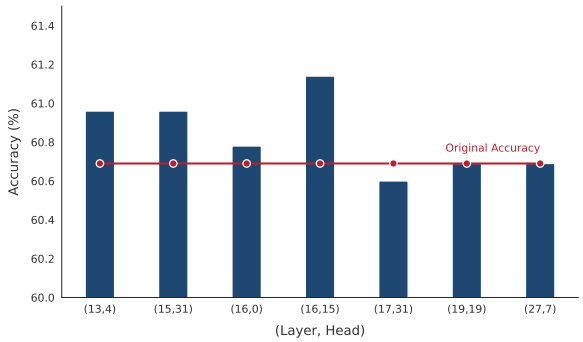


Figure 7: Accuracy shifts under temporal attention enhancement. Red line denotes the baseline performance.

Model	Acc (%)		Cons (%)	
	Value	Δ	Value	Δ
LLaMA2-7B	60.69	-	48.69	-
+ Few-shot	**91.61**	+30.92	39.59	-9.10
+ CoT	52.03	-8.66	47.25	-1.44
+ Full SFT	84.94	+24.25	10.91	-37.78
+ Ours	63.57	+2.88	**50.68**	+1.99

Table 2: Performance comparison on the temporal reasoning task. **Acc**: Accuracy. **Cons**: Consistency. Δ denotes the relative change compared to the LLaMA2-7B baseline. Our method is the only approach that simultaneously improves both metrics.

heads with higher causal effect receive proportionally stronger interventions, controlled by ω_h.

4.3.2 Results and Discussion

We evaluate our self-adaptive intervention on the curated dataset (Section 2.1) with a global intervention strength of $\beta = 0.4$. To benchmark performance, we compare our method against the base LLaMA2-7B model and three standard enhancement strategies: few-shot prompting, Chain-of-Thought (CoT) (Wei et al., 2022), and full-parameter supervised fine-tuning (SFT). Implementation details for prompts and training configurations are provided in Appendices D.1 and D.2.

The comparative results are presented in Table 2. Specifically, regarding *CoT*, the performance unexpectedly degrades (Acc: 52.03%). We attribute this to the explicit nature of temporal relations in the dataset; in such contexts, constructing elaborate reasoning chains tends to introduce hallucinated constraints or extraneous noise rather than clarifying the logic. Conversely, while *Few-shot* and *SFT* achieve substantial gains in accuracy, they suffer from a severe regression in consistency (e.g., SFT drops to 10.91%). Such sharp divergence between high accuracy and low consistency represents a classic manifestation of the "Reversal Curse" (Berglund et al., 2024). During Few-shot and SFT, LLMs tend to learn superficial correlations such as detecting the mere existence of a relationship between two events, rather than genuinely mastering the underlying temporal logic (e.g., distinguishing "before" from "after"). Consequently, their high accuracy appears superficial, failing to generalize to logical negations.

In contrast to the brittleness of these baselines, **our self-adaptive intervention achieves the most robust improvement as the only method to simultaneously enhance both accuracy** (↑ 2.88%) **and consistency** (↑ 1.99%). This dual improvement empirically validates our mechanistic insight: by intrinsically rectifying suboptimal attention patterns—specifically, suppressing structural noise and amplifying temporal cues—we improve the reliability of the reasoning process itself rather

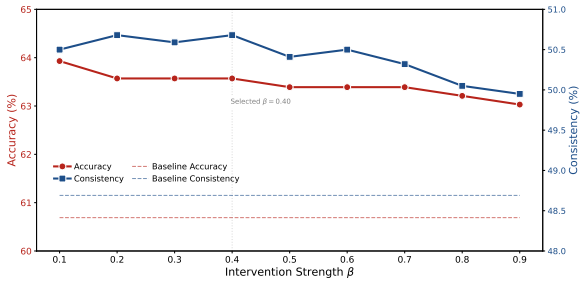


Figure 8: Ablation analysis on the intervention strength β. Accuracy and consistency versus β.

than fitting the output space. Crucially, the adaptive intervention on specific attention distributions provides a *preliminary exploration* of the model's potential temporal reasoning capability. This distinguishes our work from standard optimization strategies; instead of enhancing performance through extensive fine-tuning or heavy context prompting, we aim to elicit the model's latent reasoning power. Additionally, the adaptive scaling mechanism effectively prevents over-intervention on high-impact heads, ensuring stability. Extended results on additional LLMs demonstrating generalizability are detailed in Appendix C.3.

We also conduct an ablation study on the intervention strength β. Figure 8 shows an inverted U-shaped relationship with β, peaking at $\beta = 0.4$ (63.57% accuracy, 50.68% consistency) and declining for $\beta \geq 0.5$. This indicates moderate intervention ($\beta = 0.2$–0.4) optimally balances attention redistribution, while excess values disrupt equilibrium. The plateau between $\beta = 0.2$–0.4 demonstrates the framework's robustness, with $\beta = 0.4$ achieving the optimal trade-off between accuracy and consistency improvements.

5 Related Work

5.1 Mechanistic Interpretability

Interpreting the inner mechanism of LLMs has raised increasing attention in recent years (Räuker et al., 2023; El-Gayar et al., 2024). Most studies in mechanistic interpretability focus on either features or circuits (Rai et al., 2024). The former decodes human-interpretable input properties from the model's activation (e.g., probing (Gurnee et al., 2023; Antverg and Belinkov, 2022), sparse autoencoder (Sharkey et al., 2022)), while the latter regards a LLM as a computational graph to identify a sub-graph that responsible for a specific task or behavior (e.g., visualization (Lieberum et al., 2023), causal mediation analysis (Meng et al., 2022; Wang et al., 2023a; Syed et al., 2023)). This work primarily builds upon research on circuits, including attention analysis and a causal intervention method known as path patching. Specifically, path patching has been employed to explain LLM behaviors across a range of tasks, such as indirect object identification (Wang et al., 2023b), mapping answer text to answer labels (Lieberum et al., 2023), and mathematical computation (Zhang et al., 2024).

5.2 Event Temporal Reasoning

Benchmarks and evaluation frameworks for the event temporal reasoning have evolved to diagnose LLMs' capabilities and limitations, for instance, COTEMPQA (Su et al., 2024) introduced a hierarchical benchmark for four co-temporal scenarios, revealing that even state-of-the-art models like GPT-4 struggle with relationship reasoning. Similarly, (Fatemi et al., 2025) decomposed temporal reasoning into semantic understanding and arithmetic computation, exposing LLMs' reliance on prior knowledge rather than intrinsic reasoning. Further research has also tackled multi-aspect challenges related to event temporal reasoning, (Fang et al., 2024a) introduced a method to detect knowledge conflicts arising from mismatches between prior knowledge and event relations, using counterfactual data augmentation to reduce hallucination. (Ning et al., 2024) utilized a temporal cognitive tree to reason through multiple levels of temporal relations and optimize inference via deductive reasoning and multi-task learning. (Yang et al., 2024b) proposed a dedicated framework integrating temporal information-aware Embedding and a granular contrastive reinforcement learning strategy to enhance LLMs' temporal sensitivity and reasoning. Despite these innovations, core challenges persist in interpretability and complex reasoning. LLMs exhibit "uncertainty errors" when deducing implicit temporal links (Su et al., 2024), indicating the necessity of our interpretability-based work.

6 Conclusion

In this work, we present a mechanistic analysis of the event temporal ordering task, a crucial task for evaluating the temporal reasoning ability of LLMs. By adopting an "Identify-Interpret-Verify" framework, we first identify a sparse set of attention heads located in the middle-to-upper layers that are critical for processing temporal patterns. Crucially,

via fine-grained attention pattern analysis, we reveal a specialized processing mechanism where the key heads integrate information from lower-layer heads while especially attending to temporal and structural patterns. Furthermore, we validate that these identified components directly control reasoning performance, and translate these insights into a self-adaptive intervention strategy. The resulting performance gains further demonstrate that interpreting internal model mechanisms is essential for enhancing the robustness and reliability of complex temporal reasoning tasks.

Limitations

Despite the promising findings in interpreting and improving event temporal reasoning, our work has several limitations. Firstly, our analysis relies on a constructed dataset where temporal relations are binary ("before/after") and explicit. While this controlled setting is necessary for isolating causal components, it simplifies real-world complexities involving multi-hop reasoning, vague timelines, or implicit temporal dependencies. Consequently, the verification of identified mechanism in such implicit reasoning scenarios warrants further investigation. Secondly, due to the necessity of accessing internal parameters for path patching and causal intervention, our experiments are restricted to open-weight models (e.g., LLaMA and Qwen). The lack of access to internal parameters and activations in closed-source models prevents a direct verification of our results in those settings.

Ethics Statement

This paper investigates the mechanistic interpretability of event temporal reasoning in Large Language Models. The dataset constructed for this study was generated using GPT-4o and underwent multiple rounds of manual verification to ensure high quality. We confirm that the generated data samples do not contain any personally identifiable information, offensive content, or ambiguous temporal relations. Additionally, the use of open-source models (LLaMA, Qwen) and existing benchmarks (MATRES) in this work adheres to their respective licenses and is consistent with their intended research use.

The inner mechanisms and intervention techniques uncovered in this paper provide valuable insights for understanding and improving LLM behavior in temporal reasoning tasks. Nonetheless, we acknowledge that methods capable of modulating attention patterns to steer model outputs could theoretically be misused to maliciously manipulate temporal logic or induce hallucinations at inference time. Therefore, we emphasize the importance of responsible deployment and rigorous monitoring when applying such white-box interventions. Furthermore, we do not see any other potential risks associated with this work.

Acknowledgements

This work is supported by the grants from the National Natural Science Foundation of China (No. 62376130). The authors would like to thank the organizers of ACL 2026 and the reviewers for their helpful suggestions.

References

Omer Antverg and Yonatan Belinkov. 2022. On the pitfalls of analyzing individual neurons in language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2024. The reversal curse: Llms trained on "a is b" fail to learn "b is a". In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Comput. Linguistics*, 16(2):79–85.

Omar F. El-Gayar, Mohammad Al-Ramahi, Abdullah Wahbeh, Tareq Nasralah, and Ahmed Elnoshokaty. 2024. A comparative analysis of the interpretability of LDA and LLM for topic modeling: The case of healthcare apps. In *30th Americas Conference on Information Systems: Elevating Life through Digital Social Entrepreneurship, AMCIS 2024, Salt Lake City, UT, USA, August 15-17, 2024*. Association for Information Systems.

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

Tianqing Fang, Zhaowei Wang, Wenxuan Zhou, Hongming Zhang, Yangqiu Song, and Muhao Chen. 2024a. Getting sick after seeing a doctor? diagnosing and

mitigating knowledge conflicts in event temporal reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 3846–3868. Association for Computational Linguistics.

Tianqing Fang, Zhaowei Wang, Wenxuan Zhou, Hongming Zhang, Yangqiu Song, and Muhao Chen. 2024b. Getting sick after seeing a doctor? diagnosing and mitigating knowledge conflicts in event temporal reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 3846–3868. Association for Computational Linguistics.

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. 2025. Test of time: A benchmark for evaluating llms on temporal reasoning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. 2023. Finding neurons in a haystack: Case studies with sparse probing. *Trans. Mach. Learn. Res.*, 2023.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Zhanming Jie, Jierui Li, and Wei Lu. 2022. Learning to reason deductively: Math word problem solving as complex relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5944–5955. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. 2023. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. *CoRR*, abs/2307.09458.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Qiang Ning, Hao Wu, and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. In *ACL*.

Wanting Ning, Lishuang Li, Xueyang Qin, Yubo Feng, and Jingyao Tang. 2024. Temporal cognitive tree: A hierarchical modeling approach for event temporal relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 855–864. Association for Computational Linguistics.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo M. Ponti, and Shay B. Cohen. 2023. Are large language models temporally grounded? *CoRR*, abs/2311.08398.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2024. A practical review of mechanistic interpretability for transformer-based language models. *CoRR*, abs/2407.02646.

Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. 2023. Toward transparent AI: A survey on interpreting the inner structures of deep neural networks. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning, SaTML 2023, Raleigh, NC, USA, February 8-10, 2023*, pages 464–483. IEEE.

Lee Sharkey, Dan Braun, and Beren Millidge. 2022. Taking features out of superposition with sparse autoencoders. In *AI Alignment Forum*, volume 6, pages 12–13.

Rishab Sharma, Fuxiang Chen, Fatemeh H. Fard, and David Lo. 2022. An exploratory study on code attention in BERT. In *Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension, ICPC 2022, Virtual Event, May 16-17, 2022*, pages 437–448. ACM.

Zhaochen Su, Juntao Li, Jun Zhang, Tong Zhu, Xiaoye Qu, Pan Zhou, Yan Bowen, Yu Cheng, and Min Zhang. 2024. Living in the moment: Can large language models grasp co-temporal reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 13014–13033. Association for Computational Linguistics.

Aaquib Syed, Can Rager, and Arthur Conmy. 2023. Attribution patching outperforms automated circuit discovery. *CoRR*, abs/2310.10348.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pages 14820–14835. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023a. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2717–2739. Association for Computational Linguistics.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023b. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. 2025. Duoattention: Efficient long-context LLM inference with retrieval and streaming heads. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024a. Qwen2 technical report. *CoRR*, abs/2407.10671.

Wanqi Yang, Yanda Li, Meng Fang, and Ling Chen. 2024b. Enhancing temporal sensitivity and reasoning for time-sensitive question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 14495–14508. Association for Computational Linguistics.

Zhen Yang, Ping Jian, and Chengzhi Li. 2025. Option symbol matters: Investigating and mitigating multiple-choice option symbol bias of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 1902–1917. Association for Computational Linguistics.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. 2024. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Wei Zhang, Chaoqun Wan, Yonggang Zhang, Yiu-ming Cheung, Xinmei Tian, Xu Shen, and Jieping Ye. 2024. Interpreting and improving large language models in arithmetic calculation. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Jingze Zhu, Yongliang Wu, Wenbo Zhu, Jiawang Cao, Yanqiang Zheng, Jiawei Chen, Xu Yang, Bernt Schiele, Jonas Fischer, and Xinting Hu. 2025. Layercake: Token-aware contrastive decoding within large language model layers. *arXiv preprint arXiv:2507.04404*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 1097–1100. ACM.

A Dataset

A.1 Prompt for Dataset Construction

The specific prompt for GPT-4o to generate data instances is displayed in Figure 9.

Prompt for Dataset Construction

Generate exactly one JSON entry for temporal reasoning analysis. Strictly follow this structure:
{
"input": "Answer the question according to the article.\nOnly answer yes or no.\nArticle: [CONTEXT]\nQuestion: Is Event A before/after Event B?.\nThe answer is:",
"output": "yes"
}
Requirements:
1. Article:
- Contains exactly two distinct events (Event A and Event B) with clear temporal order
- Explicitly states Event A occurs BEFORE/AFTER Event B using temporal words or phrases
- Maximum 50 words, but not less than 15 words
- No ambiguous time references
- Natural language in past/perfect tense (e.g., "I finished X then did Y")
2. Question:
- Always uses the format: "Is [Event A] before/after [Event B]?"
- Events must exactly match the phrasing in the article
3. Constraints:
- Temporal relationship must be unambiguous and correct (the correct answer to the question must always be yes)
- Cover diverse daily scenarios: cooking, travel, work, study, shopping, sports, etc.
- Avoid meta-references (e.g., "this event")
- No markdown/comments in output
Output ONLY the JSON object with no additional text.

Figure 9: Prompt for Dataset Construction.

A.2 Examples of Data Instances

Table 3 shows several examples of data instances from the dataset that we constructed.

A.3 Statistical Metrics

Two critical properties, *diversity* and *fluency* are assessed in order to further validate the quality of the constructed dataset. Diversity is quantified using Self-BLEU (Zhu et al., 2018), which measures inter-sample similarity by comparing each text against others as references. Meanwhile, fluency is evaluated through Perplexity (PPL) (Brown et al., 1990) and the results are shown in Table 5.

The Self-BLEU score indicates scenario diversity across samples, while the PPL value demonstrates strong linguistic quality. These metrics collectively validate the dataset's suitability for robust temporal reasoning analysis.

B Detailed Key Heads and Attention Patterns

B.1 Statistics of Identified Key Heads

Locations and causal effects of the identified key heads are listed in Table 6.

B.2 Attention Patterns of Low Causal Effect Heads

As discussed in Section 3.2, the key attention heads with highest causal effects primarily focus on structural and temporal tokens. Further experiments on heads with lower causal effects reveal that many of these heads do attend to event-related tokens. Figure 14 and Figure 15 illustrate two examples of less important attention heads attending to event-related tokens.

B.3 Attention Patterns of Key Heads

The attention patterns of all seven key heads are visualized, while only two of them (head (13, 4) and head (19, 19)) are displayed in Section 3.2. Attention patterns of other key heads across multiple data instances are shown from Figure 16 to Figure 20. The figures further prove our observation concluded in Section 3.2, that key heads with positive logit change after path patching focus primarily on temporal tokens, while the ones with negative logit change attend more scores to structural tokens.

B.4 Generalization to Temporal Synonyms

To verify whether the identified key heads perform genuine semantic processing rather than merely memorizing specific token IDs (e.g., "before" and "after"), we conduct an additional generalization study. We replace the original temporal relational words with synonymous phrases, specifically utilizing the templates "*Is event A prior to / subsequent to event B?*".

We measure the attention scores of the representative key heads on the new template to observe their behavior shifts:

- **Temporal Head (19, 19):** Originally identified as attending to "before/after", this head successfully shifts its focus to the new synonyms. Across the evaluation set, it assigns high average attention scores to the tokens

Input	Gold Answer
Answer the question according to the article.\nOnly answer yes or no.\nArticle: I woke up early and had my breakfast, then I went for a run before heading to work.\nQuestion: Is having breakfast before going for a run?\nThe answer is:	yes
Answer the question according to the article.\nOnly answer yes or no.\nArticle: After finishing my homework, I watched a documentary on history.\nQuestion: Is finishing the homework before watching the documentary?\nThe answer is:	yes
Answer the question according to the article.\nOnly answer yes or no.\nArticle: After I grabbed my groceries, I headed home to cook dinner.\nQuestion: Is grabbing groceries before cooking dinner?\nThe answer is:	yes
Answer the question according to the article.\nOnly answer yes or no.\nArticle: I attended the meeting and discussed the project before heading to lunch.\nQuestion: Is attending the meeting before heading to lunch?\nThe answer is:	yes
Answer the question according to the article.\nOnly answer yes or no.\nArticle: He attended a meeting, and after that, he went out for lunch.\nQuestion: Is lunch after the meeting?\nThe answer is:	yes

Table 3: Examples of the constructed dataset

Input	Gold Answer
Answer the question according to the article.\nOnly answer yes or no.\nArticle: The FAA on Friday announced it will close 149 regional airport control towers because of forced spending cuts – sparing 40 others that the FAA had been expected to shutter .\nQuestion: Is the event 'expected' happening before the event 'begin'?\nThe answer is:	yes

Table 4: A finetuning dataset instance.

Metric	Value
Self-BLEU	0.6865
PPL (LLaMA2-7B)	10.8548

Table 5: Evaluation results of diversity and fluency metrics for the constructed dataset.

i-th Layer	j-th Head	Logit Diff(%)
13	4	-10.49
15	31	-6.30
16	0	-10.31
16	15	-5.30
17	31	5.37
19	19	5.88
27	7	-6.26

Table 6: Locations and logit differences of the identified key attention heads, sorted by layer index.

prior/subsequent (0.2168) and *to* (0.1991), while largely ignoring the structural token *?* (0.0020) and other context tokens (avg = 0.0060).

- **Structural Head (13, 4):** Originally identified as attending to structural delimiters, this head maintains its specific role. It consistently focuses on the question mark *?* (0.7987), with negligible attention directed towards the semantic tokens *prior/subsequent* (0.0227) or other background tokens (avg = 0.0077).

A detailed case study is presented in Table 7, which illustrates the token-level attention score distribution of both heads on the exact same input sequence. For simplicity and clarity, tokens preceding the "Question" segment are omitted.

These results align closely with our original findings under the "before/after" setting (as illustrated in Figure 3). This robust consistency confirms that the identified key heads do not overfit to specific vo-

cabulary; rather, they successfully capture broader abstract semantic categories (i.e., temporal indicators) and underlying syntactic structures.

Token	(19, 19)	(13, 4)
Question	0.0019	0.0007
:	0.0017	0.0071
__Is	0.0068	0.0404
__gra	0.0029	0.0206
bb	0.0015	0.0001
ing	0.0221	0.0034
__gro	0.0029	0.0004
cer	0.0031	0.0001
ies	0.0079	0.0042
__prior	**0.1213**	0.0054
__to	**0.1185**	0.0036
__cook	0.0047	0.0010
ing	0.0021	0.0704
__dinner	0.0191	0.0478
?	0.0026	**0.7185**
\n	0.0005	0.0241
The	0.0006	0.0046
__answer	0.0002	0.0020
__is	0.0002	0.0023
:	0.0006	0.0019

Table 7: Attention score distribution of Temporal Head (19, 19) and Structural Head (13, 4) on a sample sequence using the temporal synonym "prior to".

B.5 Attention Redistribution

In Section 4.2, in order to validate the function of each key head, the attention reduction experiments are conducted on each key head individually. Besides the two key heads displayed in the main text, (27, 7) and (15, 31), the attention redistribution of the other ones is shown in Figure 21. The visualized results align with our conclusion in Section 4.2.

C Mechanism Analysis on other LLMs

In this section, we provide supplementary results on LLaMA3 and Qwen2 including: (i) identification of key attention heads through path patching, (ii) detailed attention patterns of the identified key heads, and (iii) self-adaptive intervention results. All results on LLaMA3 and Qwen2 align with our observations and conclusions in Section 3 and Section 4, confirming the generalization of our identified temporal reasoning mechanism.

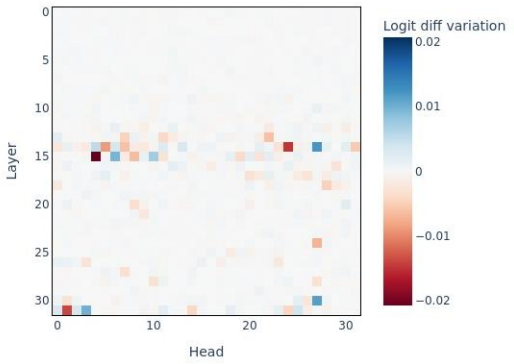


Figure 10: Path patching results on LLaMA3-8B. For each head, a darker color indicates a larger logit difference from the model before patching.

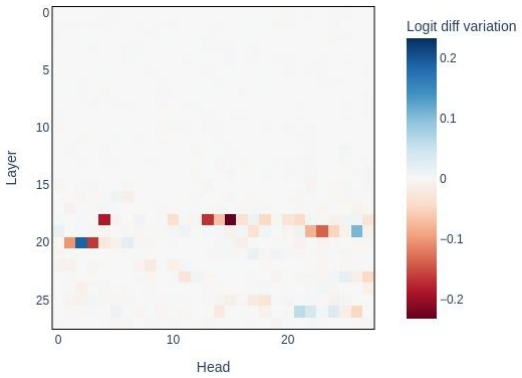


Figure 11: Path patching results on Qwen2-7B.

C.1 Path Patching results on LLaMA3 and Qwen2

Figure 10 and Figure 11 illustrate the path patching results on LLaMA3-8B and Qwen2-7B, respectively. Both visualized results align with the observations from that of LLaMA2-7B in Section 3.1.

Identified key attention heads of the two models are listed in Table 8, ranked by the absolute value of the causal effects in descending order.

C.2 Attention Patterns of Key Heads on LLaMA3 and Qwen2

For LLaMA3, we visualize the attention patterns of identified key heads (14, 5) and (14, 7), prioritizing temporal and structural tokens respectively. The visualization results are displayed in Figure 22 and Figure 23.

For Qwen2, we visualize the attention patterns of identified key heads (19, 23) and (20, 3), prioritizing temporal and structural tokens respectively. The visualization results are displayed in Figure 24 and Figure 25.

Index	Heads of LLaMA3	Heads of Qwen2
1	(15, 4)	(18, 13)
2	(14, 24)	(20, 2)
3	(14, 27)	(18, 15)
4	(15, 6)	(19, 23)
5	(14, 7)	(20, 3)
6	(14, 5)	(18, 4)
7	(15, 10)	(19, 26)

Table 8: Identified key attention heads on LLaMA3-8B and Qwen2-7B.

Model	Acc (%)		Cons (%)	
	Value	Δ	Value	Δ
LLaMA3-8B	91.43	+1.35	68.80	+1.08
Qwen2-7B	91.34	+2.16	89.09	+0.36

Table 9: Self-adaptive intervention results on LLaMA3-8B and Qwen2-7B.

C.3 Self-adaptive Intervention Results on LLaMA3 and Qwen2

The self-adaptive intervention results on LLaMA3-8B and Qwen2-7B are displayed in Table 9. Both models show consistent improvement in accuracy and consistency, further proving that the discovered LLMs' temporal reasoning mechanism is faithful and can be generalized across different model architecture.

D Experimental Settings

D.1 Prompts for Few-shot and CoT

Figure 12 and Figure 13 show the prompts that we utilize for few-shot and CoT baselines in Section 4.3, respectively.

D.2 Configurations for SFT

To establish a strong supervised fine-tuning (SFT) baseline and equip the LLaMA model with better temporal reasoning capabilities, we construct a SFT dataset derived from the MATRES dataset (Ning et al., 2018). MATRES adopts a multi-axis modeling approach and annotates temporal relations based on event start-points, which has been shown to achieve higher inter-annotator agreement.

To align the data with our objective of enhancing explicit event temporal reasoning, we exclude event pairs annotated as "Equal" or "Vague" and preserve only those with explicit "Before" or "After" relations. The structured annotations are then converted into natural language Question-Answering pairs to simulate our downstream event temporal reasoning task. Consequently, we construct 11,461 training samples from MATRES, and set a 20% ratio for validation set. We also provide a data instance from the training set in Table 4.

We perform full-parameter supervised fine-tuning on the LLaMA2-7B model using the LLaMA-Factory (Zheng et al., 2024) framework. The training was conducted for 3 epochs, and achieved a final validation loss of 0.274.

Prompt for the Few-shot Baseline

Article: After taking a shower, I brushed my teeth.
Question: Is taking a shower after brushing teeth?
The answer is: no

Article: I finished my report and then went to the gym.
Question: Is finishing the report before going to the gym?
The answer is: yes

Article: I checked my emails first thing in the morning, then attended the meeting.
Question: Is attending the meeting before checking emails?
The answer is: no

Figure 12: Prompt for the few-shot baseline.

Prompt for the CoT Baseline

Article: After taking a shower, I brushed my teeth.
Question: Is taking a shower after brushing teeth?
The answer is: The word "after" shows the sequence: shower → teeth brushing. So shower happened first. So the answer is no.

Figure 13: Prompt for the CoT baseline.

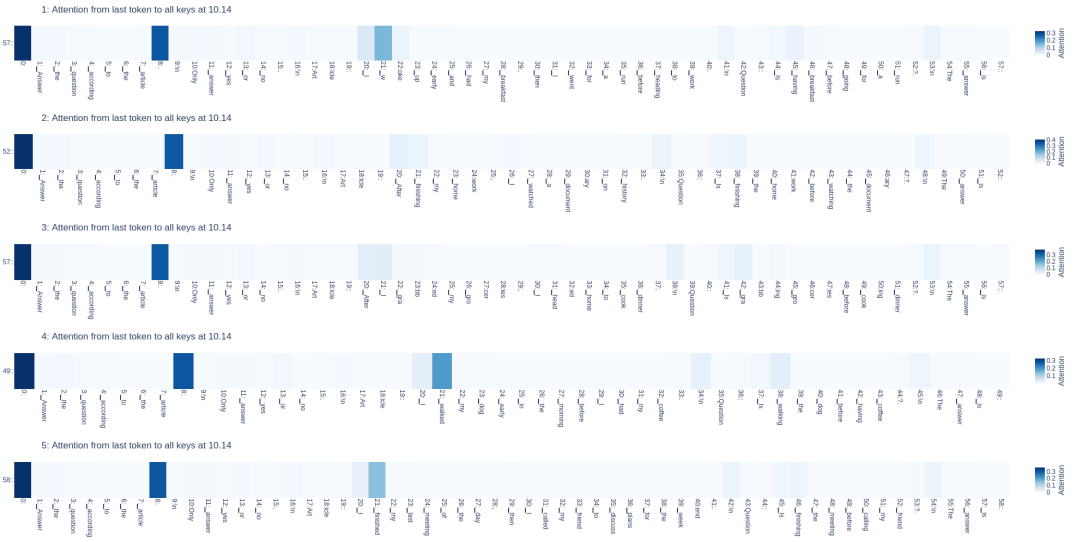


Figure 14: Attention patterns on head (10, 14), the head prioritizes effect-related tokens in the articles across five distinct data instances.

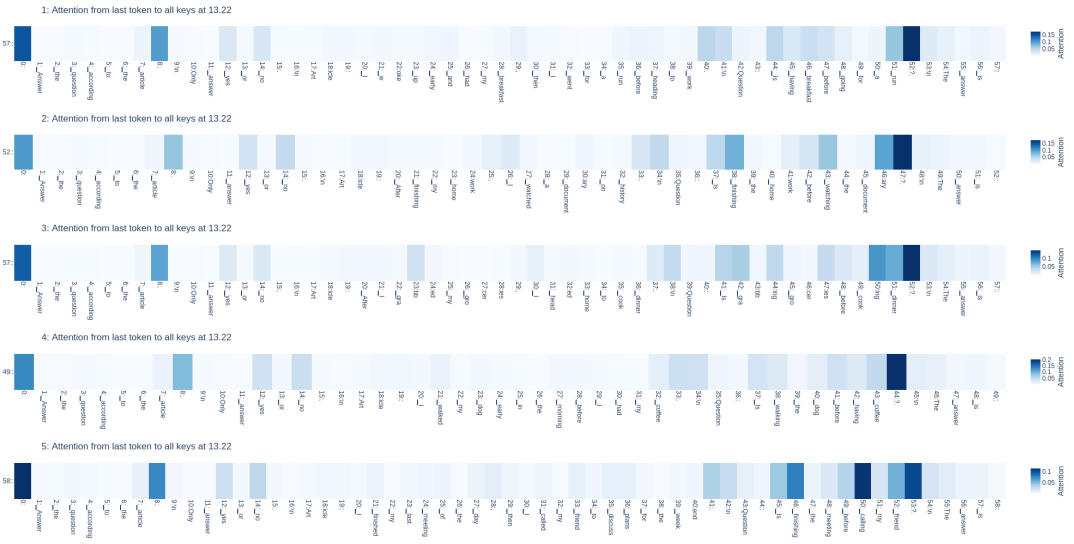


Figure 15: Attention patterns on head (13, 22), the head prioritizes effect-related tokens in the questions across five distinct data instances.

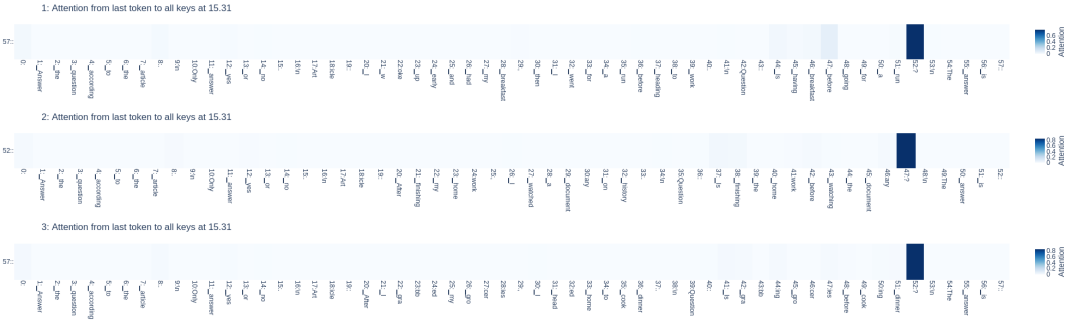


Figure 16: Attention patterns on key head (15, 31), prioritizing structural tokens.

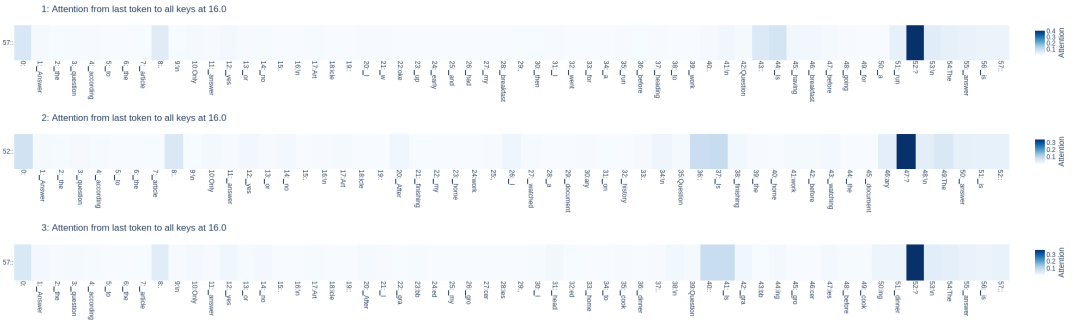


Figure 17: Attention patterns on key head (16, 0), prioritizing structural tokens.

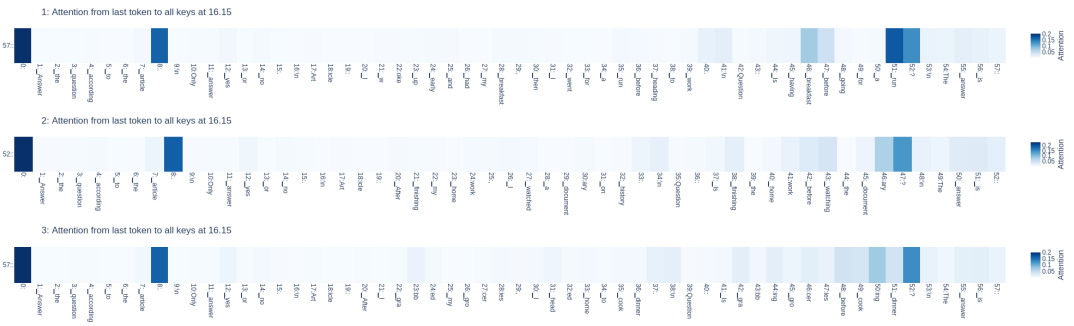


Figure 18: Attention patterns on key head (16, 15), prioritizing structural tokens.

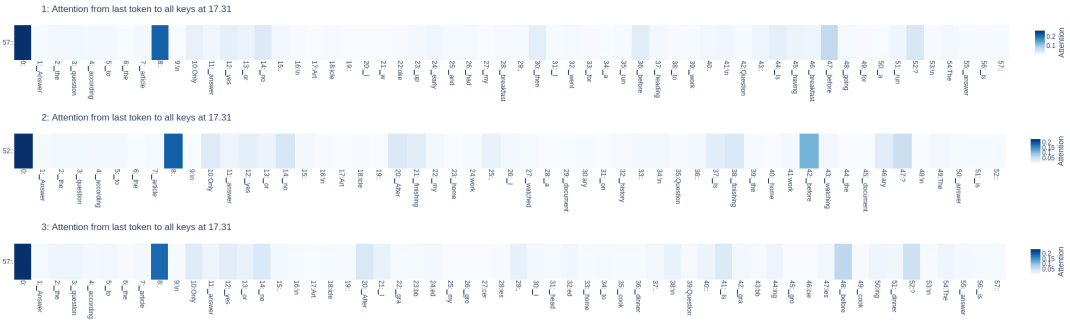


Figure 19: Attention patterns on key head (17, 31), prioritizing temporal tokens.

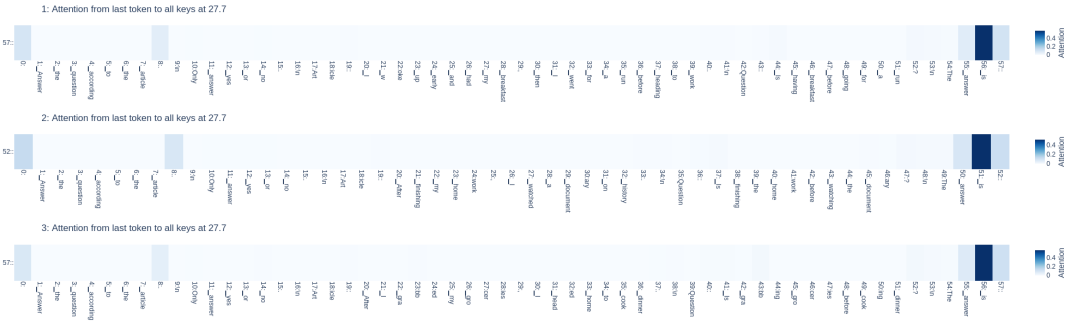


Figure 20: Attention patterns on key head (27, 7), prioritizing structural tokens.

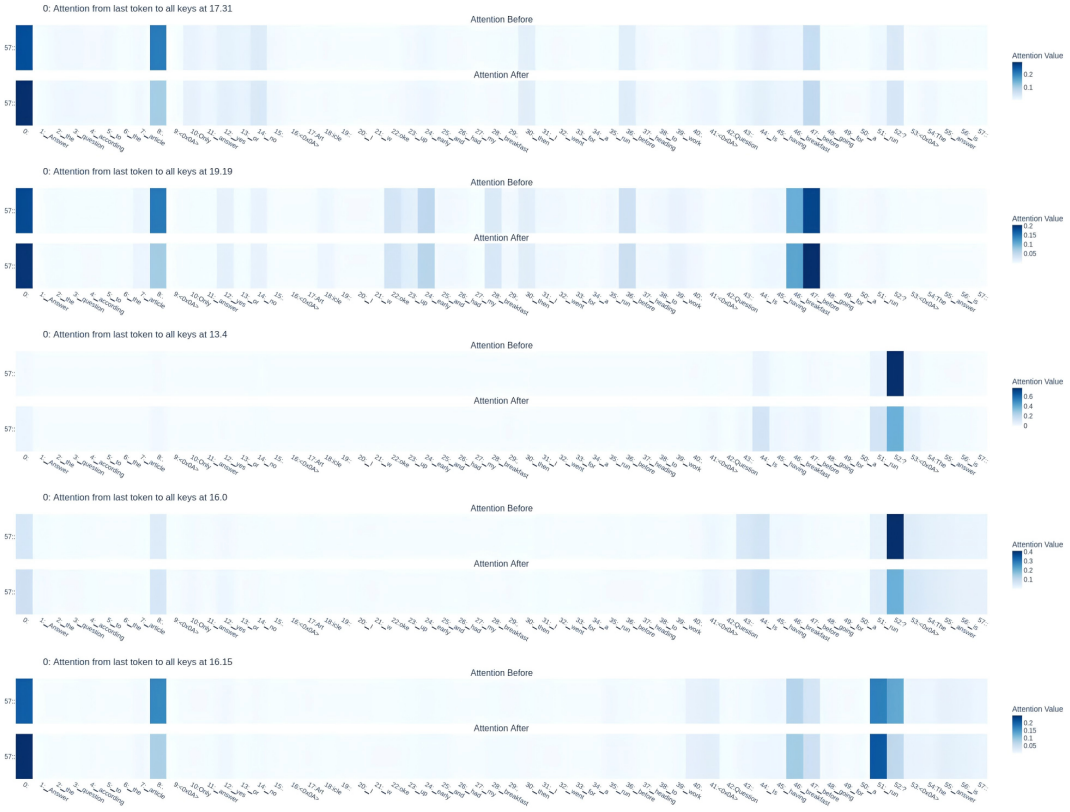


Figure 21: Attention distribution changes before and after reducing the attention on certain structural tokens, involving key heads (13, 4), (16, 0), (16, 15), (17, 31) and (19, 19) on the same data instance.

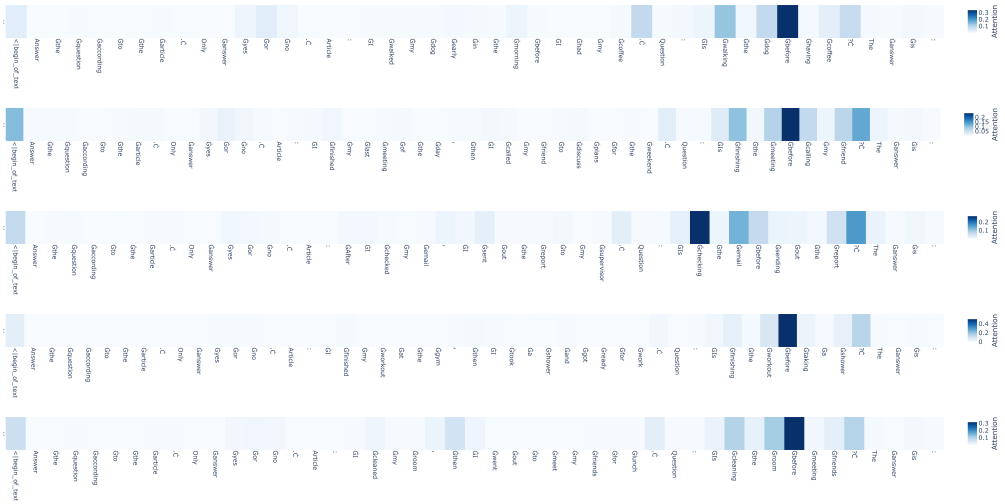


Figure 22: Attention patterns on key head (14, 5) of LLaMA3-8B, specifically attending to the temporal token (“before”) in the questions across five data instances.

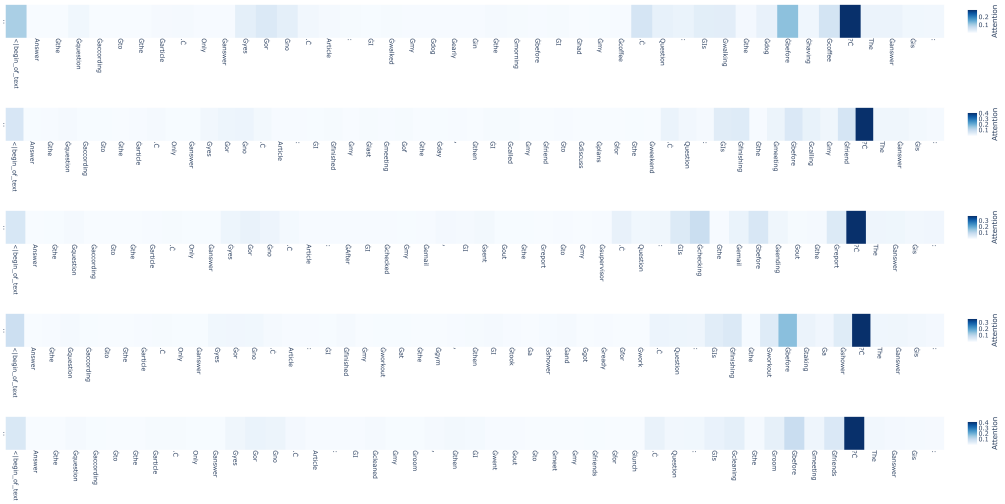


Figure 23: Attention patterns on key head (14, 7) of LLaMA3-8B, specifically attending to the structural token (“?”) in the questions across five data instances.

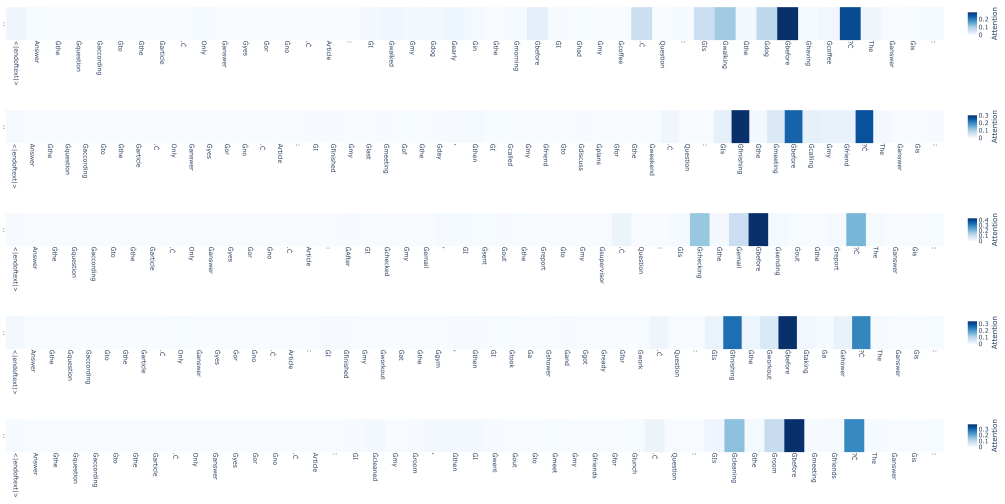


Figure 24: Attention patterns on key head (19, 23) of Qwen2-7B, specifically attending to the temporal token (“before”) in the questions across five data instances.

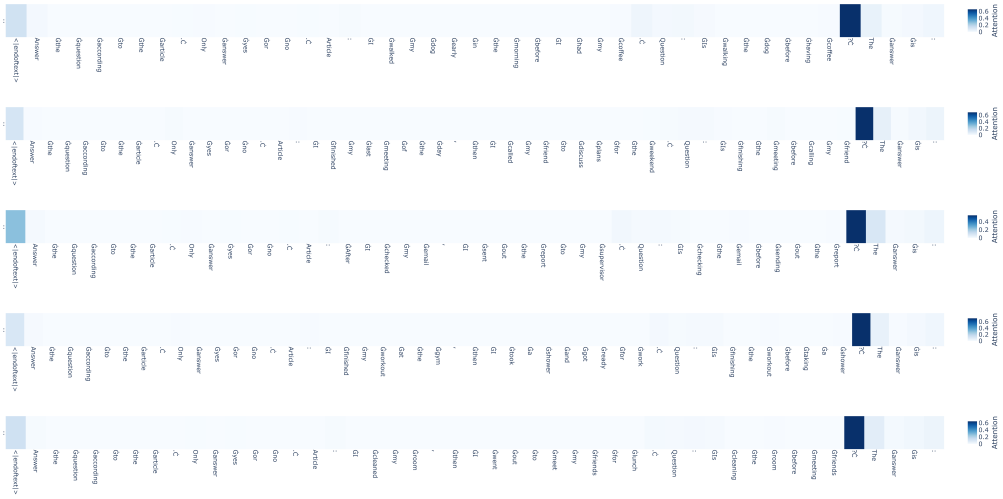


Figure 25: Attention patterns on key head (20, 3) of Qwen2-7B, specifically attending to the structural token (“?”) in the questions across five data instances.