Reading Attribution from Attention: Evidence Heads as Latent Attribution Mechanisms in LLMs

Qi Wu[1], Jianfeng Qu[2]*, Yanzhe Ji[3], Peng-Fei Zhang[4],
Siyu Li[2], Wei Chen[1], Zhixu Li[2,5], Jiajie Xu[1]
[1] School of Computer Science and Technology, Soochow University, Suzhou, China
[2] School of Smart Governance, Renmin University of China, Suzhou, China
[3] Australian National University, Canberra, Australia
[4] School of Computer Science and Technology, Tongji University, Shanghai, China
[5] School of Information, Renmin University of China, Beijing, China
wuqi7137@gmail.com, {jfqu, zhixuli}@ruc.edu.cn

Abstract

Large language models (LLMs) show strong performance in multi-document question answering, but their practical deployment is limited by unreliable and unfaithful attribution to supporting evidence. Existing prompting and training-based methods often suffer from hallucinated citations and lack interpretability in how evidence is selected. In this work, we investigate whether attribution signals are inherently encoded within transformer attention mechanisms. We introduce a sensitivity-based diagnostic that identifies a small subset of attention heads, termed **Evidence Heads**, which are highly responsive to perturbations in supporting documents. Building on this finding, we propose a training-free Attention-based Attribution framework that extracts evidence signals directly from Evidence Head attention using global and local strategies. Extensive experiments show that our method consistently outperforms strong baselines while remaining lightweight and interpretable. Further analyses confirm that Evidence Heads provide causally relevant and semantically grounded attribution signals rather than merely reflecting citation-generation behavior. Overall, our results suggest that structured attribution signals are implicitly encoded in LLM attention, and can be effectively leveraged for faithful multi-document reasoning.[2]

1 Introduction

Large language models (LLMs) have demonstrated strong capabilities [Brown et al., 2020, Lewis et al., 2020, Wei et al., 2022] in multi-document question answering, long-form generation, and knowledge-intensive reasoning. However, their practical deployment requires not only generating correct answers but also *faithfully attributing* them to supporting evidence [Madsen et al., 2022, Ferrando et al., 2023]. That is, models must identify which external documents justify each generated statement, as illustrated in Figure 1. This requirement is critical for high-stakes applications [Thorne et al., 2018, Ouyang et al., 2022] such as retrieval-augmented generation and factual verification, where unverifiable outputs undermine reliability.

Despite growing interest in answer attribution, existing approaches largely treat it as an auxiliary generation task. Methods that prompt LLMs to produce citations alongside answers [Sun et al., 2023, Li et al., 2024] often suffer from hallucinated references, formatting inconsistencies, and

*Corresponding author
[2]Our code, dataset, and the models used in this work are available at https://github.com/Xuan-Van/Attention-based-Attribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

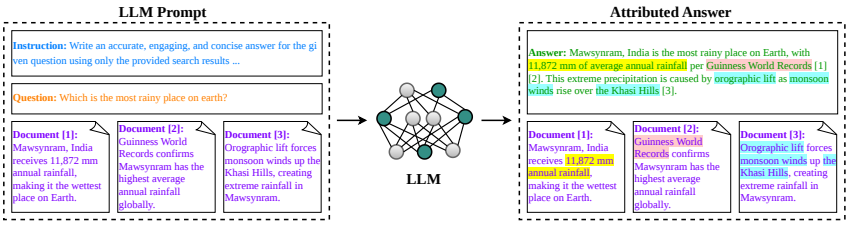

Figure 1: Overview of answer attribution for faithful generation. The model must generate answers grounded in retrieved documents and explicitly identify which document spans support each statement.

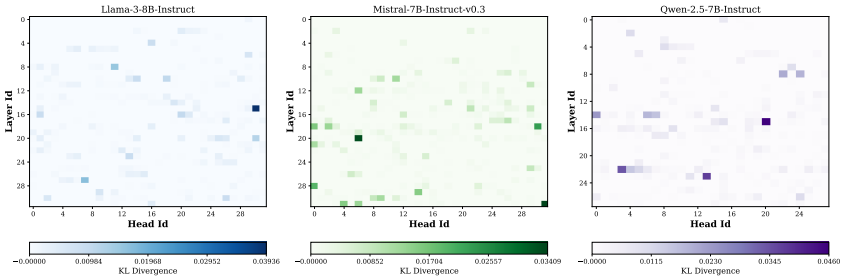

Figure 2: Head-wise distribution of Evidence Heads across three LLMs, showing variations in both the number of heads and their layer-wise allocation. The score refers to the Kullback–Leibler (KL) divergence between original and perturbed attention distributions.

weak grounding because generation and attribution are entangled [Huang et al., 2024c, Cohen-Wang et al., 2024]. Instruction-tuning approaches [Chuang et al., 2025, Yu et al., 2026] partially alleviate these issues but rely on costly annotated data and multi-stage training, while still lacking an explicit mechanism that explains how attribution is internally formed. As a result, attribution remains an emergent behavior rather than a controllable and interpretable process.

This raises a fundamental question: *can attribution be derived directly from the internal mechanisms of LLMs, rather than being imposed externally?* Recent studies on transformer interpretability [Elhelo and Geva, 2025, Yang et al., 2026] have shown that attention heads exhibit functional specialization, capturing patterns such as token copying, temporal reasoning, and syntactic alignment [Wu et al., 2025, Park et al., 2025, BINBINLIU et al., 2026]. Motivated by these findings, we ask: *do specialized attention heads responsible for evidence identification exist, and can they serve as a latent mechanism for answer attribution?*

To investigate this hypothesis, we introduce a sensitivity-based diagnostic that probes how attention behaviors change under evidence perturbations. Specifically, we compare attention distributions between original and corrupted contexts and quantify their divergence to identify heads that are highly responsive to disruptions in supporting evidence. This analysis reveals a small subset of attention heads that consistently exhibit strong sensitivity to evidence-related changes (Figure 2). We refer to this set of heads as **Evidence Heads**.

Building on this insight, we propose a training-free Attention-based Attribution framework that directly extracts attribution signals from Evidence Head attention. Instead of generating citations explicitly, our approach aggregates token-level attention into evidence-level scores, reflecting the strength of grounding between generated content and candidate documents. To accommodate the dynamic nature of evidence usage, we design both global and local attribution strategies, which respectively apply shared thresholds or adapt to sentence-specific distributions, enabling a principled balance between precision and recall. Experimental results demonstrate that our framework consistently outperforms strong baselines across multiple datasets, showing that internal attention signals can serve as a reliable source for answer attribution without additional training.

To probe the mechanism behind these performance gains, targeted head ablation and activation patching reveal that interventions on Evidence Heads systematically alter attribution behavior, sup-

porting the causal role of Evidence Heads. Their attention scores also align with natural language inference (NLI)-based sentence–document entailment signals, suggesting that Evidence Heads track semantically relevant evidence. Moreover, attribution remains informative when generated citation markers are removed and even without an explicit citation prompt, indicating that Evidence Heads capture more than citation-generation behavior. Together, these analyses support the use of Evidence Heads as an internal source of attribution signals.

Overall, our work provides a new perspective on answer attribution by uncovering a latent mechanism within LLMs. We show that attribution is not merely an external generation problem, but is partially governed by structured attention patterns that can be identified, utilized, and validated. These findings bridge interpretability and factual grounding, and suggest that carefully analyzing internal model dynamics can lead to practical and interpretable solutions for faithful multi-document reasoning.

2 Related Work

2.1 Answer Attribution

Answer attribution [Rashkin et al., 2023, Muller et al., 2023] aims to identify which supporting documents are utilized by LLMs when generating each sentence of an answer in multi-document question answering settings. Early approaches [Sun et al., 2023, Jain et al., 2023, Weller et al., 2024] rely on prompt engineering, where models are instructed to generate answers together with explicit citation markers. While simple and widely adopted, these methods suffer from several well-known limitations [Liu et al., 2023, Zuccon et al., 2023, Peskoff and Stewart, 2023], including formatting inconsistencies, missing citations, and hallucinated references. More fundamentally, such self-citations are not guaranteed to reflect the model's internal reasoning, as they are generated in the same autoregressive decoding procedure rather than derived from explicit grounding signals.

To address these issues, instruction-tuning methods [Schimanski et al., 2024, Phukan et al., 2024, Huang et al., 2024b] incorporate citation generation and introduce additional routing and alignment modules. Although effective, these approaches [Huang et al., 2024a, Chuang et al., 2025] require costly annotated data and additional training stages, limiting their applicability in resource-constrained settings. In contrast to these methods, our work investigates whether attribution signals can be directly recovered from internal attention mechanisms, without requiring additional training or hidden-state computation.

2.2 Mechanistic Interpretability of LLMs

Mechanistic interpretability seeks to explain LLM behavior by identifying its internal features, components, and circuits [Rai et al., 2024, Atakishiyev et al., 2025, Somvanshi et al., 2026]. Early work developed frameworks for analyzing information flow in Transformer circuits [Elhage et al., 2021], while later studies traced task-specific computations and localized factual associations in GPT models [Wang et al., 2022, Meng et al., 2022, Conmy et al., 2023]. Other studies characterize feed-forward layers as key–value memories [Geva et al., 2021]. Sparse autoencoders have been used to extract interpretable features from model activations, and this approach has been applied at scale to Claude 3 Sonnet [Huben et al., 2024, Templeton et al., 2026]. Together, these studies examine model behavior at the levels of features, components, and circuits, using both representational analysis and causal interventions. Attention heads are a useful unit of analysis because their outputs can be isolated and intervened on, enabling direct tests of their functional contributions.

2.3 Interpretability of Attention Heads

A substantial body of work [Zheng et al., 2025, Han et al., 2025, Qu et al., 2025] has investigated the interpretability of attention heads in transformer-based language models. Early studies [Zhang and Nanda, 2024, Tan et al., 2025] primarily focus on probing head importance through ablation-based techniques, including zero-ablation [Wang et al., 2023, Mohebbi et al., 2023, Yao et al., 2024, Jin et al., 2024] and mean-ablation [Hanna et al., 2023, McDougall et al., 2024, Kim et al., 2025], where specific attention heads are removed or replaced with baseline activations to measure performance change. Complementary approaches [Merullo et al., 2024a,b, Wiegreffe et al., 2025] such as activation patching replace intermediate activations with those from corrupted or alternative inputs, enabling causal assessment of functional components within the network.

Recent studies [Elhelo and Geva, 2025, Yang et al., 2026, Sandoval-Segura et al., 2026] collectively demonstrate that attention heads are not homogeneous but instead exhibit functional specialization during pretraining. In particular, prior work [Wu et al., 2025, Park et al., 2025, BINBINLIU et al., 2026] has identified distinct roles such as token copying, temporal recognition, and syntactic alignment in translation. Motivated by these findings, we investigate whether the attention patterns of specific attention heads during generation can serve as latent signals for evidence selection, providing a new perspective on leveraging internal model mechanisms for explainable multi-document reasoning.

3 Method

3.1 Problem Formulation of Answer Attribution

We formalize the answer attribution problem as follows. Let I denote the task instruction and q denote the user query. Given a document set $D = \{d_1, d_2, \ldots, d_m\}$ and a model-generated answer composed of n sentences $S = \{s_1, s_2, \ldots, s_n\}$, the goal of answer attribution is to identify, for each sentence s_i, a set of supporting documents $c_i \subseteq \{1, 2, \ldots, m\}$, where c_i may be empty.

Formally, we define the attribution function as:

$$\mathcal{A} : (I, q, D, S) \rightarrow C = \{c_1, c_2, \ldots, c_n\}, \quad c_i \subseteq \{1, 2, \ldots, m\}. \tag{1}$$

Each c_i represents the set of document indices in D that provide evidential support for the corresponding sentence s_i. If no document in D supports s_i, then $c_i = \varnothing$.

3.2 Identification of Evidence Heads

To identify attention heads that are potentially involved in answer attribution, we hypothesize that such heads are sensitive to the semantic integrity of supporting evidence. Specifically, when key information in the documents is corrupted, these attention heads should exhibit noticeable changes in their attention patterns.

For each sample, we construct a perturbed document set D' by replacing key entity tokens in the original document set D with length-matched unrelated entity tokens. This perturbation is designed to disrupt the underlying evidence structure while preserving surface-level fluency.

We feed (I, q, D, S) and (I, q, D', S) into the model to obtain the corresponding original and perturbed attention maps, and let $\mathcal{T}(S)$ denote the token indices of the answer. For each attention head h, we extract a submatrix of the attention weights focusing on interactions between answer tokens and entity tokens. Let $\mathcal{T}_e(D)$ denote the token indices of key entity tokens in D, and let $\mathcal{T}_{e'}(D')$ denote the token indices of unrelated entity tokens in D'. We define the corresponding attention distributions as:

$$P^h = A^h_{\mathcal{T}(S), \mathcal{T}_e(D)}, \quad Q^h = A^h_{\mathcal{T}(S), \mathcal{T}_{e'}(D')}. \tag{2}$$

To quantify the sensitivity of each attention head to the perturbation, we compute the KL divergence between the two attention distributions:

$$\mathrm{KL}(P^h \,\|\, Q^h) = \sum_i \sum_j P^h_{i,j} \log \frac{P^h_{i,j}}{Q^h_{i,j}}. \tag{3}$$

A higher KL divergence indicates that the attention pattern of head h changes more significantly when the supporting evidence is disrupted, suggesting a stronger involvement in evidence processing. All experiments are conducted on the HotpotQA dataset [Yang et al., 2018] using three models: Llama-3.1-8B-Instruct [Patterson et al., 2022] (32 layers, 32 attention heads per layer), Mistral-7B-Instruct-v0.3 [AI, 2024] (32 layers, 32 attention heads per layer), and Qwen-2.5-7B-Instruct [Yang et al., 2024, Team, 2024] (28 layers, 28 attention heads per layer).

As shown in Figure 2, only a small subset of attention heads exhibit notable sensitivity to perturbations in supporting evidence, indicating that evidence processing is selectively concentrated in specific heads. We rank all attention heads according to their KL divergence scores and select the top 1% as **Evidence Heads**. The perturbation protocol and rationale for this ratio are described in Appendix B, and the specific Evidence Head indices are reported in Table 3.

3.3 Aggregating Token-to-Sentence Evidence

After the model generates the final answer, we extract attention weights from a set of Evidence Heads H. For each head $h \in H$, we obtain a full token-level attention matrix $A^h \in \mathbb{R}^{T \times T}$, where T is the total number of tokens in the input sequence. However, these raw token-level attention signals are typically noisy and difficult to interpret directly.

From a modeling perspective, sentence-level attribution can be viewed as a structured sparsification of token-level interactions: each sentence is usually supported by only a small subset of relevant tokens, which are distributed across one or a few related documents. Therefore, evidence is not uniformly distributed over all token pairs, but instead manifests as sparse, high-confidence local alignments between specific sentence and document spans, as shown in Figure 8 in Appendix E.

Motivated by this observation, we consider interactions between each sentence s_i and each document d_j, where $\mathcal{T}(s_i)$ and $\mathcal{T}(d_j)$ denote the corresponding sets of token indices. To map token-level attention to the sentence–document level, we adopt max pooling over the token-level attention scores within each sentence–document pair, which effectively highlights localized evidence spikes:

$$A_{S,D}^{h,\max}(i,j) = \max_{t \in \mathcal{T}(s_i),\ t' \in \mathcal{T}(d_j)} A_{t,t'}^{h}. \tag{4}$$

Here, $A_{S,D}^{h,\max} \in \mathbb{R}^{n \times m}$ denotes the sentence–document attention matrix for head h, with examples of the corresponding attention distributions shown in Figure 9 in Appendix E. Further, we aggregate signals across Evidence Heads to obtain the final sentence–document level attribution score:

$$A_{S,D}(i,j) = \frac{1}{|H|} \sum_{h \in H} A_{S,D}^{h,\max}(i,j), \tag{5}$$

where $A_{S,D} \in \mathbb{R}^{n \times m}$ and $A_{S,D}(i,j)$ denotes the aggregated attribution score between sentence s_i and document d_j.

3.4 Adaptive Attribution Strategy

Since the number of supporting documents per sentence is inherently dynamic, fixed Top-k selection imposes a strong inductive bias by assuming uniform evidence sparsity. This often results in over-selection for low-evidence sentences and under-selection for high-evidence sentences. To address this issue, we directly use continuous attribution scores $A_{S,D}(i,j)$ to identify supporting documents in a data-adaptive way, rather than relying on discrete ranking cutoffs. We propose two complementary attribution strategies: global attribution and local attribution.

Global attribution assumes that evidence scores across different sentences are comparable, implying that the model induces a shared attention scale during generation. Under this assumption, a single global threshold is applied across all sentence-document pairs:

$$c_i^{\text{global}} = \{ j \mid A_{S,D}(i,j) \geq \mu + \alpha \cdot \sigma \}, \tag{6}$$

where μ and σ denote the mean and standard deviation of all scores in $A_{S,D}$, respectively, and α (default 1.0) is a scaling hyperparameter that controls the selection threshold.

Local attribution relaxes the assumption of global comparability and allows sentence-specific evidence calibration. This is motivated by the observation that different sentences may have heterogeneous evidence requirements: some rely on multiple supporting documents, whereas others may rely on a single strong signal or no external evidence at all. To accommodate this variability, we define a sentence-specific local threshold for each sentence s_i, allowing the attribution decision to adapt to its individual evidence distribution:

$$c_i^{\text{local}} = \{ j \mid A_{S,D}(i,j) \geq \mu_i + \alpha \cdot \sigma_i \}, \tag{7}$$

where $\mu_i = \mu(\{A_{S,D}(i,j)\}_{j=1}^{m})$ and $\sigma_i = \sigma(\{A_{S,D}(i,j)\}_{j=1}^{m})$ denote the mean and standard deviation of attribution scores associated with sentence s_i, respectively. This formulation allows the threshold to adapt to sentence-specific evidence distributions, enabling more flexible and robust document selection under heterogeneous grounding conditions.

Overall, by leveraging max-pooled evidence signals and adaptive thresholding strategies, we eliminate the need for external training or fixed-k assumptions. This results in a lightweight yet effective attribution framework that directly exposes latent evidence selection behavior within LLMs. The full attribution procedure is summarized in Algorithm 1 in Appendix C.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method on six widely used multi-document question answering datasets, covering both single-hop and multi-hop reasoning scenarios: ASQA [Stelmakh et al., 2022], ELI5 [Fan et al., 2019], 2WikiMultiHopQA (2WikiMQA) [Ho et al., 2020], Natural Questions (NQ) [Kwiatkowski et al., 2019], PopQA [Mallen et al., 2023], and TriviaQA [Joshi et al., 2017].

Baselines. We compare our approach with the following baselines: **Vanilla** [Gao et al., 2023b] directly prompts the LLM to generate answers with citations without any attribution mechanism. **Summary** [Gao et al., 2023b] first summarizes each retrieved document and then asks the model to generate answers with citations based on the summaries. **Snippet** [Gao et al., 2023b] provides truncated evidence snippets along with the prompt to encourage grounded generation. **MIRAGE** [Qi et al., 2024] estimates fine-grained token-level contributions by simulating predictive distributions under counterfactual contexts. **C^2-Cite** [Yu et al., 2026] integrates instruction-tuning alignment modules and gating modules to encode document-level semantics into specialized citation tokens.

Evaluation. We assess attribution quality using the Automatic AI-based Information Score (AutoAIS) [Gao et al., 2023a]. AutoAIS formulates attribution evaluation as an NLI task: each generated answer sentence is paired with its cited evidence to determine whether the evidence entails the claim. The evaluation setting follows the ALCE benchmark [Gao et al., 2023b]. More experimental details are provided in Appendices C and D.

4.2 Main Results

Table 1: F1 scores for our two attribution strategies (global and local) and the baselines across all datasets. The best results are shown in **bold** and the second-best results are underlined. Evidence Heads are fixed per model across all datasets.

Methods	ASQA	ELI5	2WikiMQA	NQ	PopQA	TriviaQA	Overall
Llama-3.1-8B-Instruct							
Vanilla	67.49	49.10	31.53	61.31	64.68	53.05	54.41
Summary	65.21	45.32	30.46	57.72	60.27	52.78	51.85
Snippet	66.26	48.29	29.98	58.91	61.34	49.60	52.32
MIRAGE	66.22	47.41	30.30	61.69	64.86	53.63	53.91
C^2-Cite	67.82	49.19	32.25	61.62	65.09	53.94	54.87
Ours (global)	67.73	**51.41**	**34.75**	66.88	67.05	56.19	57.24
Ours (local)	**71.04**	51.11	32.39	**69.18**	**67.97**	**58.39**	**58.24**
Mistral-7B-Instruct-v0.3							
Vanilla	23.27	40.54	14.17	39.01	37.76	33.11	31.38
Summary	20.32	39.79	14.69	39.57	32.92	33.38	30.03
Snippet	28.18	36.97	18.91	36.53	38.95	35.83	32.45
MIRAGE	40.69	44.45	21.06	54.45	48.92	46.33	42.57
C^2-Cite	46.82	45.32	22.87	58.87	49.03	46.02	44.72
Ours (global)	64.32	**52.60**	**25.48**	65.70	60.05	52.44	53.34
Ours (local)	**65.51**	51.69	24.78	**66.68**	**60.58**	**53.65**	**53.72**
Qwen-2.5-7B-Instruct							
Vanilla	59.90	40.09	27.12	58.79	54.66	48.74	48.11
Summary	60.07	44.30	25.45	57.13	52.78	44.61	47.28
Snippet	61.22	43.73	26.26	57.05	56.54	46.48	48.44
MIRAGE	60.66	40.88	27.11	59.51	55.40	49.09	48.66
C^2-Cite	60.56	41.06	27.82	58.95	54.96	49.40	48.68
Ours (global)	64.22	44.90	**28.42**	65.82	61.12	53.80	52.95
Ours (local)	**68.46**	**46.69**	27.31	**68.40**	**65.02**	**55.14**	**55.05**

As shown in Table 1, our method consistently outperforms all baselines across datasets, demonstrating its effectiveness in the answer attribution task. It significantly outperforms prompt-based approaches, indicating that post-hoc generation alone is insufficient for faithful attribution. It also achieves superior performance compared to prior methods without requiring hidden-state computation or additional training.

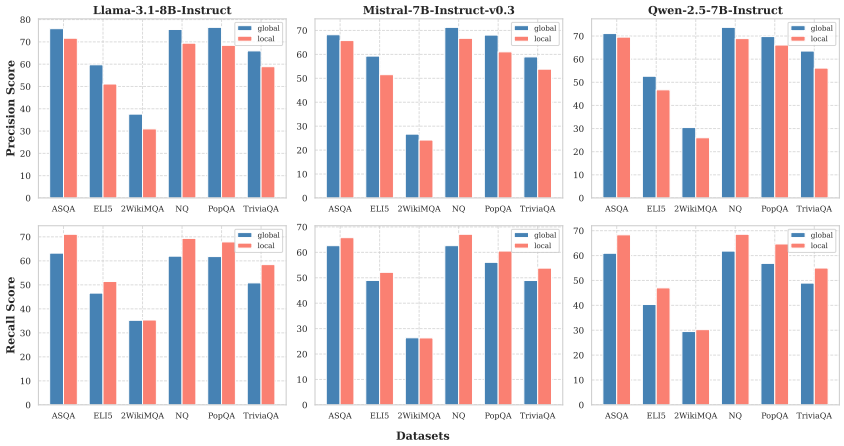

Figure 3: Bar charts showing precision (top row) and recall (bottom row) of different models on multiple datasets under two attribution strategies (global and local).

Figure 3 further illustrates the precision-recall trade-off of different attribution strategies. The global strategy achieves higher precision, producing more accurate evidence selection, while the local strategy yields higher recall, covering a broader set of relevant supporting documents. This reflects a trade-off between conservative selection and broad coverage at different levels of attribution granularity.

Overall, these results demonstrate that Evidence Heads exhibit strong cross-dataset generalization and provide a stable basis for answer attribution. By quantifying their attention signals, our method enables effective attribution without additional training, offering a transparent and consistent view of how LLMs leverage supporting documents during generation.

4.3 Ablation Study

We conduct a series of ablation studies to better understand the key components of our method. All experiments are performed using Llama-3.1-8B-Instruct on the ASQA dataset.

Layer-wise Ablation. We investigate the contribution of different Transformer layers by treating all attention heads within a given layer as Evidence Heads and evaluating attribution performance accordingly. As shown in Figure 4 (left), attribution performance peaks around layers 15–25, while both lower and higher layers are relatively less effective. This trend is consistent with the head-wise distribution of Evidence Heads, suggesting that signals responsible for answer attribution are primarily concentrated in the intermediate layers of the model.

Effect of α. We evaluate the scaling hyperparameter α under both global attribution and local attribution strategies. As shown in Figure 4 (middle), both strategies maintain stable and strong performance when $\alpha \in [0, 1]$. However, performance drops significantly when $\alpha > 1$, and the decline is more pronounced for the local attribution strategy. This suggests that the evidence scores exhibit a well-behaved distribution within a moderate scaling range and supports our design of modeling attribution decisions based on distribution-aware normalization at both global and local levels.

Effect of Pooling Granularity. We analyze the effect of different pooling granularities by varying the proportion τ of top-attended tokens used to aggregate attention scores. Specifically, we replace Equation 4 with:

$$A^{h,\tau}_{S,D}(i,j) = \frac{1}{|\Omega^{\tau}_{ij}|} \sum_{(t,t') \in \Omega^{\tau}_{ij}} A^{h}_{t,t'}, \quad \Omega^{\tau}_{ij} = \mathrm{Top}_{\tau}\left(\{A^{h}_{t,t'} \mid t \in \mathcal{T}(s_i),\, t' \in \mathcal{T}(d_j)\}\right). \tag{8}$$

Here, $\tau \in (0, 1]$. When $\tau = 1$, the aggregation reduces to mean pooling over all tokens. As $\tau \to 0$, the aggregation becomes increasingly close to max pooling.

Figure 4 (right) shows that the quality of answer attribution gradually decreases as τ increases. When $\tau > 0.5$, the performance tends to plateau due to noise dilution. This further indicates that our

attention aggregation strategy is reasonable and effective in balancing signal preservation and noise suppression.

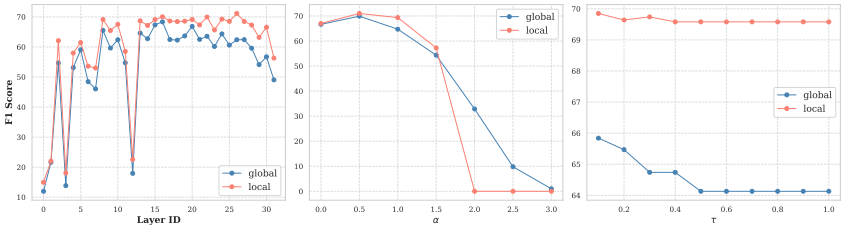

Figure 4: Ablation experiments based on global and local attribution strategies. (Left) Layer-wise attribution performance across Transformer layers. (Middle) Effect of scaling hyperparameter α. (Right) Effect of pooling granularity controlled by τ for attention aggregation.

Discussion. Overall, these ablation results highlight three key findings: (1) attribution signals are mainly concentrated in middle layers of the model, (2) stable performance relies on a properly bounded scaling hyperparameter $\alpha \in [0, 1]$, and (3) finer-grained pooling (smaller τ) consistently yields better attribution, confirming the advantage of more selective, high-saliency token aggregation over coarse averaging.

5 Analysis

5.1 Causal Role in Answer Attribution

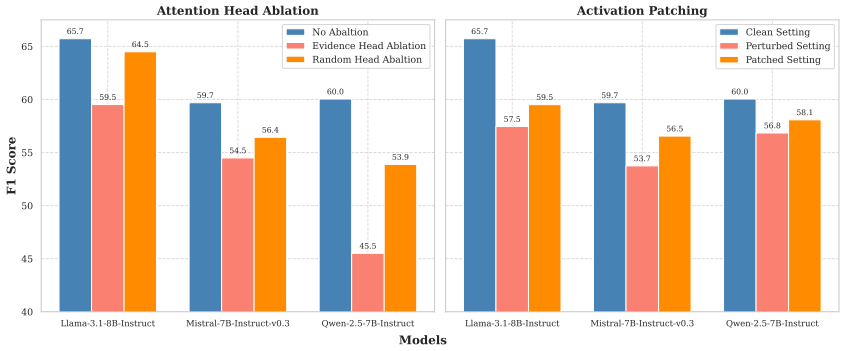

Figure 5: Causal effect of Evidence Heads on answer attribution. Left: attribution F1 scores under no ablation, Evidence Head ablation, and random head ablation. Right: attribution F1 scores under the clean, perturbed, and patched setting using Evidence Head activations.

To better understand the performance gains reported above, we examine whether the Evidence Heads identified in Section 3.2 encode meaningful signals for answer attribution. Specifically, we intervene on these attention heads through ablation or activation patching and evaluate the resulting self-citing outputs to quantify the effects of these interventions on attribution. The experiments are conducted on 500 random ASQA samples, and the evaluation follows the same setting as Section 4.1.

Attention Head Ablation. We investigate the causal importance of Evidence Heads by comparing their attention output ablation (setting the corresponding head activations to zero) against both no ablation and random ablation of the same number of attention heads. The answer attribution F1 score is evaluated using model outputs obtained under these conditions, in order to quantify how selectively ablating Evidence Heads affects attribution performance compared to other ablation conditions.

Activation Patching. To further verify causality, we conduct activation patching experiments to test whether Evidence Head activations directly capture attribution-relevant information. We evaluate three conditions: (1) a clean setting using D, (2) a perturbed setting using D', and (3) a patched setting where Evidence Head activations computed from the clean setting are injected into the model

during inference with D'. In the patched setting, only the Evidence Head activations are restored from the clean setting, while all other components of the model operate on the perturbed input. This allows us to isolate the contribution of these heads to attribution behavior.

As shown in Figure 5 (left), ablating Evidence Heads leads to a significant drop in attribution F1 score, while ablating random attention heads results in a smaller decrease, indicating that Evidence Heads play a critical role in answer attribution. Figure 5 (right) shows that replacing key entity tokens in the input leads to a noticeable drop in attribution F1 scores. However, restoring the Evidence Head activations substantially recovers the performance, suggesting that these heads provide robust and intervention-sensitive signals closely associated with evidence selection.

5.2 Entailment Alignment of Evidence Heads

Building on the causal findings above, we investigate the relationship between the attention signals of Evidence Heads and semantic grounding, in order to further assess whether these heads align with interpretable evidence selection mechanisms.

Since the attribution task is defined at the sentence–document level, we construct an appropriate semantic supervision signal by using a NLI model to estimate entailment relations between sentences and documents. Specifically, we compute the entailment probability between each sentence and each document, forming a matrix $E_{S,D} \in \mathbb{R}^{n \times m}$, where $E_{S,D}(i, j)$ denotes the probability that sentence s_i is supported by document d_j, with n being the number of sentences and m being the number of documents.

In addition to max pooling, we use mean pooling to determine whether the observed alignment depends on the pooling strategy:

$$A_{S,D}^{h,\text{mean}}(i,j) = \frac{1}{|\mathcal{T}(s_i)| \cdot |\mathcal{T}(d_j)|} \sum_{t \in \mathcal{T}(s_i)} \sum_{t' \in \mathcal{T}(d_j)} A_{t,t'}^{h}. \tag{9}$$

Here, $A_{S,D}^{h,\text{mean}} \in \mathbb{R}^{n \times m}$ denotes the mean-pooled sentence–document attention matrix for head h, with an example of the corresponding attention distribution shown in Figure 10 in Appendix E. We then compute correlation metrics between attention and entailment:

$$\rho_{\text{Pearson}}\left(A_{S,D}^{h,*}, E_{S,D}\right),\ \rho_{\text{Spearman}}\left(A_{S,D}^{h,*}, E_{S,D}\right),\ \text{AUC}\left(A_{S,D}^{h,*}, E_{S,D}\right),\ * \in \{\max, \text{mean}\}. \tag{10}$$

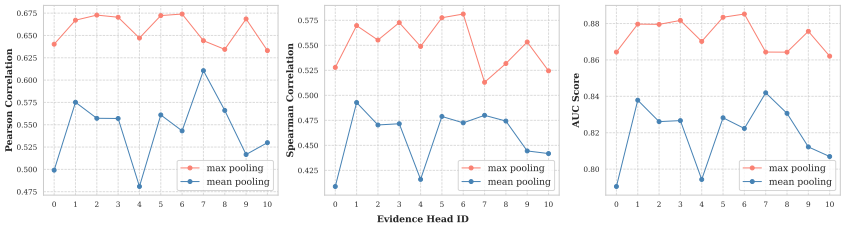

Figure 6: Correlation results for Llama-3.1-8B-Instruct between Evidence Head attention and NLI-based entailment scores. The figure compares different pooling strategies (max and mean) and evaluates alignment for each Evidence Head using Pearson correlation (left), Spearman correlation (middle), and AUC score (right). The x-axis follows the model's native (layer ID, head ID) order.

Figure 6 reports results for Llama-3.1-8B-Instruct. Across all three evaluation metrics, max pooling yields stronger alignment than mean pooling for every Evidence Head and shows lower variance across heads. From the perspective of inter-head dynamics, different Evidence Heads show noticeable variation in their alignment strengths, yet no single head dominates overall performance, suggesting that evidence modeling is distributed rather than concentrated. Overall, these results show that Evidence Head attention, particularly under max pooling, aligns closely with NLI-derived entailment signals in a stable and distributed manner, supporting its use as an interpretable evidence signal for answer attribution.

5.3 Separating Evidence Attribution from Citation Behavior

An attribution method evaluated on citation-bearing answers could appear successful simply by reusing the generated citation markers, rather than recovering evidence from the model's internal signals. Citation instructions may also shape answer generation and the associated evidence-related signals. To distinguish these effects, we compare three conditions on Qwen-2.5-7B-Instruct. The **original condition** uses answers generated with the citation prompt and retains their markers. In the **citation-markers-removed condition**, we strip markers from those same answers before attribution, testing whether the method relies on the markers. In the **no-citation-prompt condition**, we generate answers without the explicit citation requirement and apply the same attribution pipeline, testing whether evidence-related signals remain without that instruction. The prompt template for the no-citation-prompt condition is provided in Appendix Table 6.

Table 2: Attribution F1 scores on Qwen-2.5-7B-Instruct under three citation conditions.

Condition	ASQA	ELI5	2WikiMQA	NQ	PopQA	TriviaQA	Overall
Global							
Original	64.22	44.90	28.42	65.82	61.12	53.80	52.95
Citation markers removed	66.01	46.95	30.16	67.02	62.74	54.37	54.44
No citation prompt	61.17	40.65	27.40	59.91	57.24	50.34	49.35
Local							
Original	68.46	46.69	27.31	68.40	65.02	55.14	55.05
Citation markers removed	69.70	47.54	28.72	68.50	65.93	55.66	55.89
No citation prompt	64.04	39.47	25.72	62.23	59.81	51.89	50.41

As shown in Table 2, removing markers slightly improves attribution performance under both strategies, indicating that the method does not simply reuse the generated citations. Omitting the citation prompt lowers performance, but attribution remains informative. This suggests that citation instructions strengthen the observed signal, while Evidence Heads still capture useful evidence-related information without explicit citation guidance.

6 Conclusion

In this work, we investigate whether **Evidence Heads** in LLMs can serve as a latent mechanism for answer attribution. Evidence perturbations identify a small set of sensitive attention heads, which we use in a training-free attribution framework to infer answer–document support through complementary global and local strategies. Extensive experiments across multiple datasets and model families show that our method achieves strong attribution performance while remaining lightweight and interpretable. Causal interventions, NLI-based alignment, and citation-condition tests further indicate that these heads carry attribution-relevant signals, which remain informative even without explicit citation prompts. Overall, our findings suggest that structured attribution signals are embedded in LLM attention, offering a promising path toward more faithful and transparent multi-document reasoning.

7 Acknowledgments

This work is supported by the National Natural Science Foundation of China (62576236), the Key-Area Research and Development Program of Guangdong Province (2024B0101050005), the Natural Science Foundation of Jiangsu Province (BK20251823), Beijing Natural Science Foundation (4262049), Suzhou Key Laboratory of Artificial Intelligence and Social Governance Technologies (SZS2023007), Smart Social Governance Technology and Innovative Application Platform (YZCXPT2023101), the Innovation System of the Integration between Industry and Education for Smart Governance (CJRH2024101), the Leadership Talent Program (Science and Education) of SIP, the Fundamental Research Funds for the Central Universities, JLU, the Priority Academic Program Development of Jiangsu Higher Education Institutions, the open research fund of Suzhou Key Lab of Multi-modal Data Fusion and Intelligent Healthcare, the Science and Technology Development Plan of Suzhou City (Science and Technology Research Program (Medical and Health Innovation)) (SYW2025065), and the U40 Youth Academic Interdisciplinary Team Incubation Project.

References

Mistral AI. Mistral-7b-instruct-v0.3, 2024. URL https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3.

Shahin Atakishiyev, Housam KB Babiker, Jiayi Dai, Nawshad Farruque, Teruaki Hayashi, Nafisa Sadaf Hriti, Md Abed Rahman, Iain Smith, Mi-Young Kim, Osmar R ZaÃ■ane, et al. Explainability of large language models: opportunities and challenges toward generating trustworthy explanations. *arXiv preprint arXiv:2510.17256*, 2025.

BINBINLIU, Wenhan Han, Feng Chen, Yifan Zhang, Ping Guo, Haobin Lin, Bingni Zhang, Taifeng Wang, and Yin Zheng. Token alignment heads: Unveiling attention's role in LLM multilingual translation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Yung-Sung Chuang, Benjamin Cohen-Wang, Zejiang Shen, Zhaofeng Wu, Hu Xu, Xi Victoria Lin, James R. Glass, Shang-Wen Li, and Wen tau Yih. Selfcite: Self-supervised alignment for context attribution in large language models. In *Forty-second International Conference on Machine Learning*, 2025.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Mądry. Contextcite: Attributing model generation to context. *Advances in Neural Information Processing Systems*, 37: 95764–95807, 2024.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

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(1):12, 2021.

Amit Elhelo and Mor Geva. Inferring functionality of attention heads from their parameters. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17701–17733, 2025.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. Eli5: Long form question answering. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3558–3567, 2019.

Javier Ferrando, Gerard I Gállego, Ioannis Tsiamas, and Marta R Costa-Jussà. Explaining how transformers use context to build predictions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5513, 2023.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, 2023a.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, 2023b.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 5484–5495, 2021.

Senyu Han, Hongchuan Zeng, Kai Yu, and Lu Chen. Heads up! large language models can perform tasks without your instruction via selective attention head masking. In *Forty-second International Conference on Machine Learning*, 2025.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *Advances in Neural Information Processing Systems*, 36:76033–76060, 2023.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, 2020.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920. Association for Computational Linguistics, July 2022.

Chengyu Huang, Zeqiu Wu, Yushi Hu, and Wenya Wang. Training language models to generate text with citations via fine-grained rewards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2926–2949, 2024a.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuxuan Gu, Weihong Zhong, Xiachong Feng, Weijiang Yu, Weihua Peng, Duyu Tang, Dandan Tu, et al. Learning fine-grained grounded citations for attributed large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14095–14113, 2024b.

Lei Huang, Xiaocheng Feng, Weitao Ma, Liang Zhao, Yuchun Fan, Weihong Zhong, Dongliang Xu, Qing Yang, Hongtao Liu, and Bing Qin. Advancing large language model attribution through self-improving. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3822–3836, 2024c.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pages 7827–7845, 2024.

Palak Jain, Livio Baldini Soares, and Tom Kwiatkowski. 1-pager: One pass answer generation and evidence retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14529–14543, 2023.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1193–1215, 2024.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, 2017.

Geonhee Kim, Marco Valentino, and André Freitas. Reasoning circuits in language models: A mechanistic interpretation of syllogistic inference. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10074–10095, 2025.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Yanyang Li, Shuo Liang, Michael Lyu, and Liwei Wang. Making long-context language models better multi-hop reasoners. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2462–2475, 2024.

Nelson F Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, 2023.

Peng Liu, Peng-Fei Zhang, Jianfeng Qu, Ximing Li, Zhixu Li, and Pengpeng Zhao. Towards practical llm unlearning: Efficient, modular, and retain-free. In *Proceedings of the ACM Web Conference 2026*, pages 2846–2857, 2026.

Andreas Madsen, Siva Reddy, and Sarath Chandar. Post-hoc interpretability for neural nlp: A survey. *ACM Computing Surveys*, 55(8):1–42, 2022.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 9802–9822, 2023.

Callum Stuart McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. Copy suppression: Comprehensively understanding a motif in language model attention heads. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 337–363. Association for Computational Linguistics, November 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Circuit component reuse across tasks in transformer language models. In *International Conference on Learning Representations*, volume 2024, pages 18349–18377, 2024a.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Circuit component reuse across tasks in transformer language models. In *The Twelfth International Conference on Learning Representations*, 2024b.

Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. Quantifying context mixing in transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3378–3400, 2023.

Benjamin Muller, John Wieting, Jonathan H Clark, Tom Kwiatkowski, Sebastian Ruder, Livio Baldini Soares, Roee Aharoni, Jonathan Herzig, and Xinyi Wang. Evaluating and modeling attribution for cross-lingual question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 144–157, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Yein Park, Chanwoong Yoon, Jungwoo Park, Minbyul Jeong, and Jaewoo Kang. Does time have its place? temporal heads: Where language models recall time-specific information. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16616–16643, 2025.

David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David R So, Maud Texier, and Jeff Dean. The carbon footprint of machine learning training will plateau, then shrink. *Computer*, 55(7):18–28, 2022.

Denis Peskoff and Brandon M Stewart. Credible without credit: Domain experts assess generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 427–438, 2023.

Anirudh Phukan, Shwetha Somasundaram, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. Peering into the mind of language models: An approach for attribution in contextual question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11481–11495, 2024.

Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. Model internals-based answer attribution for trustworthy retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6037–6053. Association for Computational Linguistics, November 2024.

Jianfeng Qu, Wen Hua, Dantong Ouyang, and Xiaofang Zhou. A noise-aware method with type constraint pattern for neural relation extraction. *IEEE transactions on knowledge and data engineering*, 35(2):1134–1148, 2021.

Xiaoye Qu, Zengqi Yu, Dongrui Liu, Wei Wei, Daizong Liu, Jianfeng Dong, and Yu Cheng. Cooperative or competitive? understanding the interaction between attention heads from a game theory perspective. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14079–14099, 2025.

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

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, 2023.

Pedro Sandoval-Segura, Xijun Wang, Ashwinee Panda, Micah Goldblum, Ronen Basri, Tom Goldstein, and David Jacobs. Identifying and evaluating inactive heads in pretrained LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026.

Tobias Schimanski, Jingwei Ni, Mathias Kraus, Elliott Ash, and Markus Leippold. Towards faithful and robust llm specialists for evidence-based question-answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1931, 2024.

Damien Sileo. tasksource: A large collection of NLP tasks with a structured dataset preprocessing framework. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15655–15684. ELRA and ICCL, May 2024.

Shriyank Somvanshi, Md Monzurul Islam, Amir Rafe, Anannya Ghosh Tusti, Arka Chakraborty, Anika Baitullah, Tausif Islam Chowdhury, Nawaf Alnawmasi, Anandi Dutta, and Subasish Das. Bridging the black box: a survey on mechanistic interpretability in ai. *ACM Computing Surveys*, 58(8):1–35, 2026.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. Asqa: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, 2022.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. Recitation-augmented language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Tao Tan, Yining Qian, Ang Lv, Hongzhan Lin, Songhao Wu, Yongbo Wang, Feng Wang, Jingtong Wu, Xin Lu, and Rui Yan. Pear: Position-embedding-agnostic attention re-weighting enhances retrieval-augmented generation with zero inference overhead. In *Proceedings of the ACM on Web Conference 2025*, pages 1693–1702, 2025.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *arXiv preprint arXiv:2605.29358*, 2026.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, 2018.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

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

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020.

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

Zhepei Wei, Wei-Lin Chen, and Yu Meng. InstructRAG: Instructing retrieval-augmented generation via self-synthesized rationales. In *The Thirteenth International Conference on Learning Representations*, 2025.

Orion Weller, Marc Marone, Nathaniel Weir, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. "according to...": Prompting language models improves quoting from pre-training data. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2288–2301, 2024.

Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hannaneh Hajishirzi, and Ashish Sabharwal. Answer, assemble, ace: Understanding how LMs answer multiple choice questions. In *The Thirteenth International Conference on Learning Representations*, 2025.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *International Conference on Learning Representations*, volume 2025, pages 62143–62156, 2025.

Tianyu Xu, Jianfeng Qu, Wen Hua, Zhixu Li, Jiajie Xu, An Liu, Lei Zhao, and Xiaofang Zhou. Evidence reasoning and curriculum learning for document-level relation extraction. *IEEE Transactions on Knowledge and Data Engineering*, 36(2):594–607, 2023.

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, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Haolin Yang, Hakaze Cho, and Naoya Inoue. Localizing task recognition and task learning in in-context learning via attention head analysis. In *The Fourteenth International Conference on Learning Representations*, 2026.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380, 2018.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. Knowledge circuits in pretrained transformers. *Advances in neural information processing systems*, 37:118571–118602, 2024.

Yue Yu, Ting Bai, Hengzhi Lan, Li Qian, Li Peng, Jie Wu, Wei Liu, Jian Luan, and Chuan Shi. C2-cite: Contextual-aware citation generation for attributed large language models. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, pages 871–880, 2026.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024.

Peng-Fei Zhang, Guangdong Bai, Xin-Shun Xu, and Zi Huang. Data-centric challenges, techniques, and impacts: A survey on image data perturbation. *IEEE Transactions on Knowledge and Data Engineering*, 38(9):5736–5764, 2026.

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. Attention heads of large language models. *Patterns*, 6(2), 2025.

Guido Zuccon, Bevan Koopman, and Razia Shaik. Chatgpt hallucinates when attributing answers. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 46–51, 2023.

A Limitations and Impact Statement

Limitations. Our method focuses on sentence–document level attribution and leverages Evidence Heads as an empirical signal, but it does not include a circuit-level analysis of how these heads interact with other components in the model. In particular, we do not explicitly identify the underlying computation pathways or mechanistic circuits through which evidence signals are formed and propagated. As a result, while our findings demonstrate a strong correlation and a causal effect, they do not fully explain why these heads emerge or how they function as part of a broader internal reasoning circuit. Future work could incorporate mechanistic interpretability techniques, such as circuit discovery and path patching, to systematically analyze how Evidence Heads interact with other attention heads and multilayer perceptron (MLP) layers, providing a deeper causal understanding of the attribution process.

Impact Statement. This work aims to improve the interpretability and faithfulness of LLMs in multi-document question answering by providing a training-free attribution mechanism grounded in internal model behavior. By identifying and leveraging specialized attention heads for evidence selection, our approach can help users better understand how generated answers are supported by source documents, potentially increasing trust in LLM-based systems. However, like other interpretability methods, it may also be misused to over-interpret internal signals as definitive explanations of model reasoning. Care should therefore be taken when applying these methods in high-stakes domains, where attribution signals should be treated as informative but not fully definitive evidence of model decision-making processes.

B Entity Perturbation and Evidence Head Selection

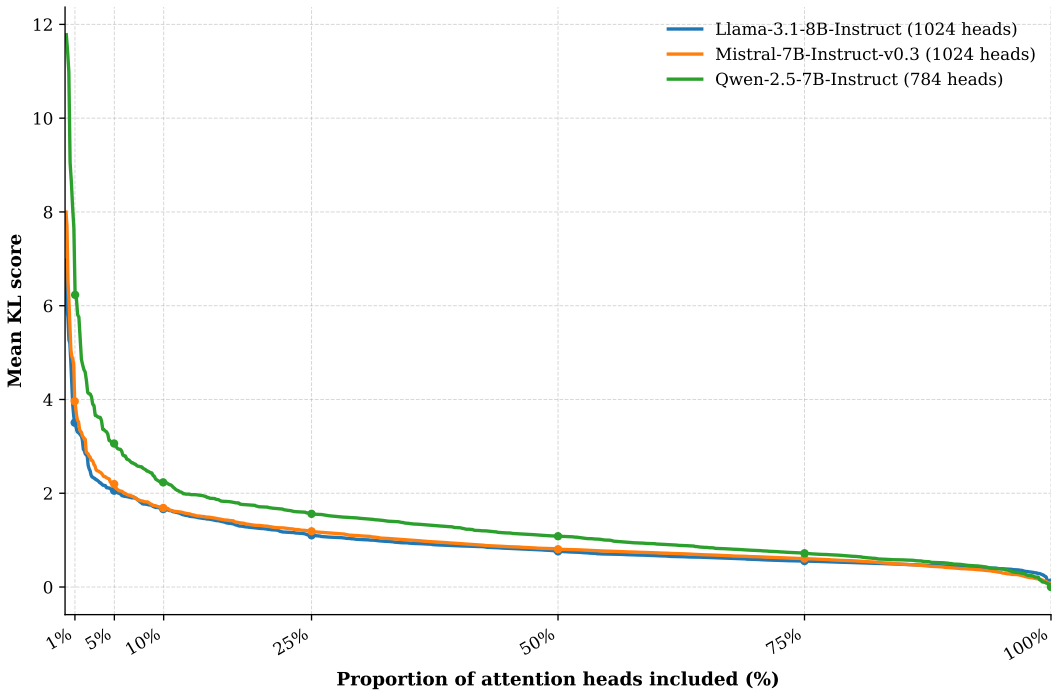

Figure 7: Ranked mean KL-divergence scores of attention heads for Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, and Qwen-2.5-7B-Instruct. Heads are sorted in descending order by mean KL divergence, and the x-axis shows the cumulative proportion of selected heads. The steep initial decline and subsequent long tail motivate selecting the top 1% as a sparse Evidence Head set.

We use the HotpotQA dataset for entity-level perturbation because its `supporting facts` annotations identify the sentences that support each answer. Most answers in HotpotQA are noun phrases, with an average length of 2.56 ± 1.84 words. We build a global replacement pool from filtered

noun entities extracted from document titles and answers, with candidates grouped by token length according to the model's tokenizer. For each sample, we use the `supporting facts` annotations to locate evidence sentences and select answer-related noun entities appearing in those sentences as perturbation targets. Each target is replaced with an entity sampled from the corresponding length group, subject to the following constraints:

- The replacement has the same number of tokenizer tokens as the target.
- The replacement comes from a different sample.
- The replacement does not appear elsewhere in the current sample.
- The replacement has no lexical overlap with the target.
- The embedding-based semantic similarity between the replacement and target is below 0.2.

The token-length constraint preserves input length and token positions after the replaced span. The other constraints reduce the chance that the replacement retains the original evidence or introduces an entity already present in the sample. If no candidate satisfies all constraints, we discard the sample rather than apply an uncontrolled perturbation. As a check, we measure semantic similarity with `all-MiniLM-L6-v2` [Wang et al., 2020]. The replacement entities have a mean similarity of 0.12 ± 0.06 to their original targets, indicating that they are generally unrelated.

For each attention head, we average its KL-divergence scores across perturbation samples and rank heads in descending order by this mean score. Figure 7 plots the ranked KL scores against the cumulative proportion of selected heads for the three models. The curves show a steep initial decline followed by a long, low-score tail, with the 1% point near the transition between these regions. This pattern indicates that evidence sensitivity is concentrated in a sparse subset of heads: selecting substantially fewer heads risks omitting heads around the transition, whereas including more heads adds many lower-sensitivity heads. We therefore use the top 1% as a practical sparse cutoff, applying the same ratio across models rather than tuning it separately for each dataset, and rounding the resulting head count up to the nearest integer. The layer IDs and head IDs corresponding to the identified Evidence Heads are reported in Table 3.

Table 3: Evidence Heads per model. Each tuple denotes (Layer ID, Head ID). Total Layers gives the number of Transformer layers.

Model	Evidence Head	Total Layers
Llama-3.1-8B-Instruct	(13,4) (13,13) (13,18) (14,13) (14,20) (14,22) (14,31) (16,1) (16,8) (17,24) (17,29)	32
Mistral-7B-Instruct-v0.3	(14,21) (14,24) (15,1) (15,7) (16,7) (16,12) (17,0) (17,24) (18,0) (18,3) (19,8)	32
Qwen-2.5-7B-Instruct	(16,0) (19,15) (19,20) (19,22) (21,5) (22,1) (22,7) (22,24)	28

C Attribution Algorithm and Implementation Details

All experiments were conducted on 8 NVIDIA V100 GPUs, each equipped with 32GB of memory. For model inference, we set the maximum input length to 64,000 tokens and the maximum generation length to 2,500 tokens. Greedy decoding was employed throughout all experiments. To ensure full reproducibility, all random seeds were fixed to 42 across all runs. Detailed prompt templates and dataset specifications for all models are provided in Appendix D.

For output post-processing, sentence segmentation was performed using the Natural Language Toolkit (NLTK) Punkt tokenizer, which provides robust and language-agnostic sentence boundary detection. In addition, attention pattern extraction was implemented via forward hooks on transformer attention modules. This design enables non-intrusive access to intermediate attention weights during inference without modifying the underlying model architecture.

To ensure clarity and reproducibility of the proposed framework, we further provide a complete algorithmic specification. In particular, Algorithm 1 describes the attention-based answer attribution

Algorithm 1 Attention-based Attribution

Require: Instruction I, query q, document set $D = \{d_1, \ldots, d_m\}$, generated answer sentences $S = \{s_1, \ldots, s_n\}$, Evidence Head set H, threshold scaling parameter α, attribution strategy $z \in \{\text{global}, \text{local}\}$

Ensure: Attribution results $C = \{c_1, \ldots, c_n\}$

Perform a forward pass of the model on the concatenated input (I, q, D, S)
Extract token-level attention matrices $\{A^h \in \mathbb{R}^{T \times T} \mid h \in H\}$
for each head $h \in H$ **do**
    for each sentence s_i and document d_j **do**
        Compute sentence-document attention via max pooling using Equation 4
    end for
end for
Aggregate across attention heads using Equation 5 to compute $A_{S,D}$
if strategy $z = \text{global}$ **then**
    Compute $\mu = \mu(A_{S,D})$ and $\sigma = \sigma(A_{S,D})$
    for each sentence s_i **do**
        Compute c_i^{global} using Equation 6
    end for
else if strategy $z = \text{local}$ **then**
    for each sentence s_i **do**
        Compute $\mu_i = \mu(\{A_{S,D}(i,j)\}_{j=1}^{m})$ and $\sigma_i = \sigma(\{A_{S,D}(i,j)\}_{j=1}^{m})$
        Compute c_i^{local} using Equation 7
    end for
end if
return $C = \{c_1, \ldots, c_n\}$

procedure, which identifies supporting evidence from a document set for each generated sentence. This is achieved by applying either a global or local adaptive thresholding strategy over the attention signals derived from Evidence Heads.

The NLI model used to compute AutoAIS in Section 4 and 5.1 is t5-xxl-true-nli-mixture [Honovich et al., 2022]. For efficiency, we use a lightweight deberta-base-long-nli model [Sileo, 2024] in Section 5.2 to enable faster inference while maintaining acceptable performance.

D Prompt Templates and Datasets

All datasets used in Table 4, including their questions, answers, and retrieved document collections, are sourced from Gao et al. [2023b], Wei et al. [2025]. For 2WikiMultiHopQA, each sample is associated with 10 candidate documents, whereas all other datasets contain 5 documents per sample.

Table 5 presents the unified prompt template for question answering with retrieval-augmented inputs and self-citation, which requires models to generate answers grounded in provided documents and to cite supporting evidence in a controlled and verifiable manner. The template is used across all models in our experiments and is designed based on the data construction and evaluation settings described above.

To make the no-citation-prompt condition reproducible, Table 6 gives its exact instruction. Compared with the unified prompt in Table 5, it omits explicit citation requirements while retaining the same question and retrieved-document format.

E Attention Distributions of Evidence Heads

To better understand how Evidence Heads align with semantic grounding signals, we visualize attention distributions at multiple levels of granularity and compare them with NLI-based entailment probabilities. Specifically, we present three complementary views based on Llama-3.1-8B-Instruct: token-level attention, sentence-document attention with max pooling, and sentence-document atten-

Table 4: Overview of question answering datasets used in our paper.

Dataset	Type	Description
ASQA	Single-hop	A long-form question answering dataset where questions are often ambiguous and require aggregating information from multiple sources to produce detailed, verifiable answers.
ELI5	Single-hop	A dataset derived from Reddit "Explain Like I'm Five", consisting of open-ended questions that require simple, long-form explanations of complex topics.
Natural Questions	Single-hop	A dataset built from real Google search queries that requires models to locate and extract or generate answers from Wikipedia documents.
PopQA	Single-hop	An open-domain QA dataset focused on popular entities, covering high-frequency factual questions that may still involve long-tail knowledge.
2WikiMultiHopQA	Multi-hop	A multi-hop question answering dataset where answering questions requires reasoning over multiple Wikipedia facts across different pages.
HotpotQA	Multi-hop	A Wikipedia-based multi-hop question answering dataset that requires reasoning across multiple documents and provides supporting facts for explainable answers.
TriviaQA	Multi-hop	A reading comprehension dataset containing trivia questions paired with evidence from multiple documents, emphasizing multi-source reasoning.

tion with mean pooling. In each visualization, the NLI probability matrix is shown as an inset in the top-left corner of the corresponding attention heatmap for direct comparison.

The raw token-level attention weights $A^{h}_{S,D} \in \mathbb{R}^{|\mathcal{T}(S)|\times|\mathcal{T}(D)|}$ (Figure 8) are highly diffuse across token pairs, making it difficult to discern structured evidence relationships. In contrast, max-pooled attention $A^{h,\max}_{S,D} \in \mathbb{R}^{n\times m}$ (Figure 9) produces a sparse and sharply localized alignment pattern, where each sentence-document pair is dominated by a small number of strong token interactions. Notably, the corresponding NLI matrix $E_{S,D} \in \mathbb{R}^{n\times m}$ closely aligns with these high-response regions, suggesting that max pooling effectively preserves the most semantically relevant evidence signals. By comparison, mean-pooled attention $A^{h,\text{mean}}_{S,D} \in \mathbb{R}^{n\times m}$ (Figure 10) yields a smoother and more distributed pattern across sentence-document pairs, potentially diluting salient alignment signals.

Overall, these visualizations suggest that Evidence Heads encode sparse and highly localized alignment structures that are better captured by extremal aggregation than by averaging. The consistent correspondence between max-pooled attention and the inset NLI probability matrices further supports the hypothesis that Evidence Heads align with interpretable semantic grounding mechanisms rather than reflect a uniform distribution of attention.

Table 5: Unified prompt template for retrieval-augmented question answering with document grounding and citation constraints.

Instruction: Write an accurate, engaging, and concise answer for the given question using only the provided search results (some of which might be irrelevant) and cite them properly. Use an unbiased and journalistic tone. Always cite for any factual claim. When citing several search results, use [1][2][3]. Cite at least one document and at most three documents in each sentence. If multiple documents support the sentence, only cite a minimum sufficient subset of the documents.

Question: {Demo Query}
Document[1](Title: {Demo Title 1}): {Demo Content1}
Document[2](Title: {Demo Title 2}): {Demo Content2}
. . .
Answer: {Demo Answer}

Question: {User Query}
Document[1](Title: {Title 1}): {Content1}
Document[2](Title: {Title 2}): {Content2}
. . .
Answer:

Table 6: Instruction used to generate answers in the no-citation-prompt condition.

Instruction: Write an accurate, engaging, and concise answer for the given question using only the provided search results (some of which might be irrelevant). Use an unbiased and journalistic tone.

Question: {Demo Query}
Document[1](Title: {Demo Title 1}): {Demo Content1}
Document[2](Title: {Demo Title 2}): {Demo Content2}
. . .
Answer: {Demo Answer}

Question: {User Query}
Document[1](Title: {Title 1}): {Content1}
Document[2](Title: {Title 2}): {Content2}
. . .
Answer:



Figure 8: Token-level attention heatmaps of Evidence Heads with the corresponding NLI probability matrix shown as an inset in the top-left corner. Rows correspond to tokens in the generated answer and columns correspond to tokens in the document set, representing fine-grained token-level interactions across different attention heads.

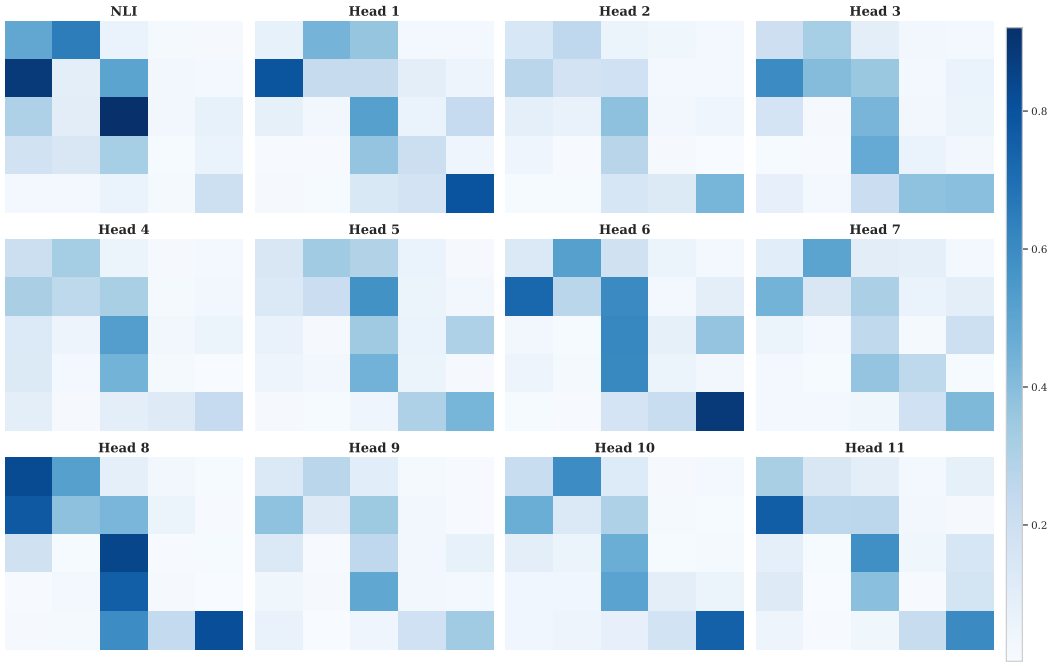

Figure 9: Sentence-document attention heatmaps obtained via max pooling over token-level interactions, with the corresponding NLI probability matrix shown as an inset in the top-left corner. Rows correspond to sentences in the generated answer and columns correspond to the document set, representing sentence-level alignment patterns across different attention heads.



Figure 10: Sentence-document attention heatmaps obtained via mean pooling over token-level interactions, with the corresponding NLI probability matrix shown as an inset in the top-left corner. Rows correspond to sentences in the generated answer and columns correspond to the document set, representing aggregated sentence-level evidence distributions across different attention heads.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's main contributions, assumptions, and scope without overstating the results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses its limitations in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not present any formal theoretical results, theorems, or proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides sufficient details on the experimental setup, including algorithm descriptions, evaluation procedures, and key hyperparameters.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper provides open access to the code and data, enabling faithful reproduction of the main experimental results.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The proposed method is training-free, and the paper clearly specifies all relevant experimental settings to understand and interpret the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper does not report error bars, confidence intervals, or statistical significance tests for the experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides sufficient information on the computational setup, including the GPU type and number used, enabling reproduction of the experiments within a comparable compute environment.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research has been conducted in accordance with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses positive and negative societal impacts in Appendix A.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper reports no high-risk data or models requiring special safeguards, as it uses open-source models and datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators or original owners of all assets used in the paper are properly credited, and their licenses and terms of use are explicitly mentioned and respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release any new datasets, codebases, or models that require separate documentation or accompanying assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve any human subjects research or crowdsourcing experiments, and therefore does not require IRB approval or equivalent ethical review.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The paper does not use LLMs as a component of the core methodology, and any potential use is limited to writing or editing assistance.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.