Reading Citations from Attention: Faithful Source Attribution in Retrieval Augmented Generation

Anonymous ACL submission

Abstract

Fine-grained citations are essential for trustworthy retrieval-augmented generation (RAG), but most systems ask models to generate citation markers as output tokens, adding formatting burden and potentially obscuring actual evidence use. We propose a training-free framework that derives sentence-level citations directly from retrieval-head attention. An offline semantic probe selects the head that best aligns generated clauses with supporting source sentences. At inference time, the model answers once without citation instructions; we aggregate the selected head's attention into a clause-to-sentence matrix and apply a peak-minus-entropy rule to cite or abstain. The method requires no retriever, verifier, or citation-specific tuning. On LongBench-Cite, it achieves 82.7–86.4 citation F1 across models, outperforming prompting, post-hoc matching, and fine-tuned citation generation while preserving answer correctness.

1 Introduction

Retrieval-augmented generation (RAG) enables language models to answer questions from external evidence rather than relying only on parametric knowledge (Lewis et al., 2020; Jiang et al., 2023). For such systems, citations are not merely a formatting convention: they are the main interface through which users verify whether each generated claim is supported by the retrieved context, trace answer provenance, and diagnose hallucinations. This makes fine-grained attribution a core requirement for trustworthy long-context generation (Rashkin et al., 2023; Bohnet et al., 2022; Gao et al., 2023). However, recent analyses show that self-generated citations can be malformed, unsupported, or disconnected from the evidence actually used by the model (Qi et al., 2024).

Most existing approaches address attribution by asking the model to generate citation markers together with the answer. These methods vary in supervision and optimization—from prompting to supervised fine-tuning and alignment—but they share the same premise: attribution is represented as surface-form output tokens (Zhang et al., 2025; Chuang et al., 2025; Yu et al., 2026). This creates two limitations. First, citation quality depends on the model's ability to learn a specialized output format, which can introduce generation burden and correctness degradation. Second, the generated marker may not reflect the model's actual evidence-use process. We instead ask whether citations can be recovered from the internal signal that the model already computes while producing the answer.

Our key observation is that long-context models often reveal evidence use through a small set of retrieval heads. Prior work shows that these heads carry sparse context-retrieval behavior and are causally linked to factual recall (Wu et al., 2025). We therefore formulate citation generation as an attention readout problem: rather than generating citation tokens, we read where the model looked when it generated each clause.

The proposed pipeline has two stages. In an offline calibration stage, we identify a citation head with semantic probing. Unlike verbatim needle-in-a-haystack probes, our probe aligns generated clauses with supporting source sentences in embedding space, so it can capture paraphrase, compression, and synthesis rather than only copied tokens. In the inference stage, the model generates a normal answer once, with no citation instruction. We record the selected head's attention, segment the answer into clauses and the context into source sentences, normalize token-level attention, and sum attention mass over each source-sentence span to obtain a clause-to-sentence citation matrix. A confidence rule then emits citations when the attention row is sharply peaked and abstains when it is diffuse.

This design addresses three practical challenges. Semantic probing selects heads that track support-

ing evidence rather than heads that only copy surface tokens. Normalize-then-sum aggregation converts token-level attention into sentence-level citation scores without penalizing longer evidence sentences. Finally, the peak-minus-entropy decision rule separates attributable factual clauses from transitional or unsupported clauses whose attention is spread across the context.

Our contributions are:

- We reframe fine-grained citation as a training-free readout from retrieval-head attention, avoiding citation-token generation and citation-specific fine-tuning.
- We introduce a complete attention-based citation pipeline that combines semantic citation-head probing, normalize-then-sum aggregation, and peak-minus-entropy cite/abstain decisions.
- We show on LongBench-Cite that attention-derived citations consistently outperform prompting, post-hoc matching, and generated citations from a fine-tuned citation model, while preserving answer correctness.

2 Related Works

Attribution in retrieval-augmented generation. Fine-grained attribution connects generated statements to the external evidence that supports them, and is therefore central to reliable RAG (Rashkin et al., 2023; Bohnet et al., 2022; Gao et al., 2023). Existing methods mainly fall into two groups. Post-processing methods attach citations after generation through lexical or semantic matching, entailment verification, or lightweight correction models, as in systems such as RAGFlow, TRUE, and CiteFix (Infiniflow, 2024; Honovich et al., 2022; Maheshwari et al., 2025). They are easy to deploy because they do not modify the generator, but they decouple attribution from the model's actual decoding process. End-to-end methods instead prompt, fine-tune, or align LLMs to emit citation markers together with the answer (Nakano et al., 2021; Zhang et al., 2025; Chuang et al., 2025; Yu et al., 2026). These methods can produce more integrated outputs, but they require citation data or optimization and can burden generation with an additional formatting task. Our work differs from both families: we do not add a post-hoc verifier or train the model to generate citation tokens, but recover citations from the attention trace produced during normal answer generation.

Retrieval heads in large language models. Interpretability studies show that attention heads can specialize into distinct functional roles, including induction, knowledge, reasoning, and safety-related heads (Olsson et al., 2022; Wang et al., 2023; Jiang et al., 2024; Wang et al., 2024; Zhang et al., 2024). In long-context settings, retrieval heads are particularly relevant: they form a small set of heads that concentrate attention on key evidence, and pruning them substantially hurts long-context recall and factuality (Wu et al., 2025). Recent systems exploit this property for long-context efficiency or robustness, for example by allocating full KV cache to retrieval heads or compressing non-retrieval heads (Xiao et al., 2025; Tang et al., 2025; Zheng et al., 2025). We build on the same mechanistic insight, but use retrieval heads for attribution. Specifically, we select a citation head with semantic probing and convert its attention into sentence-level citations through normalized attention aggregation and confidence-aware abstention.

3 Method

3.1 Overview and Motivation

Our goal is to produce sentence-level citations for the free-form answers of a long-context language model without additional training, an external retriever, or a separate verification model. The central observation is that grounding information is already present in the model's decoding process. When the model generates a faithful answer, part of its attention returns to the source sentences that support the current statement. A citation can therefore be recovered from the model's internal evidence-use trace rather than generated by an extra prediction module.

This observation leads to two design problems. The first problem is to locate where the grounding signal is concentrated in the network. Prior work on retrieval heads shows that long-context recall is carried by a small and sparse set of attention heads rather than distributed uniformly across all heads (Wu et al., 2025). The second problem is to convert a noisy token-level attention map into a discrete sentence-level citation decision, including the decision to abstain when a generated clause is not attributable to a specific source sentence. We address the first problem with a behavioral probing procedure that selects a citation head, and we

address the second problem with an attention readout that aggregates the selected head's attention into clause-to-sentence citation scores. Figure 1 summarizes the pipeline.

Notation. Given a query q and a source document D, we segment D into source sentences $\mathcal{C} = \{c_1, \ldots, c_M\}$. Source sentence c_j occupies the prompt-token span $I_j = \{a_j, \ldots, b_j - 1\}$. The model generates an answer Y autoregressively, and we segment Y into clauses $\mathcal{S} = \{s_1, \ldots, s_N\}$. Clause s_i corresponds to the decoding-step set $T_i \subseteq \{1, \ldots, T\}$. An attention head is indexed by $h = (l, k)$, where l is the layer and k is the head index. At decoding step t, head h produces an attention distribution $\alpha_t^h \in \Delta^{|D|}$ over document tokens. English source sentences are segmented with the Punkt sentence tokenizer (Kiss and Strunk, 2006), and Chinese text is segmented with punctuation rules. A merge pass absorbs fragments shorter than 15 characters into the preceding unit so that citation targets remain semantically self-contained.

3.2 Citation Head Identification

Using all attention heads washes out the sparse grounding signal. We therefore first probe the model to identify the head whose attention most reliably tracks the evidence used by generated clauses. Unlike the original needle-in-a-haystack protocol (Kamradt, 2023), which inserts a short needle passage into a long distractor context, we construct probes from documents better suited to RAG scenarios. Specifically, we randomly sample 100 Chinese and 100 English instances from the LongCite-45k training set and query the model to generate answers conditioned on the provided long-context documents.

The original retrieval-head probe rewards verbatim recall (Wu et al., 2025). A head scores well when its top-attended token coincides with the source token being reproduced. This criterion is too restrictive for citation, because a cited answer often paraphrases, compresses, or synthesizes evidence rather than copying it. We therefore replace token-level copying with sentence-level semantic alignment. The probe credits a head when its attention points to the evidence sentence that semantically supports the generated clause.

Semantic alignment. Let $\mathcal{C}_N$ denote the set of sentences in the inserted evidence passage. We embed every evidence sentence and every generated clause with the BGE-M3 sentence encoder (Chen et al., 2024). For clause s_i, the aligned evidence sentence $c_i^\star$ and its semantic confidence σ_i are

$$\begin{aligned} c_i^\star &= \arg\max_{c \in \mathcal{C}_N} \mathrm{sim}\big(\phi(s_i), \phi(c)\big), \\ \sigma_i &= \max_{c \in \mathcal{C}_N} \mathrm{sim}\big(\phi(s_i), \phi(c)\big). \end{aligned} \tag{1}$$

Here sim denotes cosine similarity. Clause s_i is treated as valid when $\sigma_i \geq \delta$ and invalid when $\sigma_i \leq \delta - \epsilon$. Clauses in the dead zone are omitted from the head score. We use $\delta = 0.7$ and $\epsilon = 0.05$.

Grounding score. For head h, let

$$m_t^h = \arg\max_{1 \leq m \leq |D|} \alpha_t^h[m] \tag{2}$$

be the document-token position that receives the largest attention at decoding step t. For a valid clause s_i, we measure how often the top-attended token falls inside the aligned evidence sentence:

$$g_h(s_i) = \frac{1}{|T_i|} \sum_{t \in T_i} \mathbf{1}\Big[m_t^h \in \mathrm{span}(c_i^\star)\Big]. \tag{3}$$

This term rewards heads that consistently point to the correct supporting sentence while the model generates the clause.

A useful citation head should also remain uncommitted when the clause is not grounded in the evidence passage. For an invalid clause, we compute the largest fraction of decoding steps whose top-attended token falls in any single source sentence:

$$\mathrm{conc}_h(s_i) = \max_{1 \leq j \leq M} \frac{1}{|T_i|} \sum_{t \in T_i} \mathbf{1}\Big[m_t^h \in \mathrm{span}(c_j)\Big]. \tag{4}$$

This concentration term penalizes heads that confidently point to a source sentence even when the generated clause lacks reliable semantic support.

Let $\mathcal{V}$ and $\mathcal{I}$ be the valid and invalid clause sets for one probing example. The score of head h is

$$\begin{aligned} r_h = &\frac{\sum_{s_i \in \mathcal{V}} \sigma_i \, g_h(s_i)}{\sum_{s_i \in \mathcal{V}} \sigma_i} \\ &- \frac{\sum_{s_i \in \mathcal{I}} (1 - \sigma_i) \, \mathrm{conc}_h(s_i)}{\sum_{s_i \in \mathcal{I}} (1 - \sigma_i)}. \end{aligned} \tag{5}$$

If either set is empty, the corresponding term is omitted. We average r_h over the probing set and select the citation head

$$h^\star = \arg\max_h \mathbb{E}_{\mathrm{probe}}\big[r_h\big]. \tag{6}$$

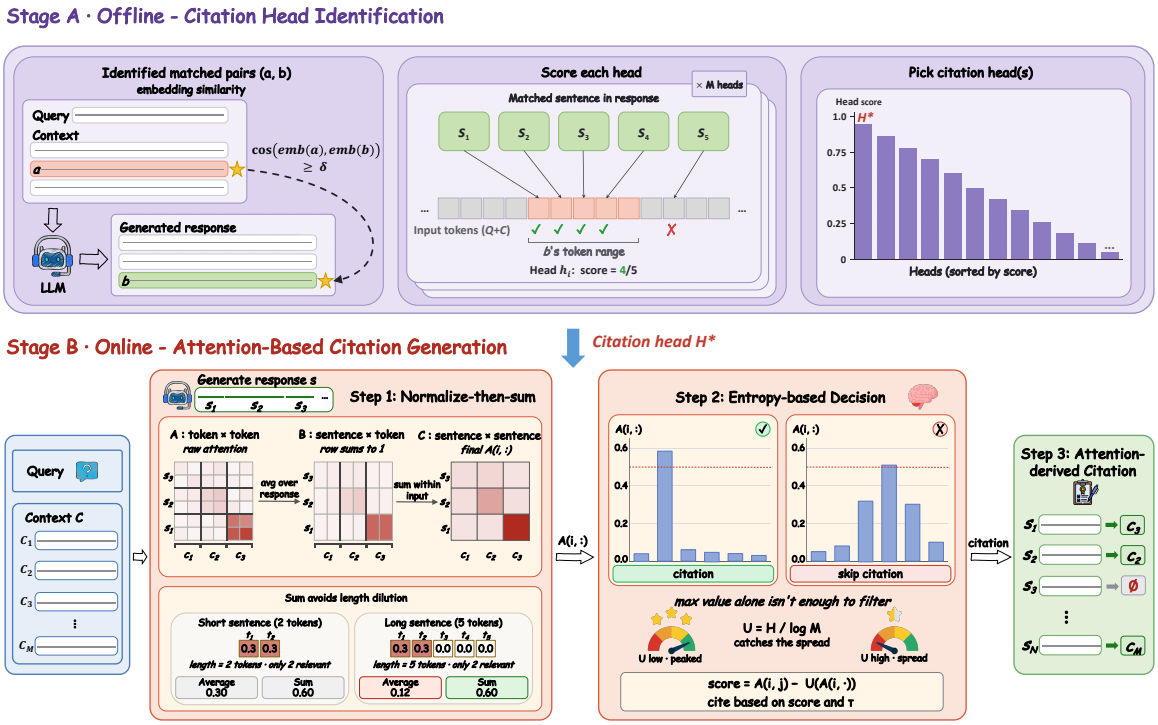

Figure 1: Overview of the training-free attention-based citation framework. Citation-head identification uses semantic probing on needle-in-a-haystack examples to select the attention head that most reliably tracks supporting evidence. During inference, the model generates the answer once while recording the selected head's attention. The method normalizes token-level attention, sums attention mass over source-sentence spans, and constructs a clause-to-sentence citation matrix. A confidence-aware decision module emits a citation when the corresponding attention row is sufficiently concentrated and abstains when the row is diffuse.

The key design is the reward-penalty structure in Eq. (5). A head is selected only when it attends to the right evidence for grounded clauses and avoids sharp spurious attention for ungrounded clauses.

3.3 Attention-Based Citation Generation

After selecting $h^\star$, inference requires only one normal decoding pass. We generate the answer with greedy decoding and record the selected head's attention over document tokens at every step:

$$a_t = \alpha_t^{h^\star} \in \mathbb{R}^{|D|}. \tag{7}$$

Citation is then a deterministic readout from the recorded attention sequence $\{a_t\}_{t=1}^{T}$.

Normalize-then-sum aggregation. The readout must convert token-level attention over decoding steps into a clause-to-sentence score. We first average the selected head's attention over the decoding steps of clause s_i and normalize the result over document tokens:

$$\bar{a}_i = \frac{1}{|T_i|} \sum_{t \in T_i} a_t, \qquad p_i = \frac{\bar{a}_i}{\sum_{m=1}^{|D|} \bar{a}_i[m]}. \tag{8}$$

We then sum the normalized attention mass inside each source-sentence span:

$$A_{ij} = \sum_{m=a_j}^{b_j - 1} p_i[m]. \tag{9}$$

The core aggregation principle is **normalize-then-sum**. Normalization makes rows comparable across clauses, while summation preserves the interpretation of A_{ij} as the total evidence-attention mass assigned from clause s_i to source sentence c_j. This avoids the length bias introduced by mean pooling, which can understate the relevance of longer evidence sentences.

We row-normalize the sentence scores to obtain a distribution over candidate source sentences:

$$\bar{A}_{ij} = \frac{A_{ij}}{\sum_{j'=1}^{M} A_{ij'}}, \qquad \sum_{j=1}^{M} \bar{A}_{ij} = 1. \tag{10}$$

The matrix $\bar{A} \in \mathbb{R}^{N \times M}$ is the clause-to-sentence citation matrix.

Confidence-aware citation decision. A clause should receive citations only when its attention is concentrated on a small set of source sentences. Clauses that are transitional, connective, or not directly attributable tend to spread their attention diffusely over many source sentences. We characterize each citation row $\bar{A}_{i,:}$ with two quantities. The *row peak* $\max_j \bar{A}_{ij}$ is the largest mass assigned to any single source sentence. The *normalized entropy* $U_i \in [0, 1]$ measures how dispersed the row is: it is the Shannon entropy H_i divided by its maximum possible value $\log M$, where M is the number of source sentences and $\log M$ is the entropy of the uniform distribution over them. U_i is close to 0 when the row is peaked on one sentence and close to 1 when the row is uniform. Because U_i already lives on the same $[0, 1]$ scale as the peak, we subtract it directly and no trade-off coefficient is needed. The clause confidence score ψ_i is the row peak minus the normalized entropy:

$$\begin{aligned} H_i &= -\sum_{j=1}^{M} \bar{A}_{ij} \log \bar{A}_{ij}, \\ U_i &= \frac{H_i}{\log M}, \\ \psi_{ij} &= \bar{A}_{ij} \; - \; U_i. \end{aligned} \tag{11}$$

We use the convention $0 \log 0 = 0$.

Rather than citing only the single peak sentence, we cite every source sentence whose attention mass is comparable to the peak. We define the candidate set

$$\mathcal{J}_i = \Big\{ j : \bar{A}_{ij} > \beta \max_{1 \le j' \le M} \bar{A}_{ij'} \Big\}, \qquad \beta = 0.5, \tag{12}$$

i.e., the source sentences whose mass exceeds half of the row peak. A clause emits citations only when its confidence score clears the threshold τ, and abstains otherwise:

$$\mathrm{cite}(s_i) = \begin{cases} \{\, c_j : j \in \mathcal{J}_i \,\}, & \psi_{ij} > \tau, \\ \emptyset, & \psi_{ij} \le \tau, \end{cases} \tag{13}$$

where τ is a fixed hyperparameter set to $\tau = -0.7$. The core decision rule is **peak-minus-entropy**. A peaked row has a high peak and low normalized entropy, so it yields a high ψ_i and a confident citation; a diffuse row has high normalized entropy, a low ψ_i, and is more likely to abstain. When a clause does cite, the half-peak candidate set lets it attribute to several supporting sentences at once rather than a single one, which matters when a clause synthesizes evidence spread across multiple source sentences.

4 Experiments

4.1 Experimental Setup

Benchmark. We evaluate on LongBench-Cite (Zhang et al., 2025), a standard benchmark for **long-context question answering with citations (LQAC)**. It comprises five datasets: MultiFieldQA-en/zh (single-document QA), HotpotQA and DuReader (multi-document QA), GovReport (summarization), and LongBench-Chat (real-world mixed tasks). Following Zhang et al. (2025), we adopt **sentence-level citations** rather than the chunk-level scheme used in prior work (Gao et al., 2023). Specifically, we segment both the source document and model responses into sentences using NLTK (Bird, 2006) for English and punctuation-based rules for Chinese. This avoids the mid-sentence truncation and coarse granularity of fixed-length chunks, yielding more precise and verifiable attributions. Dataset statistics are detailed in Appendix A.

Evaluation of citation quality. Following Zhang et al. (2025), we adopt **citation F1**—computed from citation recall (R) and citation precision (P)—as the primary metric for citation quality, and employ GPT-4o as the automatic judge to better handle the paraphrasing and synthesis prevalent in long-context QA.

Citation recall is scored per statement (0/0.5/1) and averaged over all statements. For a statement s_i with at least one citation ($\mathrm{cite}(c_i) \neq \emptyset$), we concatenate all cited snippets and prompt GPT-4o to judge whether the concatenated text *fully supports* (1 point), *partially supports* (0.5 point), or *does not support* (0 point) s_i. For statements without citations ($\mathrm{cite}(c_i) = \emptyset$), GPT-4o determines whether s_i is a functional sentence (e.g., introductory, transitional, or summary) that legitimately requires no citation—if so, recall is 1; otherwise, 0.

Citation precision is scored per citation (0/1 for irrelevant/relevant) and averaged over all citations. A cited snippet $c_{i,j}$ is deemed relevant if it entails at least some key points of s_i, i.e., partially supports the statement.

Citation F1 combines the two on a per-response basis: $\mathrm{F1}_i = (2 \cdot P_i \cdot R_i)/(P_i + R_i)$. The final re-

ported F1 is the **macro-average** over all responses, i.e., the arithmetic mean of individual F1 scores.

Evaluation of correctness. Following Zhang et al. (2025), we ask GPT-4o to rate the response against ground-truth answers via few-shot prompting (for LongBench-Chat) or zero-shot prompting (for other datasets). The rating yields the LQAC correctness score **C**. We also evaluate $\mathbf{C_{LQA}}$—the correctness in the vanilla long-context QA setting where the model receives only context and query without any citation instruction. The **correctness ratio** $\text{CR} = C/C_{\text{LQA}} \times 100\%$ then quantifies the impact of citation generation: $\text{CR} > 100\%$ indicates improvement, while $\text{CR} < 100\%$ signals degradation due to the added generation burden.

Models and baselines. We compare our training-free method against three classes of citation generation approaches:

- *In-context learning (ICL).* We use the one-shot LQAC prompt provided by Zhang et al. (2025), which demonstrates the sentence-level citation format with `<statement>` and `<cite>` tags. We apply this prompt to both *proprietary model* (GPT-4o (OpenAI, 2023)) and *open-source instruct models* (Qwen2.5-7B-Instruct (Yang et al., 2024), Qwen3-8B (Team, 2025), Llama-3.1-8B/70B-Instruct (Team, 2024), DeepSeek-V4-Flash (DeepSeek-AI, 2026)), asking them to generate answers with inline citations in a single pass.

- *Post-hoc method.* We first generate answers in the vanilla long-context QA setting (no citation instruction), then annotate citations retrospectively. We implement *BGE-M3 matching*—using the BGE-M3 sentence encoder (Chen et al., 2024) to retrieve semantically similar source sentences for each generated clause. This provides more robust similarity matching than string-based methods, as it captures semantic paraphrasing and cross-lingual alignment beyond surface-form overlap.

- *Fine-tuned models.* We report LongCite-8B (Zhang et al., 2025), which is Llama-3.1-8B fine-tuned on 45K supervised LQAC instances and generates citations as part of its output sequence.

For our method, we apply the attention-based readout pipeline to four model configurations: Qwen2.5-7B-Instruct, Qwen3-8B, and Llama-3.1-8B-Instruct (all using vanilla long-context QA prompts, i.e., no citation instruction), as well as LongCite-8B (using its native LQAC prompt). We use greedy decoding and extract citations from the top-1 citation head identified by the semantic probing procedure in §3.2. Hyperparameters and prompt templates for all methods are detailed in Appendix B and Appendix C, respectively.

4.2 Main Results

Citation quality analysis. Table 1 presents citation quality results across all methods and models.

Baseline limitations. In-context learning exhibits high variance across model capabilities: smaller models (Qwen2.5-7B, Llama-3.1-8B) achieve <25 F1, struggling with the added burden of learning citation syntax alongside answer generation. In contrast, stronger models (GPT-4o at 65.6, DeepSeek-V4-Flash at 64.6) perform substantially better, yet still fall short of producing reliable fine-grained attributions. Post-hoc BGE-M3 matching provides more stable results (65.8–68.2 F1) but hits a clear ceiling, as it cannot exploit the generator's internal reasoning traces. Supervised fine-tuning (LongCite-8B at 72.0 F1) unlocks reliable citation generation but requires costly annotated data.

Our attention-based method dominates all baselines with minimal overhead. Most critically, our training-free attention readout achieves **82.7–86.4** F1 across all four model configurations. For the same model, compared to its strongest baseline, our method achieves absolute gains ranging from **14.4** points (LongCite-8B vs. its own generated citations at 72.0) to **17.8** points (Qwen3-8B vs. post-hoc matching at 67.5). This demonstrates that attention-derived citations are not merely riding on stronger base models; they extract a reliable grounding signal that is universally present but previously underexploited. Case studies are detailed in Appendix F.

Two findings merit particular emphasis. First, applying our method to LongCite-8B yields a **14.4-point** improvement on F1 over the model's own generated citations. This indicates that even when a fine-tuned model produces malformed or inaccurate citation markers, its *internal attention* may still faithfully track the true evidence. Second, LongCite-8B under our readout outperforms Llama-3.1-8B-Instruct under the same readout by 2.9 points, despite both using identical attention aggregation. This suggests that LongCite's super-

Model	Avg. F1	Longbench-Chat			MultifieldQA			HotpotQA			Dureader			GovReport		
		R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1
One-Shot Prompting (Proprietary model)																
GPT-4o	65.6	46.7	53.5	46.7	79.0	87.9	80.6	55.7	62.3	53.4	65.6	74.2	67.4	73.4	90.4	79.8
One-Shot Prompting (Open-source models)																
Qwen2.5-7B-Instruct	23.3	19.2	11.8	10.7	37.9	37.0	34.2	11.3	12.3	9.2	42.4	32.8	29.7	15.6	16.9	15.1
Qwen3-8B	56.9	33.1	37.2	32.0	71.5	73.5	69.6	38.9	37.2	34.2	61.9	60.9	58.7	58.0	68.6	62.0
Llama-3.1-8B-Instruct	19.7	14.1	19.5	12.4	29.8	44.3	31.6	20.2	30.9	20.9	22.0	25.1	17.0	16.2	25.3	16.8
Llama-3.1-70B-Instruct	40.4	25.8	32.0	23.2	53.2	65.2	53.9	29.6	37.3	28.6	38.2	46.0	35.4	53.4	77.5	60.7
DeepSeek-V4-Flash	64.6	65.3	53.3	53.2	77.2	73.9	72.4	59.9	57.5	53.5	74.8	72.7	69.9	69.6	55.0	59.7
Post-hoc method (BGE-M3 matching)																
Qwen2.5-7B-Instruct	65.8	66.2	61.2	56.1	72.7	69.7	67.5	50.7	63.9	50.4	74.4	63.4	66.6	78.7	83.1	80.0
Qwen3-8B	67.5	62.3	65.1	56.4	77.3	68.7	68.9	50.1	75.5	55.6	73.9	61.1	65.0	82.3	83.2	82.3
Llama-3.1-8B-Instruct	68.0	68.8	62.2	60.0	78.7	67.6	68.5	51.8	64.4	53.0	84.1	61.6	68.1	**90.4**	79.2	83.9
DeepSeek-V4-Flash	68.2	64.6	64.2	57.3	79.2	70.6	71.1	49.0	57.0	47.9	**84.8**	63.6	70.6	82.3	85.7	83.6
Fine-tuned model																
LongCite-8B	72.0	62.0	79.7	67.4	74.7	93.0	80.8	59.2	72.1	60.3	68.3	85.6	73.1	74.0	86.6	78.5
Our attention-based method																
Qwen2.5-7B-Instruct	82.7	69.5	73.3	69.6	86.2	90.4	87.3	69.5	71.3	67.9	84.6	**90.0**	**86.7**	84.6	94.3	88.9
Qwen3-8B	85.3	76.9	79.4	76.7	**90.9**	93.6	**91.3**	78.7	77.7	76.6	79.9	87.4	83.0	82.3	**95.1**	87.9
Llama-3.1-8B-Instruct	83.5	74.4	80.6	76.2	88.2	**93.9**	90.0	71.3	79.1	73.3	78.4	85.1	80.9	80.6	95.0	86.7
LongCite-8B	**86.4**	**83.3**	**88.5**	**84.0**	89.7	89.9	89.2	**83.6**	**79.4**	**78.7**	83.5	88.1	85.1	88.5	94.8	**91.1**

Table 1: Citation recall (R), citation precision (P), citation F1 (F1), and citation length evaluated on LongBench-Cite benchmark. Best and second-best results are **bolded** and underlined, respectively.

vised fine-tuning not only teaches citation formatting but also sharpens the model's attention toward relevant context, making the internal grounding signal more accurate for downstream readout.

Dataset-level analysis reveals the advantage of attention-based attribution on complex reasoning tasks. HotpotQA proves most challenging for all baselines. The difficulty stems from HotpotQA's demand for *multi-hop reasoning*—models must combine, paraphrase, and infer from multiple source sentences rather than verbatim copying localized passages. Both ICL and post-hoc methods falter here because they lack access to the model's actual reasoning process. In contrast, our method achieves **67.9–78.7** F1 on HotpotQA, with gains of **7.4–58.7** points over ICL and **12.0–23.2** points over post-hoc matching. By reading citations from where the model *actually attended* during multi-hop reasoning, we bypass the opacity of paraphrased generation and recover the true evidence-use trace. This advantage is less pronounced on single-document QA datasets like MultiFieldQA, where evidence localization is simpler and all methods perform more competitively.

Correctness analysis. Table 2 reports the correctness impact of adding citations. Column **C** denotes the answer quality score under the LQAC (long-context question answering with citations) strategy, while column $\mathbf{C_{LQA}}$ denotes the score under the vanilla long-context QA strategy. For *proprietary, open-source*, and *fine-tuned models*, these two strategies are realized by varying the prompt, instructing model to generate answers with or without inline citation annotations. Most proprietary and open-source models suffer correctness degradation when forced to generate citations inline (CR $< 100\%$, red), consistent with the distribution-shift hypothesis in Zhang et al. (2025). In contrast, the fine-tuned LongCite-8B achieves CR = 107%, indicating that supervised adaptation to the LQAC format not only preserves but enhances generation quality.

Our method occupies a unique position: it incurs **zero modification to the model's output distribution**. Since citations are read from attention rather than generated as tokens, answer remains unchanged regardless of whether citations are required or not. We test our method under two prompt conditions: for Qwen2.5-7B-Instruct, Qwen3-8B, and Llama-3.1-8B-Instruct,

Model	Avg. C	Avg. C_{LQA}	CR
Proprietary model			
GPT-4o	69.4	78.2	88%
Open-source models			
Qwen2.5-7B-Instruct	58.8	65.4	90%
Qwen3-8B	68.8	67.9	101%
Llama-3.1-8B-Instruct	52.1	60.2	86%
DeepSeek-V4-Flash	79.3	80.8	98%
Fine-tuned model			
LongCite-8B	71.7	67.6	107%
Our method			
Qwen2.5-7B-Instruct	65.4	65.4	100%
Qwen3-8B	67.9	67.9	100%
Llama-3.1-8B-Instruct	60.2	60.2	100%
LongCite-8B	71.7	71.7	100%

Table 2: Correctness in LQAC setting (C), correctness in vanilla long-context QA setting (C_{LQA}), and correctness ratio (CR) of different models on LongBench-Cite. We mark the cases where adding citations improves/hurts correctness (i.e., CR $>$ 1 / CR $<$ 1) in green/red.

Settings	R	P	F1	ΔF1 (%)
Our Method	81.4	86.4	82.7	—
w/o entropy filtering	74.8	87.6	78.9	−4.6
w/o sum aggregation	76.1	82.0	77.4	−6.4

Table 3: Ablation study on Qwen2.5-7B-Instruct. We report average citation recall (R), precision (P), and F1 on LongBench-Cite. ΔF1 denotes the relative F1 drop compared to the complete method.

we use the vanilla long-context QA prompt (score C_{LQA}); for LongCite-8B, we use its native LQAC prompt(score C).

4.3 Ablation Studies

Design choices. Table 3 isolates the contribution of two core components on Qwen2.5-7B-Instruct. Removing the peak–entropy filtering (§3.3) and using raw attention peaks drops F1 by 4.6%, primarily due to a 6.6-point recall loss: without entropy regularization, more functional clauses receive false citations. Replacing normalize-then-sum with mean–mean aggregation drops F1 by 6.4%, with recall falling from 81.4 to 76.1 and precision from 86.4 to 82.0. This confirms that mean pooling dilutes attention mass over long evidence sentences, causing the model to miss key citations while also slightly degrading precision.

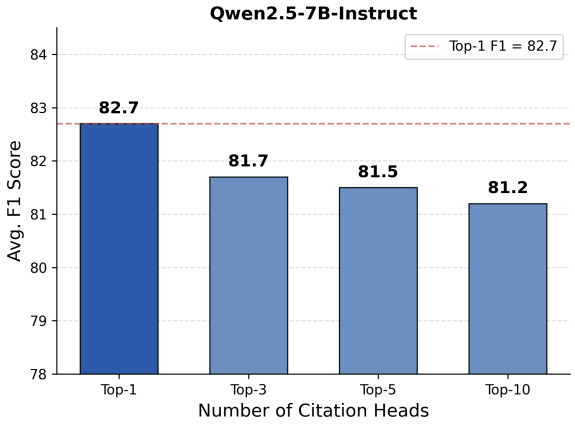

Figure 2: Effect of head ensemble size on citation F1 for Qwen2.5-7B-Instruct.

Head ensemble size. Figure 2 examines whether averaging multiple top-ranked heads improves robustness. Contrary to ensemble intuition, using Top-3, Top-5, or Top-10 heads monotonically degrades F1 from 82.7 to 81.2. The sparse nature of retrieval heads (Wu et al., 2025) means that lower-ranked heads carry non-grounding attention patterns; their inclusion introduces noise that outweighs any variance-reduction benefit. This means that the head with the highest score in citation head identification stage is indeed more accurate in the citation annotation task, validating our single-head selection strategy.

5 Conclusion

We presented a training-free framework that reads fine-grained citations from retrieval-head attention instead of generating citation tokens. Through semantic head probing, normalize-then-sum aggregation, and peak-minus-entropy abstention, the method turns internal evidence-use traces into sentence-level attributions. On LongBench-Cite, it achieves 82.7–86.4 F1 and consistently surpasses prompting, post-hoc matching, and generated citations from fine-tuned models, while preserving the original answer distribution. Future work can extend attention readout to richer head combinations and broader long-context architectures.

Limitations

While our study focuses on validating the effectiveness of calibration-driven LoRA merging across representative instruction-tuned models and task domains, several promising directions remain for future exploration. First, the current framework

can be extended to a broader range of model families, adapter architectures, and larger expert collections to further examine its scalability. Second, the calibration signals used in this work may be enriched with additional internal model statistics, enabling more fine-grained allocation at different layers, heads, or modules. Finally, beyond the evaluated benchmarks, future work may investigate how such block-aligned merging strategies interact with continual expert updates and deployment-time efficiency constraints in practical multi-domain systems.

References

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. 2024a. Longalign: A recipe for long context alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Findings of ACL, pages 1376–1395. Association for Computational Linguistics.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024b. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 3119–3137. Association for Computational Linguistics.

Steven Bird. 2006. NLTK: the natural language toolkit. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*. The Association for Computer Linguistics.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, and 3 others. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

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. 2025. SelfCite: Self-supervised alignment for context attribution in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 10839–10858. PMLR.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. 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, Singapore. Association for Computational Linguistics.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. Dureader: a chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering@ACL 2018, Melbourne, Australia, July 19, 2018*, pages 37–46. Association for Computational Linguistics.

Or Honovich, Roee Roitman, Michael Lucas, Yaakov Shalev, Roee Novikova, and Boaz Carmeli. 2022. TRUE: Evaluating factual consistency in knowledge-grounded text generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3464–3476.

Luyang Huang, Shuyang Cao, Nikolaus Nova Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1419–1436. Association for Computational Linguistics.

Infiniflow. 2024. RAGFlow: Open-source RAG engine.

Yuzheng Jiang and 1 others. 2024. Knowledge circuits in pretrained transformers. *arXiv preprint arXiv:2405.17969*.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Greg Kamradt. 2023. Needle in a haystack – pressure testing LLMs. `https://github.com/gkamradt/LLMTest_NeedleInAHaystack`. GitHub repository.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Harsh Maheshwari, Srikanth Tenneti, and Alwarappan Nakkiran. 2025. Citefix: Enhancing RAG accuracy through post-processing citation correction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 310–317, Vienna, Austria. Association for Computational Linguistics.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, and 1 others. 2021. Webgpt: Browser-driven question answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, and 1 others. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. 2024. 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, Miami, Florida, USA. Association for Computational Linguistics.

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

Hanlin Tang, Yang Lin, Jing Lin, Qingsen Han, Shikuan Hong, Yiwu Yao, and Gongyi Wang. 2025. Razorattention: Efficient KV cache compression through retrieval heads. In *International Conference on Learning Representations (ICLR)*.

Llama Team. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Qwen Team. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Boshi Wang and 1 others. 2024. How to think step-by-step: A mechanistic understanding of chain-of-thought reasoning. *arXiv preprint arXiv:2402.18312*.

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

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

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. 2025. Duoattention: Efficient long-context LLM inference with retrieval and streaming heads. In *International Conference on Learning Representations (ICLR)*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics.

Yue Yu, Ting Bai, HengZhi Lan, Li Qian, Li Peng, Jie Wu, Wei Liu, Jian Luan, and Chuan Shi. 2026. C^2-Cite: Contextual-aware citation generation for attributed large language models. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*. Association for Computing Machinery.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. 2025. LongCite: Enabling LLMs to generate fine-grained citations in long-context QA. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5098–5122, Vienna, Austria. Association for Computational Linguistics.

Yuxuan Zhang and 1 others. 2024. On the role of attention heads in large language model safety. *arXiv preprint arXiv:2410.13708*.

Rui Zheng and 1 others. 2025. Improving contextual faithfulness of large language models via retrieval heads-induced optimization. *arXiv preprint arXiv:2501.13573*.

Dataset	Source	Avg. Len	#Data
MultiFieldQA-en	Multi-field	4,559	150
MultiFieldQA-zh	Multi-field	6,701	200
HotpotQA	Wikipedia	9,151	200
DuReader	Baidu Search	15,768	200
GovReport	Gov. Report	8,734	200
LongBench-Chat	Real-world	35,571	50

Table 4: Dataset statistics in LongBench-Cite. Avg. Len denotes average words (English) or characters (Chinese).

Method	Threshold Meaning	Value
Post-hoc (BGE-M3)	Cosine similarity of BGE-M3 embeddings	0.70
Our method	Aggregated attention score − normalized entropy	−0.70
Our method w/o entropy filtering	Aggregated attention score	0.15

Table 5: Thresholds selected via linear search on the validation set.

A Dataset Details

LongBench-Cite statistics. Table 4 summarizes the five datasets in LongBench-Cite, including MultiFieldQA-en/zh (Bai et al., 2024b), HotpotQA (Yang et al., 2018), DuReader (He et al., 2018), GovReport (Huang et al., 2021) and LongBench-Chat (Bai et al., 2024a).

B Hyperparameters

We conduct a linear search for the optimal threshold on a validation set of 50 examples, which is a 1/20 subset of LongBench-Cite. The selected thresholds of all methods are reported in Table 5.

For the post-hoc BGE-M3 matching, we retrieve source sentences whose embedding cosine similarity to the generated clause exceeds 0.70. For our method, under the peak-minus-entropy rule, a candidate sentence $j \in \mathcal{J}_i$ must satisfy $\psi_{ij} \geq -0.70$ to be emitted as a citation; this negative threshold reflects the trade-off between peak confidence and entropy regularization. When ablating the entropy term, the raw aggregated attention score threshold is set to 0.15, which is substantially higher due to the lack of entropy penalization.

C Prompt Templates

We use three distinct prompts for model inference across all experiments.

Vanilla long-context QA prompt. For the vanilla long-context QA strategy (C_{LQA}), we use a simple text-based question-answering prompt that instructs the model to answer based on the provided document without any citation requirements.

LQAC one-shot prompt. For the LQAC strategy in ICL experiments, we adopt the one-shot prompt provided by Zhang et al. (2025), which includes a demonstration example showing the sentence-level citation format with `<statement>` and `<cite>` tags.

LQAC zero-shot prompt. For the fine-tuned LongCite-8B model, we use the same LQAC prompt structure but remove the one-shot demonstration example, relying on the model's fine-tuned capacity to generate citations directly. It is consistent with the prompt template used in LongCite-45k training data.

The full text of each prompt is provided below.

C.1 Vanilla long-context QA prompt

Vanilla long-context QA prompt

<book> `[document text here]` </book>
Based on the content of the book, answer the following question in detail.
Question: `[question here]`
Answer:

C.2 LQAC one-shot prompt

LQAC one-shot prompt (for ICL)

Please answer the user's question based on the given document. When a factual statement S in your response uses information from some chunks in the document (i.e., <C{s1}>-<C{e1}>, <C{s2}>-<C{e2}>, ...), please append these chunk numbers to S in the format "<statement>{S}<cite>[{s1}-{e1}][{s2}-{e2}]...</cite></statement>". For other sentences such as such as introductory sentences, summarization sentences, reasoning, and inference, you still need to append "<cite></cite>" to them to indicate they need no citations. You must answer in the same language as the user's question.
Here is an example:
`[an example here]`
Now get ready to handle the following test case.
[Document Start]
`[document text here]`
[Document End]
[Question]
`[question here]`
[Remind]
Please answer the user's question based on the given document. When a factual statement S in your response uses information from some chunks in the document (i.e., <C{s1}>-<C{e1}>, <C{s2}>-<C{e2}>, ...), please append these chunk numbers to S in the format "<statement>{S}<cite>[{s1}-{e1}][{s2}-{e2}]...</cite></statement>". For other sentences such as such as introductory sentences, summarization sentences, reasoning, and inference, you still need to append "<cite></cite>" to them to indicate they need no citations. You must answer in the same language as the user's question.
[Answer with Citations]

C.3 LQAC zero-shot prompt

LQAC zero-shot prompt (for LongCite-8B)

Please answer the user's question based on the given document. When a factual statement S in your response uses information from some chunks in the document (i.e., <C{s1}>-<C{e1}>, <C{s2}>-<C{e2}>, ...), please append these chunk numbers to S in the format "<statement>{S}<cite>[{s1}-{e1}][{s2}-{e2}]...</cite></statement>". You must answer in the same language as the user's question.
[Document Start]
`[document text here]`
[Document End]
`[question here]`

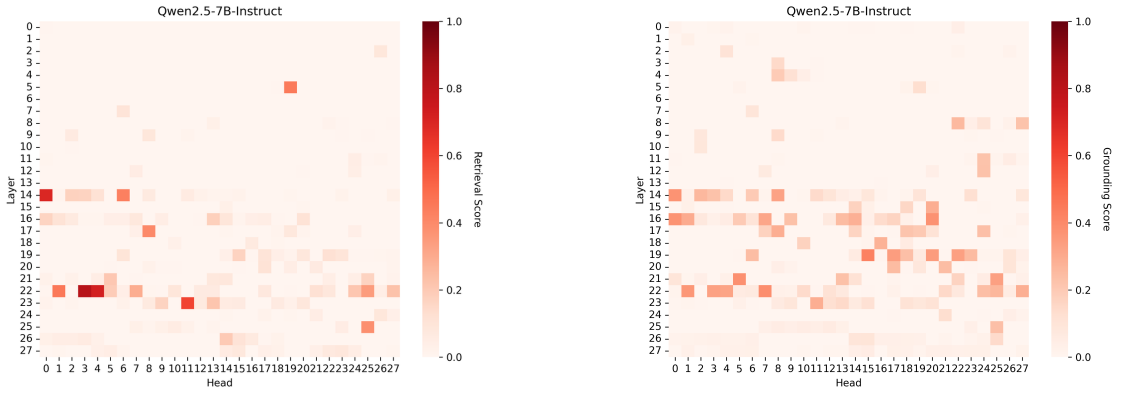

Figure 3: Attention head score heatmaps for Qwen2.5-7B-Instruct. **Left:** Token-level needle-in-a-haystack probe (Wu et al., 2025), scoring heads by verbatim copying accuracy. **Right:** Our semantic citation-head probe (§3.2), scoring heads by sentence-level grounding. Darker colors indicate higher scores.

Rank	Head (Layer, Index)
1	(19, 15)
2	(22, 7)
3	(21, 5)
4	(16, 20)
5	(14, 0)
6	(16, 0)
7	(22, 1)
8	(19, 20)
9	(19, 17)
10	(19, 22)

Table 6: Top-10 retrieval heads on Qwen2.5-7B-Instruct identified our semantic probing procedure.

D Citation Head Analysis

Comparison of probing methods. Figure 3 compares the attention head activation patterns of Qwen2.5-7B-Instruct under two different probing paradigms. The left heatmap follows the verbatim needle-in-a-haystack protocol from Wu et al. (2025), where heads are scored by token-level copying accuracy (i.e., whether the top-attended token matches the source token being reproduced). The right heatmap uses our semantic probing procedure described in §3.2, which rewards attention to the entire evidence sentence that semantically supports the generated clause.

Both methods reveal the **sparse** nature of citation-relevant heads: only a small fraction of all heads exhibit strong grounding signals. Several heads are activated under both probes (e.g., head (22, 3)), confirming that some citation heads do perform verbatim recall. However, the overall patterns differ substantially—our semantic probe detects additional heads in upper layers that attend to paraphrased or synthesized evidence, which the token-level probe misses.

Top citation heads. Table 6 lists the top-10 citation heads (layer index, head index) identified on Qwen2.5-7B-Instruct. Table 7 lists the top-1 citation head for each model in our experiments.

Intrinsic property of citation heads. Wu et al. (2025) establishes that retrieval heads are an **intrinsic property** of the base model: they emerge from large-scale pretraining and remain stable through subsequent training stages (continued pretraining, SFT, RLHF). Their experiments show Pearson correlations > 0.8 between base and chat variants of the same model family, while cross-family correlations are < 0.1.

We leverage this invariance for our experiments on LongCite-8B, which is fine-tuned from Llama-3.1-8B. Since the retrieval head set remains unchanged through supervised fine-tuning (Wu et al., 2025), we use the **same top-1 citation head** identified on the base Llama-3.1-8B-Instruct for LongCite-8B

Model	Layer Index	Head Index
Qwen2.5-7B-Instruct	19	15
Qwen3-8B	23	10
Llama-3.1-8B-Instruct	13	18
LongCite-8B	13	18

Table 7: Top-1 citation head identified by our semantic probing procedure for each model.

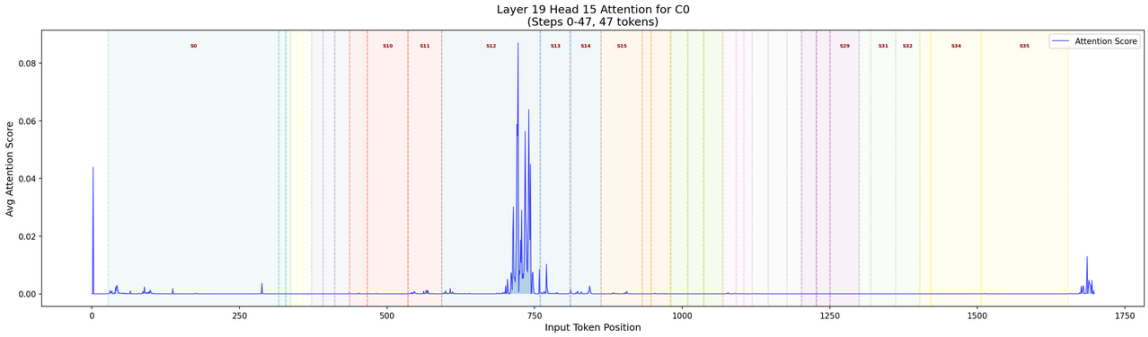

Figure 4: Attention distribution of the top-1 citation head when generating a cited clause c_0.

evaluation. This avoids re-probing the fine-tuned model and demonstrates that our method generalizes across training stages without additional overhead.

E Attention Distribution Analysis

To verify that the selected citation head indeed tracks grounding evidence during inference, we visualize its token-level attention distribution on real examples from the LongCite-45k dataset.

Case 1: Cited clause with concentrated attention. Figure 4 shows the attention distribution of the top-1 citation head when generating a factual clause c_0 that requires citation. The x-axis denotes input token positions, and the y-axis shows the average attention weight. Colored vertical bands mark sentence boundaries in the source document. In this case, c_0 should cite the latter half of source sentence s_{12}. The attention vector exhibits a pronounced peak precisely over the token span of s_{12}'s latter half, with negligible attention mass on unrelated sentences. This confirms that the citation head localizes supporting evidence with high spatial precision.

Case 2: Non-cited clause with diffuse attention. Figure 5 shows the attention distribution when generating a functional clause c_5 that requires no citation. Although c_5 lacks direct supporting evidence in the source, the citation head still activates over multiple related sentences—but the attention mass is *diffuse* across several locations rather than concentrated on a single span.

These two observations directly motivate the design choices in §3.3. **First**, for cases like c_0 where only part of a sentence serves as evidence, our **normalize-then-sum aggregation** (Eq. 9) accurately reflects the total attention mass projected onto each source sentence. A mean-aggregation alternative would dilute the evidence-bearing tokens by averaging over the entire sentence span, making the citation score sensitive to sentence length and causing the model to miss partial-sentence evidence. **Second**, the contrast between Figures 4 and 5 validates the **peak-minus-entropy criterion** (Eq. 11). The cited clause produces a *peaked* distribution with low entropy, yielding $\psi_{ij} > \tau$ and a confident citation; the functional clause produces a *diffuse* distribution with high entropy, yielding $\psi_{ij} < \tau$ and abstention. Without entropy regularization, both cases might trigger citations based solely on attention magnitude, leading to false positives on transitional sentences.

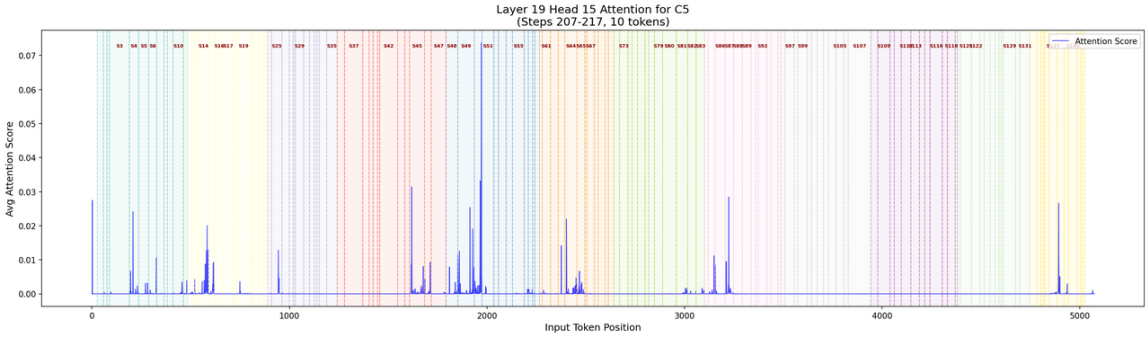

Figure 5: Attention distribution of the top-1 citation head when generating a non-cited clause c_5.

F Case Study

F.1 Multi-hop reasoning on HotpotQA

Table 8 presents a qualitative example from LongBench-Cite, illustrating how our attention-based citation pipeline operates on a real query-context-response triple.

Query	The Panel with striding lion was one of many that lined **the way north of the gate to the inner city of what location**?
Context	... <c148>: "Passage 6: *Ishtar Gate* **The Ishtar Gate was the eighth gate to the inner city of Babylon** (in the area of present-day Hillah, Babil Governorate, Iraq)." , <c149>: "It was constructed circa 575 BCE by order of King Nebuchadnezzar II on the north side of the city.", <c150>: "It was part of a grand walled processional way leading into the city.", <c151>: "The original structure was a double gate with a smaller frontal gate and a larger and more grandiose secondary posterior section.", ... <c155>: "*History* King Nebuchadnezzar II reigned 604–562 BCE, the peak of the Neo-Babylonian Empire.", ... <c444>: "Passage 8: *Panel with striding lion* The Panel with striding lion (MA 31.13.1) is a panel of Neo-Babylonian glazed ceramic bricks or tiles dated to 604–562 B.C., now in the Metropolitan Museum of Art, New York.", <c445>: "It was one of many that **lined the Processional Way north of the Ishtar Gate.**" , <c446>: "It was excavated by R. Koldewey in 1902, and at the Staatliche Museen zu Berlin from 1926, before coming into the possession of the Met in 1931.A large group of such figures is part of the Processional Way leading to the Ishtar Gate, a centrepiece display of the Pergamon Museum in Berlin.", ...
Response (our method)	The Panel with striding lion was one of many that lined the way north of the gate to the inner city of **Babylon**. **[148][149][445]**
Response (Posthoc)	The Panel with striding lion was one of many that lined the way north of the gate to the inner city of **Babylon**. **[no citation]**
Response (ICL)	<statement> The Panel with striding lion was one of many that lined the way north of the gate to the inner city of **Babylon**. <cite> **[149-151][155-155]** </cite></statement>

Table 8: Qualitative example from LongBench-Cite on Qwen3-8B.

This table demonstrates a case from the HotpotQA subset of LongBench-Cite, evaluated on Qwen3-8B under our method, one-shot in context learning, and post-hoc BGE-M3 matching. The query asks about the location of the Panel with striding lion, with the correct answer being **Babylon**. All three methods produce the correct answer, yet their citation annotations differ substantially.

This case exemplifies a **2-hop reasoning** problem: the model must first locate the Panel with striding lion at the Ishtar Gate (corresponding to source sentences c444–c446), then trace the Ishtar Gate to the

city of Babylon (corresponding to c148–c149). The core evidence supporting the final answer is c148 and c445.

Our method correctly identifies all three supporting sentences (c148, c149, c445), capturing both hops of the implicit reasoning chain. Although the model does not explicitly output its reasoning path, the attention signal faithfully reflects which context tokens were accessed during answer generation.

Post-hoc BGE-M3 matching fails to emit any citation. Because the response involves summarization and synthesis, no single source sentence exhibits sufficient semantic similarity to the generated clause, causing the similarity threshold to filter out all candidates.

ICL prompting produces an incorrect citation span. While the answer itself is correct, the model self-generates [149-151][155-155], which misses the critical c148 (establishing Babylon as the city) and c445 (linking the lion panel to the Ishtar Gate), while incorrectly including c155 (irrelevant historical background).

F.2 Attention-grounded vs. self-generated citations on LongCite-8B

Table 9 demonstrates a case where our attention readout corrects a spurious self-generated citation from LongCite-8B. For the same generated clause, we compare the citation produced by the model's own output token against the citation recovered from our attention-based pipeline.

Query	The 53rd National Hockey League All-Star Game took place at the indoor arena that was completed in what year?
Context	... <c47>: "Passage 3: *53rd National Hockey League All-Star Game* **The 2003 National Hockey League All-Star Game took place on February 2, 2003 at Office Depot Center in Sunrise, the home of the Florida Panthers.**", ... <c238>: "Passage 8: *BB&T Center (Sunrise, Florida)* FLA Live Arena (previously known as the National Car Rental Center, Office Depot Center, BankAtlantic Center, and BB&T Center) is an indoor arena located next to Sawgrass Mills in Sunrise, Florida.", ...
Response (our method)	Based on the passage, **the 53rd National Hockey League All-Star Game took place at the Office Depot Center in Sunrise, Florida**, which was completed in 1998. **[47]** .....
Response (response)	<statement> Based on the passage, **the 53rd National Hockey League All-Star Game took place at the Office Depot Center in Sunrise, Florida**, which was completed in 1998. <cite> **[238-238]** </cite></statement> .....

Table 9: Qualitative example from LongBench-Cite on Longcite-8B.

For the clause stating that the game took place at the Office Depot Center, the model's self-generated citation points to c238, which merely describes the arena's current name and location without mentioning the All-Star Game. This is a **spurious citation**: while semantically related to the arena, it provides no direct evidence for the specific claim.

In contrast, our attention readout identifies c47 as the supporting source. This sentence explicitly states that "**The 2003 National Hockey League All-Star Game took place . . . at Office Depot Center in Sunrise**", directly grounding the generated clause. The attention signal thus reveals that the model internally attended to the correct evidence during generation, even though its surface-form citation output was erroneous.

This case supports our hypothesis that LongCite's supervised fine-tuning teaches citation *formatting* (producing «cite» tags) more effectively than it teaches accurate *grounding*. The internal attention trace provides a more faithful signal of evidence use than the model's own citation tokens, suggesting that attention-based readout can serve as a lightweight verification layer even for fine-tuned citation models.