CITEPROBE: FROM LLM ATTENTION TO CITATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sentence-level citations make Retrieval-Augmented Generation (RAG) outputs easy to verify by identifying the source sentences that support a generated answer. Yet existing methods face a trade-off between effectiveness and efficiency. Perturbation-based approaches require repeated LLM inference, while attention-based approaches typically rely on fixed aggregation rules that obscure informative variations across layers and heads. Can accurate citations instead be directly decoded from the attention weights without any additional pass? We show that they can, and that signals are concentrated in a few heads. To leverage this, we introduce CITEPROBE, a lightweight approach that formulates sentence-level citation as a readout problem over the attention weights of a frozen LLM. For each candidate sentence, CITEPROBE constructs an attention profile capturing the attention it receives from response tokens across individual layers and heads, thereby preserving the full layer–head structure. A lightweight, model-specific linear probe then learns to map these attention profiles to citation scores. This formulation exploits citation-relevant signals already present in the attention weights and scores all candidate sentences using a single generation pass, avoiding the repeated LLM inference required by perturbation-based methods. Experiments across six LLMs and four datasets show that CITEPROBE consistently outperforms existing attribution methods while substantially reducing inference cost. These results demonstrate that sentence-level citation signals are directly decodable from structured LLM attention, offering an effective and efficient approach to citation attribution. We also show that CITEPROBE can be applied to black-box LLMs under surrogacy mode. We release CITEPROBE as open source at https://www.github.com/citeprobe/citeprobe.

1 INTRODUCTION

Large Language Models (LLMs) are often equipped with Retrieval-Augmented Generation (RAG) to perform analysis and reporting tasks over large document collections (Lewis et al., 2020; Edge et al., 2024). While this may spare users the laborious work of perusing the source documents, verifying the answers can require doing that work all over again, while skipping verification comes at the expense of safety. Therefore, feasible verification of LLM-RAG output remains an open problem (Choi et al., 2026; Schreieder et al., 2026; Wei et al., 2026). To address this, sentence-level citations have attracted increasing interest, closing this gap by identifying the source sentences that support an answer (Zhang et al., 2025; Wei et al., 2026).

Existing sentence-level citation methods can be divided into two categories, perturbation-based and attention-based methods. Perturbation-based attribution methods like ContextCite (Cohen-Wang et al., 2024) and ARC-JSD (Li et al., 2026) estimate which context sentences influence the model's behavior. Attention-based methods like AttnTrace (Wang et al., 2026) rely on reading model internals and use model latent states to estimate source influence. However, these methods fail to balance the trade-off between effectiveness and efficiency of this task. Specifically, perturbation-based methods require repeated inferences, making them computationally expensive. Attention-based methods reduce repeated inferences, they use coarse-grained aggregation rules, which obscure informative variation across layers and heads, leading to suboptimal performance. An ideal solution should perform well over large contexts while remaining lightweight enough for practical use. Therefore, a question arises: can we perform sentence-level citation **without LLM fine-tuning, modifications to the LLM backbone, or additional inference, while achieving good performance**?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To achieve this goal, we investigate whether we can read signals for sentence-level citation directly from the attention weights produced in a standard forward pass. An LLM autoregressively computes each next-token distribution from representations of the preceding context and partial response. In RAG, the prefix contains retrieved documents, the computation is conditioned on source tokens and may retain signals useful for locating evidence within the documents, such as attention patterns that reflect how generated tokens interact with different source tokens. However, such signals may not be distributed uniformly across an LLM's layers and attention heads, since different heads exhibit distinct behaviors, and prior analyses suggest that a relatively small subset can carry disproportionately useful information (Voita et al., 2019; Kovaleva et al., 2019). We also empirically find that a small subset of heads (we term "citation heads") is sufficient for strong sentence-level citation performance (Section 3.2), motivating us to develop a lightweight method that distinguishes informative heads from all available heads.

Based on this idea, we introduce CITEPROBE, a lightweight readout of frozen attention for sentence-level citation. First, we define an aggregation of the attention each candidate sentence receives from the answer tokens separately for every layer and head. Then, we train a lightweight linear probe on evidence annotations, which is used to turn each aggregated attention grid into a citation confidence score without repeated context perturbations, resulting in efficiency. Finally, we evaluate CITEPROBE across four open-source and two closed-source LLMs on four public datasets, where it outperforms the current SOTA baseline ARC-JSD while requiring only 16.3% of its total runtime. Our analysis reveals that beyond leveraging a subset of citation heads, the learned probe also found "counter-citation heads", whose higher attention to a source sentence is learned as a signal against citing it. Consequently, besides outperforming baselines, CITEPROBE also more often identifies better citations omitted by imperfect annotations, as judged by an LLM.

In summary, our contributions are threefold. **First**, we show that viable sentence-level citation signals can be measured from latent attention states, and the key is to identify informative attention heads. **Second**, we introduce CITEPROBE, which provides a model-specific, supervised readout for sentence-level citation with a frozen LLM and a simple linear probe. **Third**, we conduct extensive experiments on six LLMs and four public datasets, with results demonstrating the superior performance of CITEPROBE in both accuracy and efficiency.

2 RELATED WORK

Existing sentence-level methods are mainly training-free and can be divided into two categories: perturbation-based and attention-based methods.

Perturbation-based methods estimate a context unit's influence by measuring how the model's behavior changes when that unit is removed. Their scores therefore characterize behavioral dependence under a particular intervention, rather than directly assessing whether a sentence semantically supports the generated answer. ContextCite fits a surrogate over random context ablations (Cohen-Wang et al., 2024), while ARC-JSD deletes one candidate at a time and measures Jensen-Shannon divergence in the output distribution (Li et al., 2026). Other perturbation-based approaches use approximation or span search (Liu et al., 2025; Wang et al., 2025). Perturbation-based methods require repeated evaluations of modified contexts. In contrast, CITEPROBE predicts annotated supporting evidence from a single attention trace over the unmodified context and generated answer.

Attention-based methods score source tokens or sentences using the model's attention weights, sometimes combined with gradients. Think-Twice (Hsu et al., 2026) scores sentences from head-averaged attention over selected tokens and re-scores against a deliberately incomplete answer to recover missed sources. Grad-SAM (Barkan et al., 2021) combines gradients with self-attention to explain Transformer predictions, which requires a backward pass. AttnTrace (Wang et al., 2026) instead scores long-context sources using top-K token attention and repeated context subsampling. These existing approaches require backward passes, repeated context evaluations, or ablation-based training supervision. Instead, our method trains a lightweight linear probe to score citations directly from the same inference pass, completely avoiding heavy computational overheads that made prior methods difficult to adopt.

Moreover, there are also some generative methods (Chuang et al., 2025; Xia et al., 2025; Wei et al., 2026) that fine-tune LLMs to generate sentence-level citations, but they typically require 4–8

NVIDIA A100 GPUs, substantially raising the hardware barrier. In contrast, our method can be economically trained on a single RTX 4090 GPU, making it easier to apply. Although Guidance (Zhan et al., 2026) is a training-free generative method, it requires an LLM to preprocess source documents into a prescribed structure, making it computationally heavy as well (see Appendix K).

3 METHODOLOGY

We build CITEPROBE on the hypothesis that citation-relevant information is present in individual attention heads but may be obscured by all-head averaging. To verify it, we first formulate a head-specific representation of attention that preserves the contribution of individual heads. Then, we conduct a preliminary exploration to determine whether selecting informative heads alone is sufficient for citation prediction. Finally, we complete the formulation with a lightweight linear readout that learns how to integrate these head-specific signals, simplifying and streamlining the overall procedure. We formally define the task as follows:

Task definition. We define sentence-level citation as the post-hoc identification of source sentences that support a generated answer. Let M be an LLM, q a question, $\mathcal{C} = (s_1, \ldots, s_n)$ a context segmented into candidate source sentences, and y an answer generated by M. We refer to each question-context-answer triple $(q, \mathcal{C}, y)$ as an example, indexed by e. Each candidate sentence s_j is assigned a binary evidence label $\eta_j \in \{0, 1\}$, where $\eta_j = 1$ if s_j is annotated as supporting evidence for y, or 0 otherwise. We use $\mathcal{S}$ to denote a set of attention heads in M, with $k = |\mathcal{S}|$. Our objective is to assign a score $r_j = r(s_j \mid q, \mathcal{C}, y, M)$ to each candidate sentence, aiming to place claim-supporting source sentences above non-supporting candidates, thereby identifying citations.

3.1 FROM HEAD-SPECIFIC ATTENTION WEIGHTS TO CITATION SCORES

First, we construct a layer-head representation that preserves head identity. We index a head by the pair $i = (l, h)$ of its layer l and its position h, and let $\alpha_{t,u}^{(i)}$ be the attention from answer token t to source token u at head i. Let $\mathcal{T}_y$ denote the answer-token positions and $\mathcal{T}_j$ the positions belonging to candidate sentence s_j. For each head, we summarize how much the generated answer attends to a candidate sentence by

$$G_{i,j} = \sum\nolimits_{t \in \mathcal{T}_y} \sum\nolimits_{u \in \mathcal{T}_j} \alpha_{t,u}^{(i)}. \tag{1}$$

Stacking these values gives a layer-by-head grid $G_j \in \mathbb{R}^{L \times H}$ for sentence s_j, with L layers and H heads per layer. Unlike simple attention averaging, it preserves head identity, so that each head remains tied to a specific computation in the frozen LLM. Then, to measure whether a head assigns unusually high attention to a sentence relative to the others, we standardize each head's sentence scores within each example e to isolate differences in attention scale:

$$\widetilde{G}_{i,j}^{(e)} = \frac{G_{i,j}^{(e)} - \mu_i^{(e)}}{\sigma_i^{(e)}}, \tag{2}$$

where $\mu_i^{(e)}$ and $\sigma_i^{(e)}$ are the mean and standard deviation of $G_{i,j}^{(e)}$ over the candidate sentences $j \in \mathcal{C}$ in example e. Finally, for a given set $\mathcal{S}$ of k selected attention heads, we obtain the score of each candidate sentence by averaging the standardized scores of those heads:

$$\text{score}_k(s_j) = \frac{1}{k} \sum\nolimits_{i \in \mathcal{S}} \widetilde{G}_{i,j}^{(e)}. \tag{3}$$

3.2 HEAD SELECTION EXPLORATION

To verify the feasibility of our formulation, we first identify the best-performing attention heads by measuring how strongly their attention weights correlate with the benchmark evidence labels. This tests whether citation-relevant information can be recovered by distinguishing the contributions of specific heads ("citation heads"), before introducing a learned readout. For comparison, we use the average-attention baseline from ContextCite (Cohen-Wang et al., 2024), which scores each sentence by uniformly averaging its raw attention weights $G_{i,j}$ across all available heads.

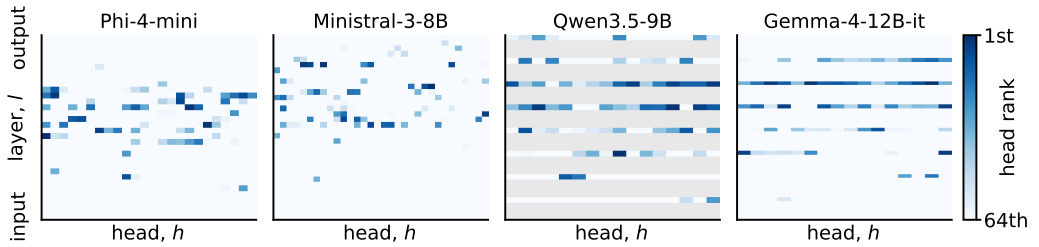


Figure 1: Locations of selected citation heads across layers. Gray rows in Qwen3.5-9B are excluded linear-attentions. Gemma-4-12B-it's selected heads concentrate in periodic full-attention layers.

Selecting citation heads. For each head i, we compute the point-biserial correlation between its sentence scores $\widetilde{G}_{i,j}^{(e)}$ and the binary evidence labels $\eta_j^{(e)}$ within each training example e, and aggregate following standard practice using Fisher's z-transform to correct for skewness (Fisher, 1921):

$$\rho_{i,e} = \mathrm{Corr}_j\left(\widetilde{G}_{i,j}^{(e)}, \eta_j^{(e)}\right), \qquad \bar{\rho}_i = \text{Fisher-z}_e(\rho_{i,e}). \tag{4}$$

We rank heads by $\bar{\rho}_i$ and sweep over increasing values of k, expanding $\mathcal{S}$ to include the top k heads in ranked order. At each k, we average their scores using Equation 3 and measure mAP and Top-1 on held-out examples. We refer to the selected heads as *citation heads*.

Observations. We investigate the effects of head selection for sentence-level citations across four open-source LLMs on four public datasets (see Section 4.1 for details) and use the harmonic mean of mAP and Top-1 accuracy as an evaluation metric. The results in Figure 2 show that selecting a subset of heads outperforms uniform all-head averaging, and adding more heads eventually degrades performance. Figure 1 further shows that most of this gain emerges with only a few top-ranked heads that are generally concentrated in the middle layers. However, strong performance from head selection alone could also imply overfitting the head selection to a particular dataset. We examine this through the overlap in top-k heads across datasets. Figure 2 (bottom) shows that the chance-corrected overlap is low at small k, meaning substantial selection variations across datasets. This stabilizes above 80% at $k = 16$. Together, these results show that **a small subset of heads is sufficient for impressive sentence-level citation performance** (details in Appendix A).

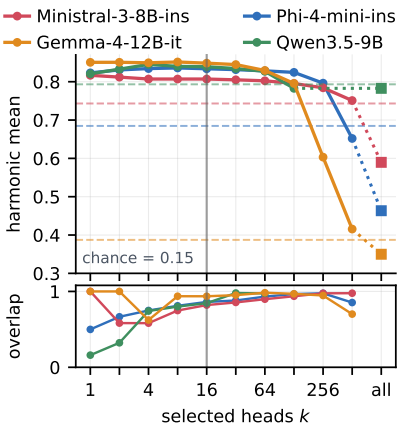


Figure 2: **Top:** harmonic mean of mAP and Top-1 versus top-k heads. Dashed lines denote each model's uniform all-head average. **Bottom:** chance-corrected overlap of the top-k heads.

3.3 CITEPROBE: A LIGHTWEIGHT READOUT FOR SENTENCE-LEVEL CITATION

Our head selection exploration shows that attention contains citation signals that vary substantially across heads. To capture this, we propose a lightweight method, named CITEPROBE, which uses logistic regression as the canonical readout due to its computational efficiency and simplicity. Moreover, the learned weights provide natural head-level interpretability, while the underlying LLM remains entirely frozen, thereby reducing training overhead. The workflow of CITEPROBE is illustrated in Figure 3. Specifically, it uses the sentence-level attention representation from Equations 1 and 2, flattens the layer-head grid into a feature vector x_j, and predicts whether sentence s_j supports the answer. Formally:

$$x_j = \mathrm{vec}(\widetilde{G}_j), \qquad \text{where } \widetilde{G}_j = \left[\widetilde{G}_{i,j}\right]_{i=(l,h)} \in \mathbb{R}^{L\times H}, \qquad \widehat{p}_j = \mathrm{sigmoid}(w^\top x_j + b). \tag{5}$$

Each probe coefficient $w_{l,h}$ gives the signed contribution of one identifiable head. Beyond our canonical linear readout, to test whether added capacity helps, we also train MLP, CNN, Transformer, and RetNet readouts. For the CNN we retain the unflattened layer-head grid so that convolutions can exploit its spatial structure.

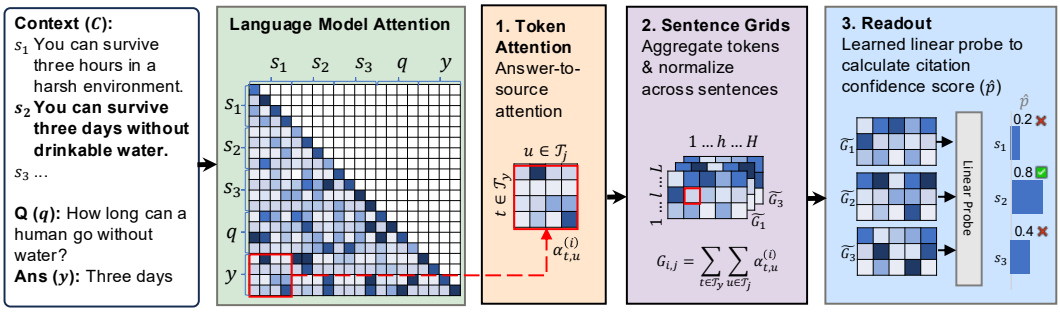


Figure 3: Workflow of CITEPROBE: Attention to Citation from a frozen LLM in a single inference pass. (1) For each layer and head, attention from answer tokens to each source sentence is summed. (2) Aggregation over answer tokens yields a normalized $L \times H$ grid per sentence. (3) A lightweight linear probe predicts the citation score.

Training. We use class-weighted sentence-wise binary cross-entropy as the loss function for training, which is computed as follows:

$$\mathcal{L} = -\frac{1}{n} \sum_j \left[\lambda \eta_j \log \widehat{p}_j + (1 - \eta_j) \log(1 - \widehat{p}_j) \right], \tag{6}$$

where $\lambda = n_-/n_+$ balances the contributions of supporting sentences (n_+) and non-supporting sentences (n_-) in the training data. The resulting score $\widehat{p}_j \in (0, 1)$ supports both ranking candidates and selecting citations by applying a threshold.

Inference. At inference, CITEPROBE reuses attention from the original generation pass, requiring no additional model pass, unless in surrogate mode. We further develop a fused kernel that accumulates sentence-level attention sums without materializing the full attention matrix, enabling compatibility with FlashAttention (Dao et al., 2022) and similar implementations (Appendix C).

4 EXPERIMENTS

We conduct experiments on four datasets and six LLMs from different families to verify the effectiveness of our method, with the goal of answering five research questions (RQs):

RQ1: Does CITEPROBE improve citation performance compared with baselines?
RQ2: What does CITEPROBE's learned weighting contribute beyond head selection?
RQ3: Is CITEPROBE's performance gains driven by lexical matching?
RQ4: What explains failures to cite correctly?
RQ5: How does CITEPROBE's computational cost compare with that of baselines?

4.1 EXPERIMENTAL SETUP

Dataset. We evaluate on four benchmarks whose supporting-evidence annotations can be mapped to candidate source sentences: HotpotQA (Yang et al., 2018) and MuSiQue (Trivedi et al., 2022) for multi-hop question answering, English TyDiQA-GoldP (Clark et al., 2020) for single-document question answering, and WiCE (Kamoi et al., 2023) for claim verification. From each benchmark, we sample 500 training, 50 validation, and up to 1,000 test examples, subject to dataset availability. The exception is TyDiQA-GoldP, which yields 366 usable English test examples after filtering out contexts containing fewer than three sentences.

Experimental settings. We adopt a leave-one-out evaluation across the four datasets. To improve robustness to training randomness, we train each probe with three random seeds and report averaged results. Checkpoints are selected by the harmonic mean of validation mAP and Top-1.

Evaluation metrics. We report mean average precision (mAP), following AttriBoT's evaluation (Liu et al., 2025), and Top-1 accuracy, following ARC-JSD (Li et al., 2026). Overall scores are unweighted means across the four target datasets.

Backbones and baselines. We choose LLMs from different families and with varying scales as backbones, comprising Phi-4-mini-instruct (Microsoft, 2025), Ministral-3-8B-Instruct (Mistral

Table 1: Overall performance comparison, where ↑ indicates that higher values are better. Among numerical results, **bold** and underlined values denote the best and second-best results for each backbone, respectively. Baseline groups are marked by ♣ (perturbation-based), ♥ (attention-based attribution), and ♦ (diagnostic baselines). For Claude Opus 5 and GPT-5.6 Sol, non-BM25 results use the best surrogate models denoted in superscripts, M for Ministral, G for Gemma, and Q for Qwen.

Backbone	Method	HotpotQA		MuSiQue		TyDiQA		WiCE		Mean	
		mAP↑	Top-1↑	mAP↑	Top-1↑	mAP↑	Top-1↑	mAP↑	Top-1↑	mAP↑	Top-1↑
Phi-4-3.8B	ContextCite-32 ♣	0.682	0.799	0.637	0.829	0.871	0.798	0.671	0.736	0.715	0.790
	ARC-JSD ♣	0.821	0.851	0.743	0.824	0.887	0.817	0.732	0.751	0.796	0.811
	AttnTrace ♥	0.340	0.178	0.298	0.165	0.784	0.654	0.526	0.439	0.487	0.359
	BM25 ♦	0.783	0.838	0.762	0.835	0.887	0.810	0.685	0.679	0.779	0.790
	Top-16 heads ♦	**0.842**	0.857	0.777	0.861	0.897	0.837	0.791	0.810	0.827	0.841
	CITEPROBE (Ours)	0.834	**0.865**	**0.789**	**0.874**	**0.908**	**0.852**	**0.793**	**0.835**	**0.831**	**0.856**
Ministral-3-8B	ContextCite-32 ♣	0.721	0.782	0.680	0.830	0.847	0.761	0.678	0.708	0.732	0.770
	ARC-JSD ♣	0.803	0.803	0.773	0.810	0.838	0.745	0.704	0.695	0.779	0.763
	AttnTrace ♥	0.787	0.805	0.748	0.791	0.819	0.691	0.743	0.759	0.774	0.762
	BM25 ♦	0.778	0.821	0.786	0.833	0.860	0.765	0.696	0.698	0.780	0.779
	Top-16 heads ♦	**0.835**	0.848	0.781	0.815	0.869	0.786	0.760	0.763	0.811	0.803
	CITEPROBE (Ours)	0.831	**0.869**	**0.811**	**0.863**	**0.874**	**0.793**	**0.773**	**0.817**	**0.822**	**0.835**
Qwen3.5-9B	ContextCite-32 ♣	0.331	0.435	0.249	0.407	0.747	0.608	0.382	0.366	0.427	0.454
	ARC-JSD ♣	0.809	0.776	0.770	0.773	0.838	0.733	0.723	0.694	0.785	0.744
	AttnTrace ♥	0.774	0.823	0.718	0.801	0.895	0.826	0.760	0.793	0.787	0.811
	BM25 ♦	0.797	**0.878**	**0.835**	**0.900**	**0.915**	**0.868**	0.739	0.781	0.822	**0.857**
	Top-16 heads ♦	**0.831**	0.867	0.809	0.844	0.906	0.846	0.791	0.820	**0.834**	0.844
	CITEPROBE (Ours)	0.817	0.869	0.823	0.876	0.894	0.823	**0.793**	**0.851**	0.832	0.855
Gemma-4-12B	ContextCite-32 ♣	0.468	0.646	0.357	0.569	0.784	0.684	0.475	0.533	0.521	0.608
	ARC-JSD ♣	0.803	0.822	0.756	0.820	0.869	0.795	0.701	0.706	0.782	0.786
	AttnTrace ♥	0.226	0.146	0.289	0.296	0.887	0.830	0.265	0.145	0.417	0.354
	BM25 ♦	0.798	0.870	0.812	0.883	0.912	0.854	0.732	0.765	0.814	0.843
	Top-16 heads ♦	**0.856**	**0.914**	0.786	0.884	0.911	0.861	0.778	0.798	0.833	0.864
	CITEPROBE (Ours)	0.850	0.896	**0.815**	**0.897**	**0.918**	**0.884**	**0.784**	**0.814**	**0.842**	**0.873**
Claude Opus 5	ContextCite-32^M ♣	0.706	0.801	0.713	0.861	0.766	0.624	0.547	0.529	0.683	0.704
	ARC-JSD^G ♣	0.834	0.846	0.797	0.852	0.844	0.748	0.621	0.574	0.774	0.755
	AttnTrace^Q ♥	0.829	0.866	0.832	**0.902**	0.799	0.664	0.694	0.696	0.788	0.782
	BM25 ♦	0.803	0.862	0.837	0.900	0.841	0.726	0.649	0.617	0.782	0.776
	Top-16 heads^Q ♦	**0.858**	**0.882**	0.834	0.871	0.819	0.693	0.692	0.670	0.801	0.779
	CITEPROBE (Ours)^M	0.832	0.880	**0.848**	0.901	**0.873**	**0.796**	**0.741**	**0.777**	**0.824**	**0.839**
GPT-5.6 Sol	ContextCite-32^M ♣	0.648	0.836	0.612	0.862	0.825	0.767	0.633	0.775	0.680	0.810
	ARC-JSD^Q ♣	**0.843**	0.886	**0.806**	**0.932**	0.930	0.897	0.774	0.803	0.838	0.879
	AttnTrace^Q ♥	0.802	0.896	0.777	0.930	0.915	0.871	0.794	0.848	0.822	0.886
	BM25 ♦	0.700	0.837	0.737	0.886	0.911	0.841	0.697	0.752	0.761	0.829
	Top-16 heads^M ♦	**0.843**	**0.915**	0.776	0.913	**0.939**	0.909	**0.813**	**0.860**	**0.843**	0.899
	CITEPROBE (Ours)^G	0.814	0.908	0.799	0.925	0.934	**0.911**	0.797	0.858	0.836	**0.901**

AI, 2025), Qwen3.5-9B (Qwen Team, 2026), and Gemma-4-12B-it (Team, 2026). CITEPROBE is trained for each LLM, using attention over the final answer span while excluding any preceding reasoning trace. For the Qwen3.5 family, which interleaves standard attention with linear-attention layers, we exclude reading from linear-attention layers as they do not expose the token-to-token attention weights required by Equation 1. Furthermore, we include Claude Opus 5 and GPT-5.6 Sol to test if CITEPROBE can be applied to black-box LLMs, operating them in surrogacy mode by teacher-forcing their outputs into an open backbone. For baselines, we select ContextCite-32 (Cohen-Wang et al., 2024) and ARC-JSD (Li et al., 2026) as perturbation-based baselines, and AttnTrace (Wang et al., 2026) as an attention-based baseline. For diagnostic baselines, we include top-16-head (from Section 3.2), and BM25 (Robertson & Zaragoza, 2009) to assess lexical matching for citation.

Training details. Our CITEPROBE is trained using the AdamW optimizer with weight decay 0.01 and batch size 128 on a device with two RTX4090 GPUs. We grid search the learning rate over $[10^{-5}, 3\times10^{-5}, 10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}]$, selecting the checkpoint based on the best validation mAP and Top-1 (harmonic mean) (see details in Appendix D.1).

4.2 MAIN RESULTS (RQ1)

Table 1 shows the experimental results. Overall, CITEPROBE recovers strong sentence-level citation signals from frozen attention and outperforms baselines in most cases. BM25 is a strong baseline

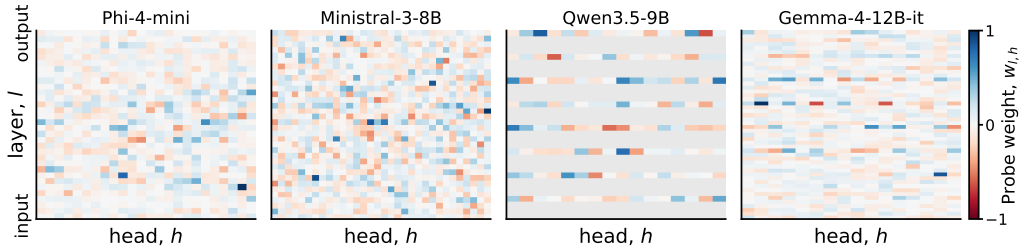


Figure 4: CITEPROBE coefficients $w_{l,h}$, each fit scaled by its largest magnitude and averaged over the twelve leave-one-out fits. Blue is positive, red negative. Linear-attention layers in gray.

because answers often share wording with their supporting citations. However, lexical overlap alone does not establish support: a sentence can repeat the answer's terms without providing the required evidence. Despite this, CITEPROBE achieves higher mean mAP than BM25 for all six LLMs.

BM25 remains particularly competitive on Qwen3.5-9B, narrowly exceeding CITEPROBE in mean Top-1 (0.857 versus 0.855). These results suggest that lexical matching is a nontrivial shortcut in this task, which we examine further in Section 4.4.

Among the attribution baselines, ContextCite-32 and AttnTrace are weaker on average and less consistent across LLMs, while ARC-JSD is the strongest of them. Averaged over the 24 model-dataset pairs, CITEPROBE achieves 0.831 mAP and 0.860 Top-1 accuracy, compared with 0.792 and 0.790 for ARC-JSD, a gain of 3.9 and 7.0 points. Whereas ARC-JSD estimates attribution through repeated context perturbations, CITEPROBE reads supporting-evidence signal directly from the frozen attention representation. This verifies that strong sentence-level citation ranking can be obtained without repeatedly modifying the input context. Moreover, because each reader is trained on three datasets and evaluated on the held-out fourth, these gains hold under cross-dataset transfer rather than in-domain fitting.

4.3 IN-DEPTH ANALYSIS (RQ2)

To understand why CITEPROBE performs well, we conduct experiments in terms of head sensitivity, learned weights, and readout architectures.

Head sensitivity. To test whether CITEPROBE benefits from retaining a broader set of heads, we apply recursive feature elimination (RFE), repeatedly removing heads with the smallest absolute coefficients and refitting. As shown in Figure 5, there is a modest performance improvement as more heads are retained. Our learned probe uses both positive and negative weights, with heads in blue raising the evidence score ("citation heads") and heads in red lowering it ("counter-citation heads"). Unlike simple head selection, CITEPROBE can use a larger set without a substantial loss in performance, simplifying the procedure by avoiding a separate pruning step. In summary, beyond top-k head selection, CITEPROBE leverages substantially all heads, and looks not just for heads that agree with a citation, but also those that object.

Weights visualization. As shown in Figure 4, our learned probe uses both positive and negative weights, with heads in blue raising the evidence score ("citation heads") and heads in red lowering it ("counter-citation heads"). To examine whether these weights simply replicate our prior head-selection, we compute the Spearman correlation between head-selection scores $\bar{\rho}_i$ (Equation 4) and signed probe coefficients w_i. As shown in Table 2, the mean Spearman correlations range from 0.06 to 0.18, revealing that CITEPROBE does not simply reproduce the head-selection ranking.

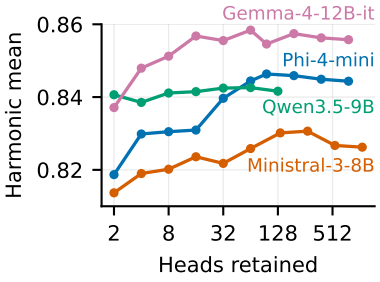


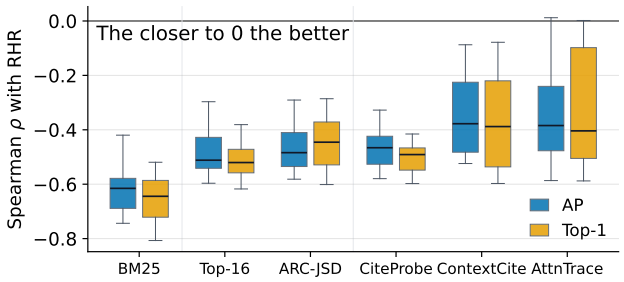
Figure 5: Head sensitivity analysis. Qwen only has 128 standard heads.

Table 2: Correlation analysis in CITEPROBE.

Model	HP	MSQ	TyDi	WiCE
Phi-4-mini	0.13	0.11	0.08	0.12
Ministral-3-8B	0.06	0.08	0.07	0.06
Qwen3.5-9B	0.13	0.18	0.13	0.14
Gemma-4-12B	0.18	0.17	0.15	0.15

Table 3: Harmonic mean of mAP and Top-1 of different readout architectures in CITEPROBE.

Readout	Phi	Ministral	Qwen	Gemma
MLP	0.843	0.826	0.851	0.852
CNN	**0.848**	**0.829**	**0.853**	0.839
Transf.	0.819	0.792	0.831	0.845
RetNet	0.812	0.794	0.840	0.844
Linear	0.843	0.828	0.843	**0.857**



Method	ρ_{AP}	ρ_{top-1}
BM25	−0.615	−0.644
Top-16 heads	−0.512	−0.520
ARC-JSD	−0.484	−0.446
CITEPROBE	−0.466	−0.491
AttnTrace	−0.384	−0.404
ContextCite-32	−0.378	−0.388

Figure 6: Spearman correlation between RHR and each method's outcome. Correlations closer to zero indicate less sensitivity to lexical competition. Table values are medians.

Effects of readout architectures. Table 3 shows that different readout architectures in CITEPROBE do not provide meaningful gains despite adding non-linearity and more trainable parameters. This indicates that the citation signal is largely linearly decodable, and the remaining error lies in other areas rather than in the architecture and capacity of the reader.

4.4 LEXICAL COMPETITION (RQ3)

BM25 is also lightweight and performs strongly, raising a further question of whether the probe merely exploits word overlap between answers and evidence. We investigate this by examining how citation performance changes when unannotated sentences also match the answer's wording. To quantify lexical competition, we define the Red Herring Ratio (RHR) using BM25 and use it to measure how well a method overcomes lexical distractions in citation. This is done by scoring the similarities between the generated answer and candidate sentences and ranking the distractors against the labels. Concretely, for each example e, let b_e^+ be the highest BM25 score among annotated supporting sentences and b_e^- the highest among unannotated candidates (distractors), to compute $\mathrm{RHR}_e = b_e^-/b_e^+$. A ratio near 1 means that the strongest distractor matches the answer about as well as the strongest annotated sentence. We use $\mathrm{RHR_e}$ to compute the Spearman correlations $\rho_{\mathrm{AP}} = \mathrm{Corr}_e(\mathrm{RHR}_e, \mathrm{AP}_e)$ and $\rho_{\mathrm{Top\text{-}1}} = \mathrm{Corr}_e(\mathrm{RHR}_e, \mathrm{correct}_e)$ across examples, where AP_e is the average precision for example e and $\mathrm{correct}_e$ indicates whether the top-ranked sentence is annotated. The closer ρ is to zero, the less sensitive the method is to lexical competition.

Figure 6 shows the correlations between the RHR and AP/Top-1 are predominantly negative. This is expected, as a distractor becomes more lexically similar to the answer, it becomes harder to differentiate. All methods get hampered by higher RHR, including perturbation methods. CITEPROBE exhibits resilience to lexical distractions, with weaker negative correlations than the Top-16-head baseline on both metrics, and ARC-JSD on ρ_{AP}, with smaller variance. ContextCite and AttnTrace yield correlations even closer to zero, but this should not be interpreted as robustness, as it may reflect their weaker overall performance and reduced sensitivity to both annotations and lexical competition. Their greater variance in ρ, as shown by the longer boxplots (Figure 6), also supports this.

4.5 FAILURE ANALYSIS (RQ4)

To understand the reasons for failed citations, we employed LLM-as-a-judge (Claude Opus 5) to classify each incorrect Top-1 selection into two broad categories: *Ranking Error* and *Valid but Unlabeled*. A *Ranking Error* is a genuinely incorrect selection, divided into four cases: (1) the selected citation contains the key entity or phrase but does not support the answer (*Entity Match, No Fact*); (2) it supports a different step of a multi-hop question (*Wrong Hop*); (3) it comes from the relevant passage but does not contain the supporting fact (*Adjacent, No Fact*); and (4) it is *Unrelated* (most severe). In contrast, a *Valid but Unlabeled* case is LLM-judged to be correct, but counted as incorrect by the benchmark due to imperfect labeling (limitations in Appendix G.2). We also further classify these cases into *Better*, *Equal*, or *Weaker* than the best annotated citation.

Table 4 shows that CITEPROBE's citation performance is substantially stronger than what the main results in Table 1 alone suggest, as more than half (54.8%) of its benchmark-failed Top-1 selections are actually LLM-judged to be valid. This is significantly higher than for all other baselines. In addition, almost half of these valid citations selected by CITEPROBE are LLM-judged to be better

Table 4: LLM-categorized Top-1 failures. We focus on performance in *Unrelated*, *Ranking Error* subtotal, *Valid but Unlabeled - Better*, and *Valid but Unlabeled* subtotals.

	Ranking error (%)					Valid but Unlabeled (%)			
Method	Entity	Hop	Adjac.	**Unrel.**	**Subtotal**	**Better**	Equal	Weaker	**Subtotal**
ContextCite-32	27.3	8.3	17.2	25.6	78.4	7.7	5.3	8.6	21.6
ARC-JSD	28.2	12.7	15.9	4.0	60.9	11.8	8.7	18.5	39.1
BM25	28.5	12.5	9.7	1.8	52.5	15.6	15.4	16.5	47.5
AttnTrace	28.5	5.9	24.6	26.8	85.7	4.6	4.5	5.2	14.3
Top-16 heads	23.6	11.9	13.8	4.2	53.5	15.7	13.0	17.8	46.5
CITEPROBE	21.9	13.3	8.4	**1.7**	**45.2**	**22.9**	13.3	18.5	**54.8**

than the best benchmark label. This means that not only is CITEPROBE performing substantially better than what is measured, but it is also being judged to have selected citations that are better than those provided by the benchmark itself. Furthermore, CITEPROBE is the least severe in terms of its *Ranking Error*, with only 1.7% of its errors in the *Unrelated* category.

4.6 EFFICIENCY COMPARISON (RQ5)

To examine the efficiency of CITEPROBE, we compare its training and inference cost with that of baselines on Gemma-4-12B-it using a single RTX 4090, with the results shown in Table 5. CITEPROBE incurs a one-time training cost of 242.3 minutes but only 12.1 minutes of inference overhead across the four datasets. Even including training, its total cost is 254.4 minutes, compared with 383.2 minutes for AttnTrace, 418.7 minutes for ContextCite-32, and 1,556.0 minutes for ARC-JSD. The efficiency advantage becomes more pronounced as the number of source sentences increases, since perturbation-based methods must repeatedly process modified contexts. In summary, CITEPROBE shifts the computational burden from repeated per-example inference to a one-time training stage, enabling efficient citation attribution at scale.

Table 5: Efficiency comparison by training and inference cost (min). "–" denotes training-free.

Method	**Train**	**Inference**				
		HQA	**MuS**	**TyDi**	**WiCE**	**Total**
AttnTrace	–	107.9	157.0	22.0	96.3	383.2
ContextCite-32	–	130.8	168.4	17.3	102.2	418.7
ARC-JSD	–	409.8	833.1	6.8	306.3	1556.0
CITEPROBE	242.3	**3.1**	**5.6**	**0.4**	**3.0**	**254.4**

5 CONCLUSION

In this paper, we find that the frozen attention of an LLM already contains citation-relevant signals and that learning to combine citation and counter-citation heads is key to extracting them for sentence-level citation. Motivated by this finding, we propose CITEPROBE, a lightweight method, which learns to directly decode citation signals from the attention weights of a frozen LLM. Experimental results show the effectiveness of CITEPROBE: **(1)** it outperforms existing methods across different LLMs and datasets, achieving strong citation performance under cross-dataset transfer; **(2)** it avoids repeated LLM inference by reusing the attention weights from the original generation pass, substantially reducing the computational cost; and **(3)** it can be extended to black-box LLMs through model surrogacy while maintaining competitive performance. Overall, our results demonstrate that sentence-level citation signals can be effectively and efficiently decoded from structured LLM attention, providing a lightweight solution for the sentence-level citation task.

6 LIMITATIONS

Despite its effectiveness, our CITEPROBE has several limitations that point to directions for future work. First, our study is limited to English, and whether citation and counter-citation heads exhibit similar patterns across languages remains unknown. Future work could evaluate the cross-lingual transferability of the learned probe. Second, the benchmarks contain annotation imperfections. Section 4.5 identifies unlabeled sentences that an LLM judge rates as stronger evidence than the best labeled citations, suggesting that revisiting these annotations could improve both training data and evaluation reliability. Third, CITEPROBE identifies sentences that support an answer according to annotated evidence, but does not distinguish between whether the model causally relied on them or recalled from its memory. Future work could extend CITEPROBE to incorporate causal attribution.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic datasets, evaluate benchmark results (LLM-as-a-judge), and to perform grammar and typo checks on our writing. We have not used generative AI tools to propose hypotheses, to design experiments, create theoretical models, mathematical claims or proofs, all of which are original contributions of the authors. Additionally, we used generative AI tools to assist in writing software code, and in finding related work. We have reviewed all AI-assisted work, such as performing sampling checks on LLM-as-a-judge results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The implementation of CITEPROBE, together with the code needed to reproduce our experiments on open-weight LLMs and public benchmarks, is available at https://www.github.com/citeprobe/citeprobe. The data splits, training configuration, and evaluation populations are documented in Section 4.1 and Appendices D and E.

REFERENCES

Oren Barkan, Edan Hauon, Avi Caciularu, Ori Katz, Itzik Malkiel, Omri Armstrong, and Noam Koenigstein. Grad-sam: Explaining transformers via gradient self-attention maps. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 2882–2887, 2021. doi: 10.1145/3459637.3482126.

Yee Man Choi, Xuehang Guo, Yi R. Fung, and Qingyun Wang. CiteGuard: Faithful citation attribution for LLMs via retrieval-augmented validation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6241–6257, 2026. doi: 10.18653/v1/2026.acl-long.282. URL https://aclanthology.org/2026.acl-long.282/.

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 10839–10858. PMLR, 2025. URL https://proceedings.mlr.press/v267/chuang25a.html.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, 2020. doi: 10.1162/tacl_a_00317. URL https://aclanthology.org/2020.tacl-1.30/.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. Contextcite: Attributing model generation to context. *Advances in Neural Information Processing Systems*, 37:95764–95807, 2024. doi: 10.52202/079017-3035.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 16344–16359, 2022. doi: 10.52202/068431-1189.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. arXiv preprint arXiv:2404.16130, 2024. URL https://arxiv.org/abs/2404.16130.

Ronald A. Fisher. On the "Probable Error" of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32, 1921.

Tz-Huan Hsu, Sian-Yao Huang, Che-Yu Lin, and Cheng-Lin Yang. Enhancing attention-based context attribution via token selection and think-twice mechanism. In Ricardo Campos, Adam Jatowt, Yanyan Lan, Mohammad Aliannejadi, Christine Bauer, Sean MacAvaney, Avishek Anand,

Zhaochun Ren, Suzan Verberne, Nan Bai, and Masoud Mansoury (eds.), *Advances in Information Retrieval*, pp. 490–498, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-21300-6.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. WiCE: Real-world entailment for claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7561–7583, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.470. URL https://aclanthology.org/2023.emnlp-main.470/.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. URL https://arxiv.org/abs/2001.08361.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of bert. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4365–4374, 2019. doi: 10.18653/v1/d19-1445. URL https://doi.org/10.18653/v1/d19-1445.

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. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Ruizhe Li, Chen Chen, Yuchen Hu, Yanjun Gao, Xi Wang, and Emine Yilmaz. Attributing response to context: A jensen-shannon divergence driven mechanistic study of context attribution in retrieval-augmented generation. In *International Conference on Learning Representations (ICLR)*, pp. 146805–146837, 2026. URL https://openreview.net/forum?id=7bHjHAOXH9.

Fengyuan Liu, Nikhil Kandpal, and Colin Raffel. AttriBoT: A bag of tricks for efficiently approximating leave-one-out context attribution. In *International Conference on Learning Representations (ICLR)*, pp. 16687–16710, 2025. URL https://openreview.net/forum?id=9kJperA2a4.

Microsoft. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025. URL https://arxiv.org/abs/2503.01743.

Mistral AI. Ministral 3 8B. Official model documentation, 2025. URL https://docs.mistral.ai/models/ministral-3-8b-25-12. Accessed September 6, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

Tobias Schreieder, Tim Schopf, and Michael Färber. Attribution, citation, and quotation: A survey of evidence-based text generation with large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30956–31000, 2026. doi: 10.18653/v1/2026.acl-long.1430. URL https://aclanthology.org/2026.acl-long.1430/.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023. URL https://arxiv.org/abs/2307.08621.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808, Florence, Italy, 2019. doi: 10.18653/v1/P19-1580. URL https://aclanthology.org/P19-1580/.

Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, Ruicheng Yin, Changze Lv, Xiaoqing Zheng, and Xuanjing Huang. Searching for best practices in retrieval-augmented generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17716–17736, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.981. URL https://aclanthology.org/2024.emnlp-main.981/.

Yanting Wang, Wei Zou, Runpeng Geng, and Jinyuan Jia. TracLLM: A generic framework for attributing long context LLMs. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3845–3864, 2025. ISBN 978-1-939133-52-6. URL https://www.usenix.org/conference/usenixsecurity25/presentation/wang-yanting.

Yanting Wang, Runpeng Geng, Ying Chen, and Jinyuan Jia. AttnTrace: Contextual attribution of prompt injection and knowledge corruption. In *2026 IEEE Symposium on Security and Privacy (SP)*, pp. 4320–4338. IEEE, 2026. doi: 10.1109/SP63933.2026.00198.

Jingxuan Wei, Xingyue Wang, Yanghaoyu Liao, Jie Dong, Yuchen Liu, Caijun Jia, Bihui Yu, and Junnan Zhu. GenProve: Learning to generate text with fine-grained provenance. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5027–5048, 2026. doi: 10.18653/v1/2026.acl-long.228. URL https://aclanthology.org/2026.acl-long.228/.

Sirui Xia, Xintao Wang, Jiaqing Liang, Yifei Zhang, Weikang Zhou, Jiaji Deng, Fei Yu, and Yanghua Xiao. Ground every sentence: Improving retrieval-augmented LLMs with interleaved reference-claim generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 969–988, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-naacl.55. URL https://aclanthology.org/2025.findings-naacl.55/.

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*, pp. 2369–2380, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Yirui Zhan, Chenhao Xu, and Jun Gao. Guidance: Sentence-level citation enforcement via prefix-tail guidance during LLM decoding. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=cEA9yswzZf.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. LongCite: Enabling LLMs to generate fine-grained citations in long-context QA. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5098–5122, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.264. URL https://aclanthology.org/2025.findings-acl.264/.

CONTENTS

1 Introduction **1**

2 Related Work **2**

3 Methodology **3**

3.1 From Head-Specific Attention Weights to Citation Scores 3

3.2 Head Selection Exploration 3

3.3 CiteProbe: A Lightweight Readout for Sentence-level Citation 4

4 Experiments **5**

4.1 Experimental Setup 5

4.2 Main Results (RQ1) 6

4.3 In-depth Analysis (RQ2) 7

4.4 Lexical competition (RQ3) 8

4.5 Failure Analysis (RQ4) 8

4.6 Efficiency Comparison (RQ5) 9

5 Conclusion **9**

6 Limitations **9**

A Additional Head-Selection Analyses **15**

A.1 Performance Peaks with Few Heads 15

A.2 Stability and Location of Selected Heads 16

A.3 Pooling the Per-Example Correlations 16

B Sentence-Length Normalization Ablation **17**

C Fused Kernel Proof for CiteProbe Readout **18**

C.1 Online Softmax with Sentence Sums 18

C.2 Exact Recovery of the Sentence Grid 18

D Readout Training and Hyperparameter Search **19**

D.1 Training Configuration 19

D.2 Head Weight Distribution and Pruning 19

E Evaluation Benchmarks and Population **20**

F Readout Architecture Ablation **21**

F.1 Hyperparameter Search and Architecture 21

F.2 Training and selection 22

F.3 Results 23

G Analysis of Citation Failures **23**

G.1 Failure Taxonomy and Judging Protocol 23

G.2 Imperfect Benchmark Labelling 25

G.3 Detailed Breakdown of Valid Citations Not Labelled 25

H Lexical Competition Across Models and Benchmarks **27**

I Computational Efficiency Analysis **28**

J Model Surrogacy: Citing a Closed Model's Answer **28**

J.1 Setup 28

J.2 Results 29

K Comparison with Guidance – A Full End-To-End RAG and Citation System **30**

L Citation Behavior on Correct and Incorrect Answers **31**

L.1 Setup 32

L.2 Correctly Answered Examples 32

L.3 Incorrectly Answered Examples 32

A Additional Head-Selection Analyses

A.1 Performance Peaks with Few Heads

The main-text analysis shows that a small subset of heads outperforms all-head averaging. The full sweep in Table 6 shows that performance peaks early. The mean harmonic score peaks at 0.8340 for $k = 4$, closely followed by 0.8336 at $k = 8$, and remains within 0.006 of the peak from $k = 1$ to $k = 32$. Adding more heads eventually reduces performance, most sharply for Gemma-4-12B-it. This suggests that selecting informative heads matters more than the exact size of a small retained set, while including too many lower-ranked heads dilutes the useful signal. Figure 7 reports mAP and Top-1 separately.

Table 6: Harmonic mean of dataset-averaged test mAP and Top-1 across head counts and ranking targets, on correctly answered examples. Bold marks $k = 4$ and its evidence-label mean, the highest mean score. Underlining marks $k = 8$ as the second-best. Qwen3.5-9B exposes only 128 heads, so larger settings and their four-model means are left blank.

Model	k=1	k=2	k=4	k=8	k=16	k=32	k=64	k=128	k=256	k=512	all
Phi-4-mini-instruct											
Evidence label	.8231	.8308	.8343	.8359	.8338	.8315	.8283	.8244	.7963	.6519	.4634
ContextCite-256	.8156	.8204	.8228	.8267	.8286	.8286	.8277	.8252	.7958	.6317	.4634
Ministral-3-8B-Instruct											
Evidence label	.8167	.8120	.8070	.8070	.8070	.8050	.8024	.7954	.7841	.7509	.5897
ContextCite-256	.7956	.7996	.8006	.8044	.8043	.8018	.8004	.7931	.7850	.7475	.5897
Qwen3.5-9B											
Evidence label	.8206	.8328	.8452	.8400	.8392	.8344	.8271	.7827	–	–	.7827
ContextCite-256	.7669	.8064	.8070	.8062	.8098	.8180	.8160	.7827	–	–	.7827
Gemma-4-12B-it											
Evidence label	.8506	.8507	.8494	.8515	.8484	.8447	.8294	.7955	.6032	.4156	.3498
ContextCite-256	.8224	.8437	.8414	.8427	.8434	.8399	.8293	.7671	.5699	.4025	.3498
Mean											
Evidence label	.8278	.8316	.8340	.8336	.8321	.8289	.8218	.7995	–	–	.5464
ContextCite-256	.8001	.8175	.8180	.8200	.8215	.8221	.8184	.7920	–	–	.5464

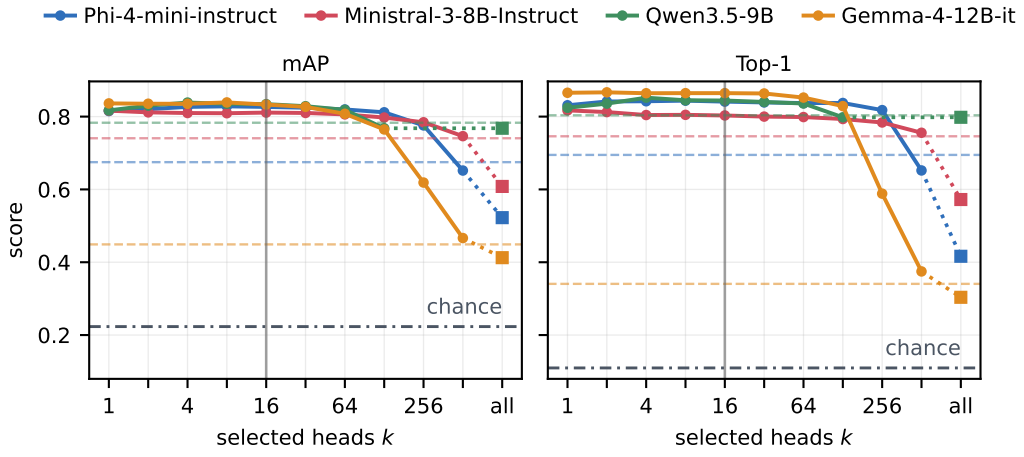


Figure 7: mAP (left) and Top-1 (right) against the number of selected heads, averaged over the four datasets and restricted to correctly answered examples. Dashed lines mark raw all-head averaging, dotted tails run to the normalized all-head result, and the vertical rule marks $k = 16$. Chance is 0.223 for mAP and 0.110 for Top-1.

We also test label-free head selection using ContextCite-256 ablation weights (Cohen-Wang et al., 2024) as the ranking target. We use 256 rather than 32 ablations to give ContextCite a larger sampling budget, aiming for more accurate ranking targets. To limit computation, we use 50 calibration examples per dataset, which still require $50 \times 256 = 12{,}800$ ablation evaluations. Evidence-label ranking uses 500 training examples but requires no such ablations. The two ranking targets show similar performance trends across the sweep, with four-model mean harmonic scores differing by less than 0.03 from $k = 1$ to $k = 128$ (Table 6), supporting our use of evidence-label ranking. This keeps CITEPROBE independent of ContextCite, given that ContextCite needs to be used as a comparison baseline.

A.2 STABILITY AND LOCATION OF SELECTED HEADS

For each pair of leave-one-out splits a and b, overlap is the fraction of selected heads shared by both sets, $|\mathcal{S}_a^{(k)} \cap \mathcal{S}_b^{(k)}|/k$, ignoring head order. Figure 8 shows the computed overlap across k. At $k = 4$ the sets are small, so a single disagreeing head moves a pair by 25 points, and the MuSiQue/WiCE pair shares only half of its heads on every LLM. At $k = 16$ every pair on every LLM shares at least 69% of its heads, and at $k = 32$ at least 78%. Retaining more heads therefore makes the selected set less sensitive to which dataset is held out, with relatively minimal change in performance. Since the training datasets overlap across splits, this measures selection stability rather than an independent test of transfer.

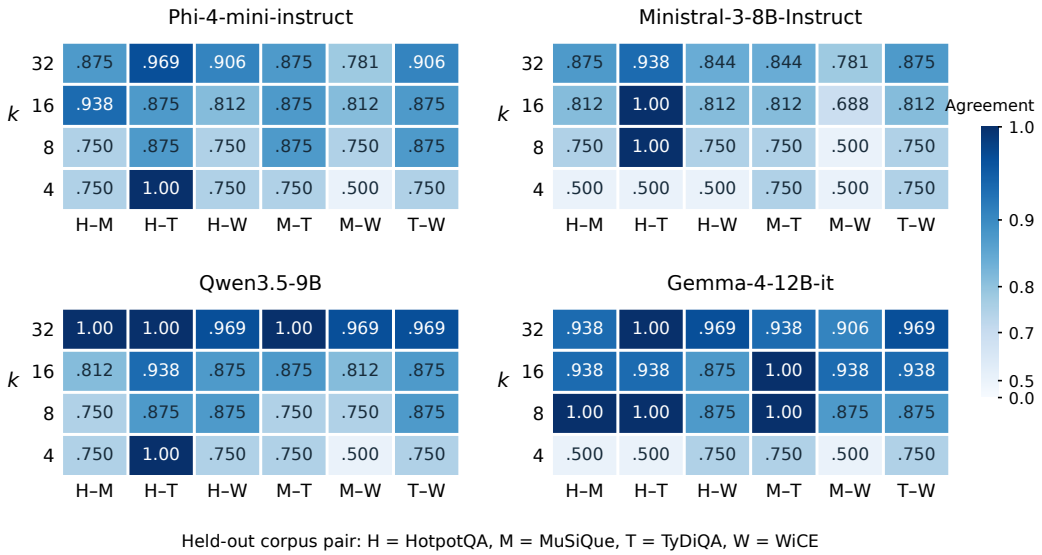


Figure 8: Agreement between head selections across leave-one-out splits. Each split ranks heads using its three training datasets. Agreement is the fraction of top-k heads shared by the two splits holding out the indicated datasets. Each panel shows one model, with rows indicating k and columns indicating pairs of held-out datasets. Cell values report agreement; darker shading indicates greater overlap.

Architecture also appears to constrain where useful heads occur (Figure 9). Gemma-4-12B-it has full attention in eight of its 48 layers, yet all sixteen selected heads lie in those layers in every split. This concentration suggests that access to the full context helps locate supporting evidence.

A.3 POOLING THE PER-EXAMPLE CORRELATIONS

To obtain one ranking score per head in Equation 4, we pool correlations after applying Fisher's z transform. We use weights $a_e = \max(n_e - 3, 0)$, where n_e is the number of candidate sentences, so examples with more candidates contribute more to the ranking:

$$\text{Fisher-z}_e(\rho_{i,e}) = \tanh\left(\frac{\sum_e a_e \operatorname{arctanh} \rho_{i,e}}{\sum_e a_e}\right). \tag{7}$$

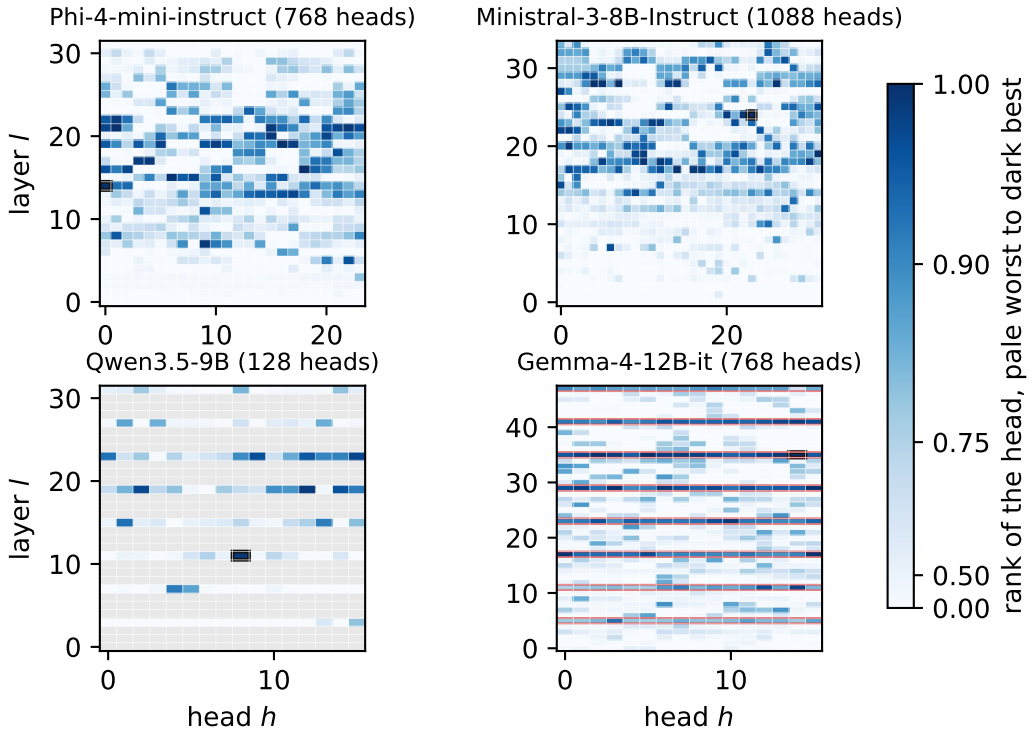


Figure 9: Head ranks over the full layer-head grid, averaged over the four leave-one-dataset-out selections and normalized by each model's number of attention heads that expose weights. Darker cells indicate higher ranks, and the outlined cell is the highest-ranked head. The color scale is compressed at the pale end. Gray cells carry no attention weights; for Qwen3.5 these are the 24 linear-attention layers among its 32 layers. Red outlines in Gemma-4-12B-it mark its eight full-attention layers.

The weights are motivated by the approximate variance $1/(n_e - 3)$ of Fisher-transformed correlations under bivariate normal sampling. For our binary evidence labels and sentences sharing a context, this is a weighting heuristic rather than an exact variance model. Examples with at most three candidates receive zero weight. Applying $\tanh$ returns the pooled score to the correlation scale.

B Sentence-Length Normalization Ablation

A potential concern with Equation 1 is that candidate sentences have different lengths, while the score is obtained by summing attention over their tokens. We conduct this ablation to determine whether normalizing attention scores by sentence length improves performance. We begin by generalizing Equation 1 as follows:

$$G_{i,j}^{(\beta)} = \frac{1}{|\mathcal{T}_j|^{\beta}} \sum_{t \in \mathcal{T}_y} \sum_{u \in \mathcal{T}_j} \alpha_{t,u}^{(i)}, \tag{8}$$

where setting $\beta = 0$ recovers Equation 1 exactly, while setting $\beta = 1$ gives mean attention per source token. We test by learning β with the same fitting procedure, learning-rate grid, and random seeds.

Table 7 shows that full sentence-length normalization ($\beta = 1$) consistently degrades performance, reducing mAP by 0.030 and Top-1 by 0.037 on average relative to $\beta = 0$. Learning β leads to partial normalization, but its average performance is effectively unchanged from $\beta = 0$, with mAP lower by 0.0006 and Top-1 differing by less than 0.0001. Thus, full normalization removes useful signal,

Table 7: Sentence-length normalization ablation. mAP and Top-1 are means over the four held-out datasets. Learned β is reported as mean $\pm$ standard deviation across fits.

Model	Normalization	β	mAP	Top-1
Phi-4-mini-instruct	None	0	.832	.856
	Learned	0.416 ± 0.054	.825	.862
	Full	1	.807	.823
Ministral-3-8B-Instruct	None	0	.826	.834
	Learned	0.433 ± 0.055	.831	.841
	Full	1	.786	.788
Qwen3.5-9B	None	0	.836	.854
	Learned	0.421 ± 0.013	.835	.854
	Full	1	.801	.813
Gemma-4-12B-it	None	0	.842	.871
	Learned	0.342 ± 0.087	.841	.867
	Full	1	.801	.829

while partial normalization offers no measurable benefit. We therefore retain the simpler $\beta = 0$ formulation.

C Fused Kernel Proof for CiteProbe Readout

FlashAttention uses tiled computation and online-softmax rescaling to avoid materializing the full attention matrix, reducing memory traffic and peak memory usage (Dao et al., 2022). In this paper, we use eager attention to materialize the weights for CiteProbe readouts. However, eager mode is not necessary: a fused kernel can accumulate the sentence sums during the tiled forward pass using the same rescaling as the attention output. We provided a HuggingFace implementation as a proof-of-concept for the fused kernel described below.

C.1 Online Softmax with Sentence Sums

For a fixed head i and query t, let z_u be the attention weight for key u, with masked weights set to $-\infty$, so that $\alpha_{t,u}^{(i)} = e^{z_u} / \sum_v e^{z_v}$. Online softmax maintains a running maximum m and normalization sum ℓ. For each new key tile T, it updates

$$m' = \max\left(m, \max_{u \in T} z_u\right), \qquad c = e^{m - m'}, \qquad \ell' = c\ell + \sum_{u \in T} e^{z_u - m'}. \tag{9}$$

The factor c puts previous contributions on the new scale, since $e^{z_u - m} c = e^{z_u - m'}$. We apply the same update to each sentence sum:

$$M_j' = cM_j + \sum_{u \in T \cap \mathcal{T}_j} e^{z_u - m'}. \tag{10}$$

Initially, $m = -\infty$ and $\ell = M_j = 0$; when $m' = -\infty$, we set $c = 1$ and all tile contributions to zero.

Algorithm 1 applies these updates to one head and one query block alongside the attention output. Let $g_u = j$ for keys in sentence s_j, and $g_u = 0$ for all other keys. All unmasked keys contribute to the softmax denominator; only unmasked candidate keys contribute to M.

C.2 Exact Recovery of the Sentence Grid

After processing keys U with at least one unmasked key, the accumulators satisfy

$$m = \max_{u \in U} z_u, \qquad \ell = \sum_{u \in U} e^{z_u - m}, \qquad M_j = \sum_{u \in U \cap \mathcal{T}_j} e^{z_u - m}. \tag{11}$$

Algorithm 1 Fused CITEPROBE sentence sums.

Require: query block Q, keys K, values V, sentence ids g_u, tile width B

$m \leftarrow -\infty, \quad \ell \leftarrow 0, \quad O \leftarrow 0, \quad M \leftarrow 0$ ▷ M: queries × sentences
for each key tile T of at most B keys **do**
  $S \leftarrow$ weights z_u for $u \in T$ (scaled $QK_T^\top$), with masked entries set to $-\infty$
  $m' \leftarrow \max(m, \operatorname{rowmax} S)$
  $c \leftarrow \exp(m - m'), \quad P \leftarrow \exp(S - m')$ rowwise; if $m' = -\infty$, set $c = 1$, $P = 0$
  $\ell \leftarrow \ell c + \operatorname{rowsum} P$
  $O \leftarrow Oc + PV_T$
  $M \leftarrow Mc$ ▷ rescale sentence sums
  $M[:, g_u] \mathrel{+}= P[:, u]$ for $u \in T$ with $g_u > 0$ ▷ accumulate sentence sums
  $m \leftarrow m'$
end for
return $O/\ell, \quad M/\ell$ rowwise division

These identities hold for the first tile containing an unmasked key and are preserved by the rescaling and addition above. After the final tile, assuming at least one key is unmasked,

$$\frac{M_j}{\ell} = \frac{\sum_{u \in \mathcal{T}_j} e^{z_u - m}}{\sum_u e^{z_u - m}} = \sum_{u \in \mathcal{T}_j} \alpha_{t,u}^{(i)}. \tag{12}$$

Averaging over answer queries therefore recovers $G_{i,j}$ in Equation 1, exactly in real arithmetic, without materializing the attention matrix. A HuggingFace implementation of Algorithm 1 reproduces the eager grids on 100 test captures per LLM up to fp16 rounding.

D READOUT TRAINING AND HYPERPARAMETER SEARCH

This section details the training configuration for the CITEPROBE.

D.1 TRAINING CONFIGURATION

CITEPROBE's training configuration is listed in Table 8. The learning rate is selected once per LLM based on the harmonic mean of validation mAP and Top-1 pooled over all four holdouts and three seeds, which avoids target-specific hyperparameter selection under the leave-one-dataset-out protocol.

Table 8: Readout and training configuration of CITEPROBE. Parameter counts are one weight per head plus a global bias; seed spread is the largest standard deviation of test mAP over the three seeds in any model-dataset cell.

Setting	Value
Optimizer	AdamW, weight decay 0.01
Batch size	128 sentence-label pairs
Epochs	Early stopping after 30 epochs without improvement
Checkpoint	Epoch with the highest harmonic mean of validation mAP and Top-1
Learning-rate grid	$\{10^{-5}, 3\times10^{-5}, 10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}\}$
Selected rate	10^{-4} Phi-4-mini-instruct, Gemma-4-12B-it; 3×10^{-4} Ministral-3-8B-Instruct; 10^{-2} Qwen3.5-9B
Seeds	3, results averaged
Seed spread	≤ 0.008 mAP

D.2 HEAD WEIGHT DISTRIBUTION AND PRUNING

Section 4.3 shows that CITEPROBE combines attention heads differently from the Top-16-head baseline. Here, we examine whether fitting the full attention profile nevertheless produces a sparse readout.

The fitted weights are distributed across heads. For each fit, we rank the available heads by absolute probe coefficient and measure their cumulative coefficient mass. Figure 10 shows that the sixteen largest coefficients account for only 6.4 to 33.3% of the total absolute coefficient mass. Reaching 50% requires between 29 and 240 heads on average, depending on the LLM. The probe therefore does not collapse to a sparse approximation of the Top-16-head baseline. These coefficient shares describe the fitted readout rather than the predictive importance of individual heads, since correlated heads may divide their contributions across several coefficients.

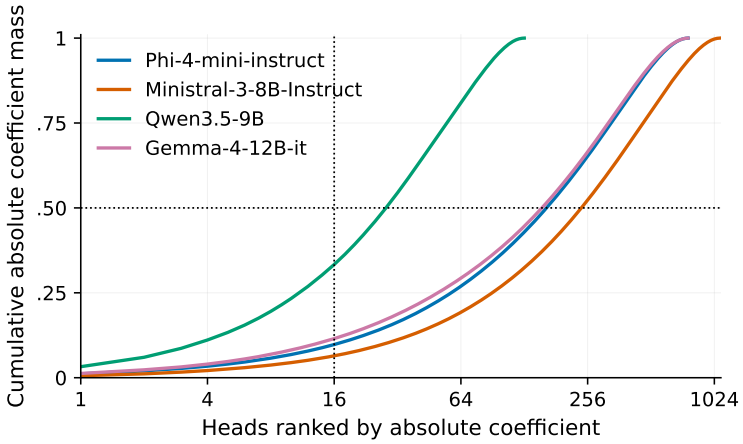


Figure 10: Cumulative absolute coefficient mass after ranking heads by coefficient magnitude, averaged over four held-out datasets and three seeds. Dotted lines mark sixteen heads and 50% of the coefficient mass. Coefficient mass describes the fitted weights rather than the predictive importance of individual heads.

Explicit pruning provides no consistent benefit. We test whether recursive feature elimination improves the probe by repeatedly removing heads with the smallest absolute coefficients and refitting. Table 9 separates the averaged curve in Figure 5 by LLM and held-out benchmark. Although intermediate widths fluctuate, the full probe outperforms the two-head probe in thirteen of the sixteen pairs, with a mean harmonic-score gain of 0.014. Selecting the retained width using validation performance improves mean test mAP by less than 0.001 and Top-1 by 0.001, which is comparable to the variation across training seeds. Explicit pruning therefore adds a separate selection and refitting stage without providing a consistent performance gain. We consequently retain the complete attention profile in CITEPROBE.

E EVALUATION BENCHMARKS AND POPULATION

HotpotQA. Multi-hop question answering over Wikipedia. Each question needs two paragraphs. Annotators marked the supporting sentences, and we use those as the evidence labels.

MuSiQue. Multi-hop question answering with two to four hops and twenty paragraphs per question. Only the supporting paragraphs are annotated, not the sentences. A sentence is labeled as evidence if it lies in a supporting paragraph and contains that hop's answer.

TyDiQA-GoldP. Single-passage question answering. We use the English subset. A sentence is labeled as evidence if it contains the annotated answer span.

WiCE. Claim verification. Each claim comes from Wikipedia and is checked against the web page it cites. Annotators marked the supporting sentences, and we use those as the evidence labels. WiCE has claims but no questions, so we used Claude Opus 5 to write one question for each claim.

Population. Before evaluating citations, we generate an answer with each LLM and retain only correctly answered examples. Because the benchmark evidence annotations support the correct

Table 9: Harmonic mean of mAP and Top-1 under recursive feature elimination. Full denotes the complete attention profile, and Δ is the difference between the full and two-head probes.

		Heads retained							
Model	Benchmark	2	4	8	16	32	64	Full	Δ
Phi-4-mini	HotpotQA	.807	.832	.846	.844	.844	.855	.849	+.042
	MuSiQue	.815	.812	.799	.807	.815	.828	.830	+.015
	TyDiQA	.869	.878	.875	.872	.881	.877	.882	+.013
	WiCE	.781	.795	.799	.798	.816	.815	.814	+.033
Ministral-3-8B	HotpotQA	.842	.845	.843	.852	.858	.846	.846	+.004
	MuSiQue	.821	.823	.827	.827	.818	.834	.836	+.015
	TyDiQA	.829	.839	.826	.825	.819	.832	.828	.000
	WiCE	.761	.767	.781	.788	.788	.789	.791	+.030
Qwen3.5-9B	HotpotQA	.852	.845	.857	.855	.856	.859	.851	-.001
	MuSiQue	.822	.834	.838	.833	.841	.848	.842	+.020
	TyDiQA	.872	.856	.849	.862	.852	.843	.854	-.018
	WiCE	.814	.817	.818	.814	.818	.817	.816	+.002
Gemma-4-12B	HotpotQA	.822	.865	.878	.877	.875	.878	.869	+.046
	MuSiQue	.847	.833	.834	.847	.853	.851	.853	+.006
	TyDiQA	.893	.895	.896	.901	.896	.903	.902	+.009
	WiCE	.783	.795	.795	.799	.796	.800	.797	+.014

answer, including incorrect generations would confound citation ranking with answer correctness. Table 10 reports the number of retained examples over the number with captured model outputs. Overall, the models retain between 63.1% and 84.2% of captured examples, leaving 2,123 to 2,833 examples per model for citation evaluation. The reported ranking metrics are therefore conditional on the underlying LLM producing a correct answer.

Table 10: Number of examples models answered correctly, retained for citation evaluation. TyDiQA contains 366 usable examples after filtering short contexts, while other denominator differences arise from prompts exceeding the configured prompt-token limit.

Model	HotpotQA		MuSiQue		TyDiQA		WiCE		Overall
	Count	%	Count	%	Count	%	Count	%	%
Phi-4-mini-instruct	743/1000	74.3	381/1000	38.1	263/366	71.9	736/1000	73.6	63.1
Ministral-3-8B-Instruct	795/1000	79.5	612/970	63.1	243/366	66.4	794/1000	79.4	73.3
Qwen3.5-9B	910/1000	91.0	801/998	80.3	311/366	85.0	811/1000	81.1	84.2
Gemma-4-12B-it	861/998	86.3	699/989	70.7	288/366	78.7	799/1000	79.9	78.9

F Readout Architecture Ablation

We ablate the citation readout architecture by replacing the linear probe with an MLP, CNN, Transformer, or RetNet. For each architecture, we follow CiteProbe's procedure for constructing attention profiles, training a model-specific readout, and evaluating transfer to a held-out dataset.

F.1 Hyperparameter Search and Architecture

Hyperparameter search. For MLP and CNN, we grid-search seven learning rates $\{10^{-5}, 3\times10^{-5}, 10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}\}$ and three batch sizes $\{128, 256, 512\}$. For the more computationally expensive Transformer and RetNet readouts, we use Bayesian optimization with Optuna's TPE sampler, searching the same learning-rate and batch-size choices. All reported configurations are trained with three random seeds, and results are averaged across seeds. Bayesian search trials use one seed before the selected configuration is refitted with three.

MLP. The MLP uses fully connected hidden layers with ReLU activations, followed by a scalar linear output. We conduct a grid search over hyperparameters: layers $\in \{1, 2\}$ and widths $\in \{1, 2, 4, 8, 16, 32, 64, 128\}$.

CNN. The CNN uses two 3×3 convolutional layers with ReLU activations, followed by a scalar linear output. Following standard CNN practice, it takes the $L \times H$ grid directly, omitting the vec operation in Equation 5 to preserve layer and head positions. We conduct a grid search over hyperparameters: widths $\in \{1, 2, 4, 8, 16, 32, 64, 128\}$.

Transformer. The Transformer uses a standard decoder-only architecture with self-attention, positional embeddings, and ReLU activations. We use a standard feed-forward expansion factor of four (Kaplan et al., 2020). We conduct a Bayesian search over hyperparameters: blocks $\in \{1, 2, 3\}$, heads $\in \{2, 4\}$, and widths $\in \{16, 32, 64, 128\}$. We reduced the width search to reduce compute.

RetNet. The RetNet-style readout uses bidirectional multi-scale retention (Sun et al., 2023), learned positional embeddings, a SiLU gate, and GELU feed-forward layers. We use a feed-forward expansion factor of four and omit xPos rotation. We conduct a Bayesian search over hyperparameters: blocks $\in \{1, 2, 3\}$, heads $\in \{2, 4\}$, and widths $\in \{16, 32, 64, 128\}$.

F.2 Training and selection

All readouts use class-weighted binary cross-entropy and AdamW with weight decay 0.01. Training stops after 30 steps without harmonic mean (mAP and Top-1) improvement. Consistent with the selection procedure for the linear probe, after evaluation, the hyperparameter configuration that yields the best harmonic mean across all four datasets is selected. Table 11 lists the selected recipe and kept epoch per holdout.

Table 11: Selected recipe of every readout and LLM in Table 12, with the kept epoch for each held-out dataset in the order HotpotQA, MuSiQue, TyDiQA, WiCE. Each holdout is fit with three seeds, and the epoch shown is the median of the three.

Readout	Hyperparameters	Params	LR	Batch	Kept epoch
Phi-4-mini-instruct					
Linear (CiteProbe)		769	1e-4	128	25, 46, 42, 19
MLP	1 layer, width 8	6161	1e-3	512	7, 12, 7, 9
CNN	channels [8, 16]	13.5k	1e-4	128	9, 20, 10, 11
Transformer	width 32, 2 blocks, 4 heads	27.3k	3e-3	128	48, 10, 19, 21
RetNet	width 16, 3 blocks, 2 heads	11.4k	1e-2	128	9, 38, 7, 22
Ministral-3-8B-Instruct					
Linear (CiteProbe)		1089	3e-4	128	34, 25, 45, 30
MLP	1 layer, width 8	8721	3e-4	256	25, 18, 20, 37
CNN	channels [8, 16]	18.7k	1e-4	128	20, 22, 27, 30
Transformer	width 64, 2 blocks, 4 heads	104.3k	3e-3	512	39, 44, 40, 33
RetNet	width 32, 2 blocks, 4 heads	29.5k	3e-3	512	10, 18, 40, 14
Qwen3.5-9B					
Linear (CiteProbe)		513	1e-2	128	23, 2, 36, 5
MLP	1 layer, width 8	4113	1e-3	512	15, 8, 16, 43
CNN	channels [8, 16]	9441	3e-4	128	24, 4, 8, 13
Transformer	width 64, 1 block, 4 heads	53.2k	1e-3	256	24, 13, 17, 17
RetNet	width 64, 2 blocks, 4 heads	111.1k	1e-2	128	22, 17, 36, 28
Gemma-4-12B-it					
Linear (CiteProbe)		769	1e-4	128	67, 52, 33, 65
MLP	1 layer, width 8	6161	1e-3	256	28, 31, 8, 9
CNN	channels [8, 16]	13.5k	3e-4	128	6, 6, 11, 9
Transformer	width 128, 1 block, 4 heads	206.7k	3e-3	256	47, 39, 28, 4
RetNet	width 64, 2 blocks, 4 heads	112.1k	1e-2	128	63, 27, 43, 33

F.3 RESULTS

Table 12 reports mAP and Top-1. Despite its simplicity, the linear probe remains competitive against more complex readouts and sometimes outperforms them. More complex readouts learn nonlinear functions, but do not improve held-out citation performance. This suggests that the citation signal available in the attention profiles can already be extracted effectively with a linear readout. The remaining errors may instead reflect limitations in the attention representation or training data quality, rather than insufficient readout capacity.

Table 12: Readout architecture ablation. Despite its simplicity, the linear probe remains competitive with more complex readouts.

Readout	HotpotQA		MuSiQue		TyDiQA		WiCE		Mean	
	mAP	Top-1	mAP	Top-1	mAP	Top-1	mAP	Top-1	mAP	Top-1
Phi-4-mini-instruct										
MLP	.838	.864	.794	.893	.906	.852	.782	.818	.830	.857
CNN	**.843**	.864	**.802**	**.898**	.905	**.853**	.791	.827	**.835**	**.861**
Transformer	.817	.838	.763	.852	.886	.826	.773	.793	.810	.828
RetNet	.822	.826	.757	.847	.888	.821	.763	.773	.807	.817
Linear (CITEPROBE)	.834	**.865**	.789	.874	**.908**	.852	**.793**	**.835**	.831	.856
Ministral-3-8B-Instruct										
MLP	.829	.852	.812	.856	**.884**	**.809**	.769	.802	.823	.830
CNN	.826	.837	**.823**	**.867**	.877	.801	**.776**	**.825**	**.826**	.832
Transformer	.790	.788	.774	.847	.857	.775	.740	.771	.790	.795
RetNet	.808	.812	.783	.846	.856	.771	.732	.749	.795	.794
Linear (CITEPROBE)	**.831**	**.869**	.811	.863	.874	.793	.773	.817	.822	**.835**
Qwen3.5-9B										
MLP	.842	.879	.835	.867	**.901**	**.839**	.795	.850	.844	.859
CNN	**.844**	**.889**	**.840**	**.878**	.895	.827	**.802**	**.852**	**.845**	**.862**
Transformer	.818	.854	.822	.848	.895	.831	.775	.813	.827	.836
RetNet	.836	.873	.826	.860	.898	.832	.785	.811	.836	.844
Linear (CITEPROBE)	.817	.869	.823	.876	.894	.823	.793	.851	.832	.855
Gemma-4-12B-it										
MLP	.832	.872	.822	**.912**	.916	.873	.775	.811	.837	.867
CNN	**.851**	.877	**.834**	.910	.858	.773	**.790**	**.819**	.833	.845
Transformer	.843	.865	.826	.888	.907	.877	.769	.791	.836	.855
RetNet	.832	.856	.826	.890	.917	.878	.769	.783	.836	.852
Linear (CITEPROBE)	.850	**.896**	.815	.897	**.918**	**.884**	.784	.814	**.842**	**.873**

G ANALYSIS OF CITATION FAILURES

A citation scored wrong is either a ranking error or a labelling gap, and this appendix separates the two. Appendix G.1 defines the failure taxonomy and the judging protocol that assigns it. Appendix G.2 documents where the benchmarks' own evidence annotations are incomplete or inferred.

G.1 FAILURE TAXONOMY AND JUDGING PROTOCOL

This appendix describes the failure taxonomy used in Section 4.5 and the procedure used to assign failure labels.

Failure-analysis population. We define a failure as an evaluated example for which the highest-ranked sentence is not annotated as supporting evidence. Since CITEPROBE is evaluated over three random seeds, an example is considered a failure when an unannotated sentence is ranked first by a majority of seeds. For each failure, we record the highest-ranked annotated sentence, its rank, and the unannotated candidates ranked above it.

Labels. Each failure is assigned exactly one label. Four labels correspond to *Ranking Error*, where the selected sentence does not support the generated answer, while three correspond to *Not Labelled*, where the selected sentence provides valid support but is absent from the benchmark annotations.

Each *Ranking Error* definition carries one representative example. QUERY is the question presented to the model, PICKED the sentence ranked first, LABELED the annotated supporting sentence it displaced, and WHY the reason the pick fails. Sentences are shortened for space.

- **Entity match, no fact**: the selected sentence contains a salient entity, number, or phrase from the question or answer but does not state the fact needed to support the answer. This includes sentences that merely introduce the relevant entity or mention a different entity with the same name.

  Example

  QUERY Matt Doheny lost his primary race to what current youngest woman ever elected to Congress?
  PICKED Elise Marie Stefanik (born July 2, 1984) is an American politician from the state of New York.
  LABELED Doheny lost the primary to Elise Stefanik, who subsequently went on to win the general election.
  WHY Names Stefanik but does not establish that she defeated Doheny.

- **Wrong hop**: for a multi-hop question, the selected sentence supports a different reasoning step from the one needed to justify the generated answer. The selected fact may be relevant to the overall reasoning chain but does not itself support the answer being cited.

  Example

  QUERY When did the rapper on *On and On and Beyond* release *Best Day Ever*?
  PICKED Best Day Ever is the fifth mixtape by American rapper Mac Miller.
  LABELED This mixtape was released online March 11, 2011.
  WHY Connects the mixtape to the rapper but omits the release date needed to answer the question.

- **Adjacent, no fact**: the selected sentence comes from the relevant passage or topic but does not state the fact needed to support the generated answer.

  Example

  QUERY Who is the current leader of South Africa?
  PICKED The President is elected by the National Assembly, the lower house of Parliament, and is usually the leader of the largest party.
  LABELED The incumbent is Cyril Ramaphosa, who was elected by the National Assembly on 15 February 2018.
  WHY Describes how the president is elected but does not identify the officeholder.

- **Unrelated**: the selected sentence has no substantive connection to the question or generated answer.

  Example

  QUERY What administrative territorial entity includes the place that KTRP is licensed to broadcast to?
  PICKED Minsk Region is one of the regions of Belarus.
  LABELED Notus is a small rural city in Canyon County, Idaho, United States.
  WHY Refers to a region in Belarus with no connection to KTRP's licensed location.

- **Not Labelled—better / equal / weaker**: the selected sentence does support the generated answer but is absent from the benchmark evidence annotations. Relative to the highest-ranked annotated sentence, the judge determines whether the selected sentence provides better, approximately equivalent, or weaker support.

Judging. We judge and categorize failures using Claude Opus 5. The prompt used for judging is as follows:

A citation method had to pick the context sentence that best supports the MODEL ANSWER. It picked PICKED SENTENCE. Human annotators had marked MARKED SENTENCE. Classify why the pick differs, using exactly one of these labels.

lexical_overlap: the picked sentence contains a key name, number or phrase from the question or the answer, but does not state the fact the answer asserts. This covers a sentence that merely introduces the entity, a fragment carrying only the name, and a sentence about a different entity that shares the name.

adjacent_no_fact: the picked sentence is on the same topic or from the same passage, does not contain the key term, and does not support the answer.

unannotated_better: the picked sentence supports the answer and is a better citation for it than the marked sentence.

unannotated_equal: the picked sentence supports the answer about as well as the marked sentence, for instance because both state the same fact.

unannotated_poorer: the picked sentence supports the answer but the marked sentence carries the specific number, name or date the answer states.

wrong_hop: the question requires two or more linked facts, and the picked sentence is evidence for a different step than the marked sentence.

unrelated: the picked sentence has no bearing on the question or the answer.

Reply with one line per example, formatted as the ### identifier, a space, then the label. No other text.

⟨identifier⟩
QUESTION: ⟨question⟩
MODEL ANSWER: ⟨generated answer, truncated to 700 characters⟩
PICKED SENTENCE: ⟨sentence ranked first⟩
MARKED SENTENCE: ⟨highest-ranked annotated sentence⟩

G.2 Imperfect Benchmark Labelling

Imperfect evidence labelling is not uncommon in question-answering and claim-verification benchmarks, particularly when the annotations are repurposed for a more fine-grained task such as sentence-level citation. For each dataset used in this work, we therefore examined how the provided evidence labels correspond to the citation targets required by our evaluation. These limitations affect all methods in our experiments equally, since every method is evaluated against the same labels. Where possible, we further examine such mismatches through our LLM-as-a-judge analysis in Appendix G.1. We summarize the dataset-specific limitations below.

HotpotQA and WiCE. HotpotQA and WiCE provide human-annotated supporting evidence at or near the sentence level. HotpotQA annotators identify supporting sentences required to answer each question (Yang et al., 2018), while WiCE annotates evidence supporting individual claims or subclaims (Kamoi et al., 2023). Although these provide the closest match to our task, human annotations need not exhaust every sentence that could validly support an answer. Consequently, a citation may be valid even when it is absent from the benchmark labels.

MuSiQue and TyDiQA. MuSiQue and TyDiQA provide evidence at a coarser granularity than required for sentence-level citation. MuSiQue identifies supporting evidence at the paragraph level (Trivedi et al., 2022), whereas TyDiQA-GoldP identifies a gold passage together with the answer span within that passage (Clark et al., 2020). ARC-JSD evaluates these datasets at the sentence level (Li et al., 2026), but does not specify in sufficient detail how the sentence-level labels are derived. We therefore infer them from the provided annotations. For MuSiQue, we identify the sentence within the labeled supporting paragraph that contains the corresponding answer string. For TyDiQA, we identify the sentence within the labeled passage that contains the annotated answer span. These inferred labels provide a practical sentence-level target, but may not capture every sentence that constitutes valid supporting evidence.

G.3 Detailed Breakdown of Valid Citations Not Labelled

Table 13 and Figure 11 show the detailed breakdown of LLM-judged valid citations that are not labeled in the benchmarks by model and individual methods. They show that CiteProbe is con-

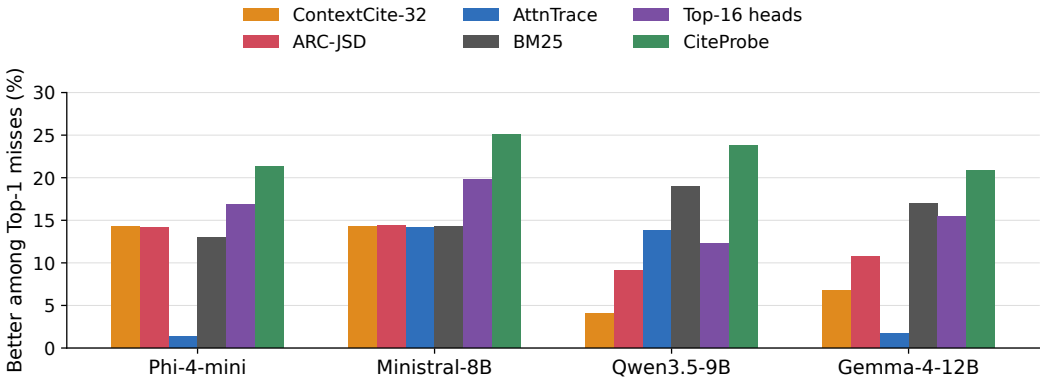


Figure 11: Top-1 misses where the selected citation is judged *better* than the best labeled citation, grouped by backbone model. Each bar is the *Better* percentage for one method in Table 13, using that model-method pair's own misses as the denominator.

sistently better at selecting valid citations that are not labeled in the benchmark across all models, as judged by Claude Opus 5.

Table 13: Top-1 failures the LLM-judge found valid despite being unlabeled.

Method	Misses	Better	Equal	Weaker	Total Valid
Phi-4-mini-instruct					
ContextCite-32	461	66 (14.3%)	51 (11.1%)	94 (20.4%)	211 (45.8%)
ARC-JSD	409	58 (14.2%)	31 (7.6%)	108 (26.4%)	197 (48.2%)
AttnTrace	1433	19 (1.3%)	31 (2.2%)	31 (2.2%)	81 (5.7%)
BM25	469	61 (13.0%)	52 (11.1%)	81 (17.3%)	194 (41.4%)
Top-16 heads	363	61 (16.8%)	44 (12.1%)	71 (19.6%)	176 (48.5%)
CITEPROBE	305	65 (21.3%)	30 (9.8%)	67 (22.0%)	162 (53.1%)
Ministral-3-8B-Instruct					
ContextCite-32	567	81 (14.3%)	49 (8.6%)	99 (17.5%)	229 (40.4%)
ARC-JSD	577	83 (14.4%)	42 (7.3%)	115 (19.9%)	240 (41.6%)
AttnTrace	549	78 (14.2%)	68 (12.4%)	86 (15.7%)	232 (42.3%)
BM25	539	77 (14.3%)	72 (13.4%)	91 (16.9%)	240 (44.5%)
Top-16 heads	481	95 (19.8%)	56 (11.6%)	86 (17.9%)	237 (49.3%)
CITEPROBE	386	97 (25.1%)	45 (11.7%)	45 (11.7%)	187 (48.4%)
Qwen3.5-9B					
ContextCite-32	1622	66 (4.1%)	39 (2.4%)	61 (3.8%)	166 (10.2%)
ARC-JSD	717	65 (9.1%)	59 (8.2%)	108 (15.1%)	232 (32.4%)
AttnTrace	542	75 (13.8%)	70 (12.9%)	84 (15.5%)	229 (42.3%)
BM25	410	78 (19.0%)	81 (19.8%)	62 (15.1%)	221 (53.9%)
Top-16 heads	651	80 (12.3%)	76 (11.7%)	101 (15.5%)	257 (39.5%)
CITEPROBE	391	93 (23.8%)	56 (14.3%)	59 (15.1%)	208 (53.2%)
Gemma-4-12B-it					
ContextCite-32	1070	73 (6.8%)	58 (5.4%)	67 (6.3%)	198 (18.5%)
ARC-JSD	573	62 (10.8%)	67 (11.7%)	91 (15.9%)	220 (38.4%)
AttnTrace	1958	33 (1.7%)	33 (1.7%)	32 (1.6%)	98 (5.0%)
BM25	424	72 (17.0%)	78 (18.4%)	70 (16.5%)	220 (51.9%)
Top-16 heads	423	65 (15.4%)	73 (17.3%)	84 (19.9%)	222 (52.5%)
CITEPROBE	342	71 (20.8%)	59 (17.3%)	93 (27.2%)	223 (65.2%)

H LEXICAL COMPETITION ACROSS MODELS AND BENCHMARKS

Section 4.4 examines whether citation methods become less accurate when an unlabeled sentence resembles the generated answer as closely as the labeled evidence. RHR is defined for each answer and candidate set and is therefore shared across citation methods. An RHR near 1 means that the strongest distractor and strongest labeled sentence have similar lexical overlap with the answer; a value above 1 means that the distractor has greater overlap. In Figure 12, the number beside each model and benchmark pair is its mean RHR, while each heatmap cell reports the correlation between RHR and the corresponding method's performance.

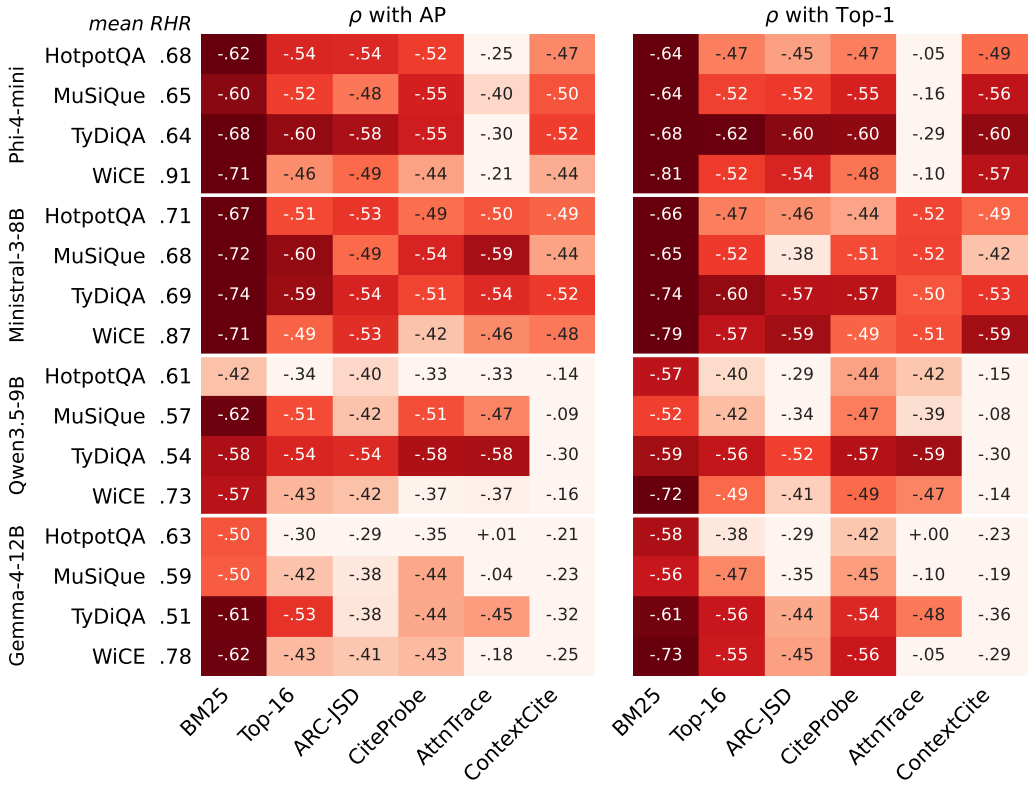


Figure 12: Lexical competition by model, benchmark, and citation method. The number beside each benchmark is the mean RHR for that model and benchmark pair. Each heatmap cell is the Spearman correlation between example-level RHR and AP or Top-1; darker cells indicate a stronger performance decline as lexical competition increases. RHR is shared across methods within each row.

Lexical competition varies more consistently across benchmarks than across LLMs. WiCE has the highest mean RHR for every LLM, ranging from 0.735 to 0.909, compared with dataset means of 0.595 to 0.657 for TyDiQA, MuSiQue, and HotpotQA. WiCE therefore presents a distinct difficulty: its unlabeled candidates often resemble the answer nearly as much as its labeled evidence does. RHR also varies across LLMs because it is computed against each model's generated answer. Qwen3.5-9B and Gemma-4-12B-it have lower mean RHRs than Phi-4-mini-instruct and Ministral-3-8B-Instruct, showing that answer wording partly determines the lexical difficulty of citation ranking.

The method-level correlations reveal a separate pattern. BM25 has the strongest negative correlation in all 32 combinations of model, benchmark, and metric, confirming that RHR captures the expected failure mode of lexical matching. CITEPROBE is less negatively correlated than BM25 in all 32 comparisons and than the Top-16-head baseline in 21. The comparison with ARC-JSD is mixed, with CITEPROBE closer to zero in 13 of 32 comparisons. Thus, CITEPROBE remains affected by lexical competition, but its performance cannot be explained as BM25-style word matching.

Learning how to combine the attention heads also reduces lexical dependence relative to equal-weight head selection.

I Computational Efficiency Analysis

Further to the analysis described in Section 4.6, which is evaluated on a specific hardware setting, we derive a generalizable formulation by expressing computational overheads in terms of the number of source sentences, C. Table 14 shows the extra inference overhead in terms of prefill tokens, expressed as a function of C. In each additional pass, AttnTrace keeps 40% ($0.4C$) of the source sentences, ContextCite-32 keeps 50% ($0.5C$) on average, while ARC-JSD removes only one. Assuming 20 tokens per sentence, the extra tokens processed are $240C$ for AttnTrace, $320C$ for ContextCite-32, and $20C(C-1)$ for ARC-JSD. At the breakeven point, ARC-JSD has lower extra prefill than both baselines for $C < 13$ with AttnTrace, and $C < 17$ with ContextCite-32. Beyond this, its quadratic cost dominates. CiteProbe does not incur extra prefill.

Table 14: Extra prefill overhead. C is the number of source sentences.

Method	Add. passes	Sent./ pass	Extra prefill
CiteProbe	0	—	0
AttnTrace	30	$0.4C$	$12C$
ContextCite-32	32	$0.5C$	$16C$
ARC-JSD	C	$C-1$	C^2-C

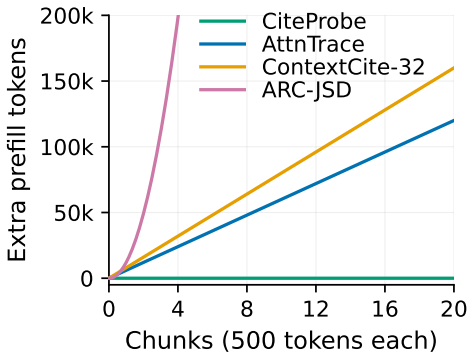


Figure 13: Extra prefill tokens processed.

To put these costs in a common RAG setting with chunks of around 500 tokens (Wang et al., 2024), Figure 13 shows how this workload grows infeasibly as more chunks are retrieved. At just 10 chunks, extra prefill reaches 60k tokens for AttnTrace, 80k for ContextCite-32, and 1.25M for ARC-JSD. This shows that CiteProbe is significantly more compute efficient.

J Model Surrogacy: Citing a Closed Model's Answer

CiteProbe reads attention weights, which closed-source model APIs do not expose. However, the same context and generated answer can instead be passed through an open-source surrogate model fitted with CiteProbe under teacher forcing. Let P denote the closed-source parent model and Q the open-source surrogate. The key premise is that the attention produced by Q can remain informative about citation even when the answer sequence itself was generated by P. In other words, the surrogate need only preserve the citation-relevant signal, rather than approximate the parent model in full. We test this through an ablation using two parent models, Claude Opus 5 and GPT-5.6 Sol, and four surrogate models: Phi-4-mini-instruct, Ministral-3-8B-Instruct, Qwen3.5-9B, and Gemma-4-12B-it, and find CiteProbe surrogacy to be feasible.

J.1 Setup

From the same four datasets, we generate answers with Claude Opus 5 and GPT-5.6 Sol. Let C be the context, q the question, and $y = P(C, q)$ the generated answer. Then, each (C, q, y) is teacher forced through the four open models to get the surrogate's attention weights for scoring, per Equations 1 to 3. The probes and head rankings are the leave-one-out checkpoints behind Table 1. We filter for only examples that the closed-source models answered correctly, which leaves 2,839 examples for Claude Opus 5 and 2,541 for GPT-5.6 Sol. Since we can also employ surrogacy with the other methods, we included them in this ablation.

J.2 RESULTS

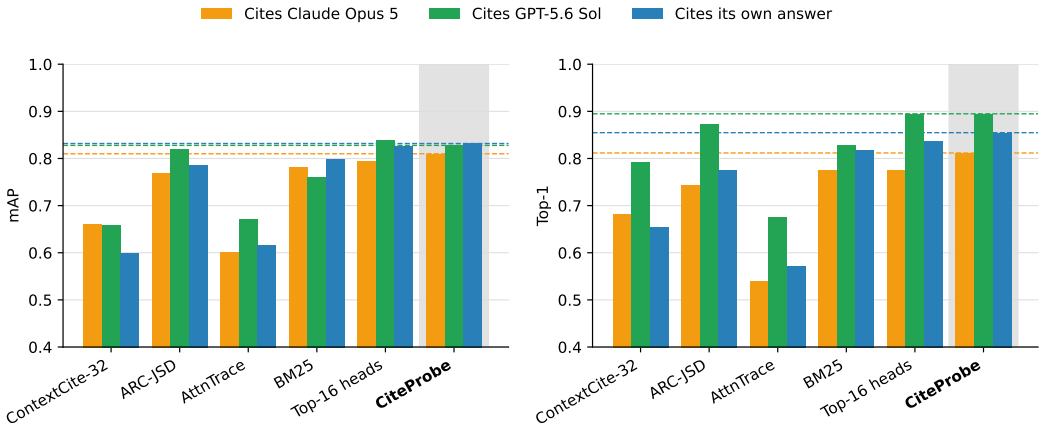


Figure 14: Citation quality under surrogacy, by method. Each bar is a method's mean over the four surrogates and the four datasets, for the two parent models and for the surrogate citing its own answer. Dashed lines mark CITEPROBE's level for each series.

The citation signal survives model surrogacy. Figure 14 shows that surrogacy works to a large extent as Claude Opus 5 and GPT-5.6 Sol yield mAP and Top-1 close to the model's own results. Specifically, on Claude Opus 5's answers (Table 15), mean mAP differs from the self-answered probe by at most 0.036, and on GPT-5.6 Sol's answers (Table 15) by at most 0.024. In the best case, Ministral-3-8B-Instruct matches its self-answered mAP and Top-1 to the third decimal on Claude Opus 5's answers. Top-1 does not move uniformly. It falls by up to 0.074 on Claude Opus 5's answers and rises for every surrogate on GPT-5.6 Sol's. A cost that reverses with the parent model points to the answer itself rather than to surrogacy, and the failure analysis below traces it to answer length. Surrogacy therefore holds.

Table 15: Citing Claude Opus 5's and GPT-5.6 Sol's answers with an open surrogate, as mean mAP and Top-1 over the four datasets. BM25 reads only text, so it has one value per answerer. For reference, the CITEPROBE *(self)* row is its self-answered performance, scored on the examples that both the surrogate and its parent model answered correctly.

Method	Phi-4-mini		Ministral-3-8B		Qwen3.5-9B		Gemma-4-12B	
	mAP	Top-1	mAP	Top-1	mAP	Top-1	mAP	Top-1
Claude Opus 5								
ContextCite-32	.655	.680	.683	.704	.679	.698	.625	.648
ARC-JSD	.763	.744	.769	.744	.767	.734	.774	.755
AttnTrace	.460	.312	.756	.751	.788	.782	.404	.312
CITEPROBE	**.802**	**.797**	**.824**	**.839**	**.808**	**.810**	**.807**	**.801**
CITEPROBE *(self)*	.833	.856	.823	.839	.835	.862	.844	.875
BM25	mAP .782, Top-1 .776							
GPT-5.6 Sol								
ContextCite-32	.644	.780	.680	.810	.669	.810	.641	.769
ARC-JSD	.813	.870	.822	.869	**.838**	.879	.804	.873
AttnTrace	.560	.492	.804	.870	.822	.886	.504	.453
CITEPROBE	**.824**	**.886**	**.828**	**.896**	.823	**.896**	**.836**	**.901**
CITEPROBE *(self)*	.836	.860	.831	.849	.847	.878	.851	.882
BM25	mAP .761, Top-1 .829							

CITEPROBE remains the best method under model surrogacy. CITEPROBE achieves the best mean mAP and Top-1 for all four surrogates on Claude Opus 5's answers, ahead of every baseline. On GPT-5.6 Sol's answers, it leads the attribution baselines on mean Top-1 for all four surrogates

and on mean mAP for three of the four surrogates. The perturbation and attention baselines follow the same pattern as in Table 1, with ContextCite-32 weakest throughout and AttnTrace unstable on Phi-4-mini-instruct and Gemma-4-12B-it.

Failure analysis under model surrogacy. Since the probe is trained on the surrogate's own attention, teacher forcing a parent's text may put it off distribution and result in degradation. Following the same setup, Table 16 shows where that degradation lands. Most error types decrease in percentage. Only *Adjacency* moves against the trend, rising from 8.4% to 14.3% on Claude Opus 5's answers and falling to 3.2% on GPT-5.6 Sol's. Answer length is the likely reason. The surrogates' own answers average 56.1 citable words and 8.4% adjacency, and the two parents sit either side of that. GPT-5.6 Sol is shorter at 15.7 words and 3.2%, while Claude Opus 5 is longer at 65.5 words and 14.3%. A longer answer restates more of the context, so the sentence beside the right one also looks relevant.

Table 16: Results of failure analysis per the failure taxonomy. "Own answer" means our consolidated results from models we experimented with in this paper (Phi-4-mini, Ministral-3-8B, Qwen3.5-9B and Gemma-4-12B), against Claude Opus 5 and GPT-5.6 Sol under model surrogacy.

		Ranking error (%)					Valid but Unlabeled (%)			
Cited answer	n	Entity	Hop	Adjac.	**Unrel.**	**Subtotal**	**Better**	Equal	Weaker	**Subtotal**
Own answer	1424	21.9	13.3	8.4	1.7	45.2	22.9	13.3	18.5	54.8
Claude Opus 5	2021	20.0	11.6	14.3	0.3	46.3	16.6	10.2	27.0	53.7
GPT-5.6 Sol	1091	24.6	9.7	3.2	0.1	37.6	25.8	18.2	18.4	62.4

Therefore, the surrogate's attention remains informative about which context sentences support the parent's answer, and no new kind of failure appears. Since CITEPROBE only needs that attention, a surrogate extends it to closed-source models for one forward pass through a small open model, without ablations, retraining, or access to the parent beyond its generated text. The extension gives up little in ranking quality. Even Phi-4-mini-instruct, the smallest surrogate at 3.8B parameters, leads every attribution baseline on both parents' answers.

K COMPARISON WITH GUIDANCE – A FULL END-TO-END RAG AND CITATION SYSTEM

Zhan et al. (2026) introduced Guidance, an end-to-end framework that includes both RAG and sentence-level citation. To use the framework, Guidance requires the source documents to be preprocessed by an LLM into its own prescribed Prefix-Tail fact-pair structure, which incurs substantial computational cost. For this reason, we could only conduct an ablation to compare the citation capabilities by applying CITEPROBE to the same models and benchmark used in their study.

For this comparison, we take the Guidance results directly from Zhan et al. (2026) and train CITEPROBE on the same three small open-source models: Mistral-7B-Instruct-v0.3, Llama-3.1-8B-Instruct, and Qwen3-8B. We use only MuSiQue, TyDi and WiCE train sets, given that HotpotQA is part of the LongBench-Cite benchmark used by Guidance. As the original model generations were not available, we reproduced the vanilla generations using greedy decoding to reduce generation variance.

CITEPROBE confidence operating threshold. For CITEPROBE, we need a calibrated confidence threshold, which we also obtain in the out-of-domain setting, using validation datasets to obtain an operating threshold that maximizes F1. The optimal thresholds obtained are 0.86 for Mistral-7B-Instruct-v0.3, 0.87 for Llama-3.1-8B-Instruct, and 0.90 for Qwen3-8B.

Citation evaluation. We then perform judging using GPT-4o per the protocol prescribed by Zhan et al. (2026). To reduce variance, we performed two separate judging runs. Table 17 shows that the variance between the runs is minimal, indicating stability in the LLM-as-a-judge protocol in this regard. Even so, we were not able to reproduce Guidance's vanilla results fully. Two out of three reproduced vanilla results are already stronger than the reported Guidance results. For Llama-3.1-8B-Instruct, the reproduced vanilla result reaches an F1 of 51.3, compared with 39.6 for the reported

vanilla model and 50.1 with Guidance. Similarly, Qwen3-8B reaches 58.4 in our reproduction, compared with 53.3 reported for vanilla and 56.2 with Guidance.

Results. CITEPROBE achieves the highest average citation F1 among all methods reported in Table 18, with Llama-3.1-8B-Instruct reaching 71.0 F1. Importantly, this performance comes without the computationally expensive document preprocessing required or the retrieval and generation-time machinery used by Guidance or NEST. Once trained, CITEPROBE operates directly on the attention produced during the original generation pass and therefore requires no additional LLM inference at test time. These results show that competitive sentence-level citation can be added to an existing white-box LLM with little more than a lightweight readout, without changing how the model generates its answer.

Table 17: GPT-4o-as-a-judge variance over two runs at a temperature of zero.

Model / file	Round 1	Round 2	Δ
Mistral-7B-Inst-v0.3			
Vanilla (repro.)	29.94	29.96	+0.03
+ CITEPROBE	60.32	60.35	+0.03
Llama3.1-8B-Inst			
Vanilla (repro.)	51.33	51.39	+0.06
+ CITEPROBE	71.01	70.80	-0.21
Qwen3-8B			
Vanilla (repro.)	58.36	58.13	-0.23
+ CITEPROBE	67.13	67.10	-0.03

Table 18: Citation recall (R), citation precision (P), and citation F1 (F1) evaluated on the LongBench-Cite benchmark, comparing CITEPROBE with results from Zhan et al. (2026) (marked with †), excluding fine-tuned models specialized to perform citation as they are unrelated to the aim of this research.

Model	Longbench-Chat			MultifieldQA			HotpotQA			Dureader			GovReport			Avg. F1
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	
Proprietary†																
GPT-4o	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	65.6
Claude-3-sonnet	52.0	67.8	55.1	64.7	85.8	71.3	46.4	65.8	49.9	67.7	**89.2**	**75.5**	77.4	**93.9**	84.1	67.2
GLM-4	47.6	53.9	47.1	72.3	80.1	73.6	47.0	50.1	44.4	73.4	82.3	75.0	82.8	93.4	**87.1**	65.4
Large open-source†																
Llama-3.1-70B	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	40.4
Mistral-Large	19.8	23.9	19.0	71.8	80.7	73.8	34.5	40.9	32.1	58.3	67.0	60.1	67.9	79.6	72.5	51.5
Small open-source																
Mistral-7B-Inst-v0.3																
Vanilla	17.4	16.9	11.8	43.6	61.3	46.4	29.4	37.6	28.9	27.4	46.0	29.6	53.3	82.7	62.1	35.8
+ Guidance	20.2	27.6	20.2	52.9	68.1	55.5	33.0	37.4	30.4	36.4	52.7	38.2	59.9	93.4	70.6	43.0
Vanilla (repro.)	15.6	25.6	14.6	41.3	55.0	43.2	15.8	18.0	13.9	29.5	40.8	27.0	43.0	75.7	50.9	29.9
+ CITEPROBE	62.9	43.2	46.2	77.8	61.8	64.8	67.1	50.1	52.0	57.7	55.3	53.8	**90.8**	72.8	80.2	59.4
Llama3.1-8B-Inst																
Vanilla	26.6	37.3	25.8	43.3	66.6	49.3	25.1	39.6	26.6	40.3	59.6	44.9	40.2	80.0	51.3	39.6
+ Guidance	44.8	49.1	36.2	57.4	77.2	62.6	31.0	51.9	35.0	46.5	60.3	48.5	57.1	90.0	68.3	50.1
Vanilla (repro.)	40.1	56.8	43.6	55.4	78.7	61.9	32.0	55.6	37.1	47.5	69.2	52.2	52.5	78.7	61.7	51.3
+ CITEPROBE	**78.9**	48.6	57.4	87.6	68.0	73.8	**78.0**	56.6	**62.4**	73.3	66.4	67.6	89.8	78.0	83.0	68.9
Qwen3-8B																
Vanilla	31.3	41.1	29.4	71.9	78.2	72.7	37.6	42.4	34.9	60.1	65.9	60.0	62.6	80.8	69.5	53.3
+ Guidance	39.4	34.1	29.1	73.1	80.0	74.1	39.0	47.0	37.8	61.8	71.3	63.8	65.6	82.4	72.2	56.2
Vanilla (repro.)	41.8	54.7	45.3	71.1	84.6	74.7	28.3	46.2	31.0	63.9	77.8	68.2	67.4	82.9	72.6	58.4
+ CITEPROBE	77.9	55.5	**61.2**	**90.4**	73.5	78.3	75.2	53.5	57.8	**74.3**	74.4	73.0	84.8	79.3	80.6	**70.2**
Other																
NEST	60.4	**71.9**	60.9	69.4	87.0	74.7	60.3	**68.5**	58.3	67.7	86.5	74.4	69.9	83.3	74.7	68.6

L CITATION BEHAVIOR ON CORRECT AND INCORRECT ANSWERS

Our citation-ranking evaluation only reports mAP and Top-1 on correctly answered examples, focusing on the ability of methods and models to rank citations. Here, we extend the analysis to both correct and incorrect answers, and analyze their classification abilities.

L.1 SETUP

We determine a citation threshold for each method using the validation splits. For ranking methods (all except CITEPROBE), we combine all sentence scores from these validation splits, compute a single mean and standard deviation, and select the z-score threshold that maximizes citation F1. For CITEPROBE, since it already outputs a confidence score, we apply it to the validation set to find the confidence threshold that maximizes F1.

L.2 CORRECTLY ANSWERED EXAMPLES

Table 19 shows the results, and CITEPROBE remains the strongest when judged on Average F1. Overall, these results reinforce that CITEPROBE reliably identifies supporting evidence across datasets.

Table 19: Citation recall (R), precision (P), and weighted average F1.

Method	HotpotQA			MuSiQue			TyDiQA			WiCE			Avg. F1
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	
Phi-4-mini-instruct													
ContextCite-32	48.0	69.8	56.9	43.9	**71.0**	54.3	75.0	62.3	68.1	52.7	41.4	46.4	54.2
ARC-JSD	55.7	**71.2**	62.5	53.2	69.3	60.2	81.2	55.0	65.6	59.2	39.5	47.4	57.2
AttnTrace	4.0	25.3	7.0	11.4	24.1	15.5	**99.4**	24.7	39.6	53.1	30.0	38.4	23.4
BM25	48.4	54.9	51.4	**77.9**	30.3	43.6	13.8	49.4	21.5	42.5	35.5	38.7	41.9
CITEPROBE (Ours)	**76.4**	66.6	71.2	62.6	69.0	**65.6**	19.5	**94.9**	32.3	58.0	**60.7**	**59.3**	**61.3**
Ministral-3-8B-Instruct													
ContextCite-32	53.5	64.7	58.6	51.6	67.5	58.5	75.5	48.5	59.1	55.3	43.0	48.4	55.3
ARC-JSD	61.9	64.0	62.9	60.5	64.7	62.5	80.1	48.3	60.3	59.1	40.0	47.7	57.6
AttnTrace	62.9	57.2	59.9	12.8	**83.2**	22.1	**97.0**	26.5	41.7	**61.3**	34.8	44.4	43.6
BM25	52.4	50.2	51.3	**79.0**	37.9	51.3	26.8	60.4	37.2	59.2	33.3	42.6	47.1
CITEPROBE (Ours)	70.6	**68.4**	69.5	71.2	66.7	**68.9**	19.6	**90.4**	32.3	54.4	**56.9**	**55.6**	**61.1**
Qwen3.5-9B													
ContextCite-32	20.2	44.0	27.7	15.5	53.6	24.1	46.4	57.5	51.3	16.0	36.7	22.2	27.7
ARC-JSD	69.9	54.8	61.4	83.6	29.6	43.8	69.5	52.8	60.0	51.1	43.6	47.0	52.2
AttnTrace	52.7	66.3	58.7	16.9	**80.0**	27.9	**97.0**	32.7	48.9	52.9	53.9	53.4	47.4
BM25	62.4	46.9	53.6	**84.7**	27.9	42.0	10.8	67.8	18.6	41.2	47.5	44.1	43.7
CITEPROBE (Ours)	**74.6**	65.7	69.9	74.3	65.8	**69.8**	22.4	**88.3**	35.7	56.9	**62.2**	**59.4**	**63.1**
Gemma-4-12B-it													
ContextCite-32	31.3	49.7	38.4	22.9	47.1	30.9	63.9	50.0	56.1	31.6	25.5	28.2	35.3
ARC-JSD	54.5	68.7	60.8	53.6	45.9	49.4	77.2	51.9	62.1	59.0	31.1	40.7	51.9
AttnTrace	42.4	7.6	12.9	16.3	5.5	8.2	**99.7**	23.3	37.8	21.3	6.6	10.1	13.5
BM25	61.4	55.3	58.2	**78.2**	31.2	44.6	16.5	54.3	25.3	52.5	40.3	45.6	47.2
CITEPROBE (Ours)	**74.1**	71.9	**73.0**	69.1	**68.8**	**68.9**	48.2	**91.2**	63.1	52.5	**64.4**	57.8	**66.3**

L.3 INCORRECTLY ANSWERED EXAMPLES

Having examined citation decisions for correct answers, we extend the analysis to incorrect answers. Specifically, we ask whether a method's strongest citation score can distinguish correctly answered examples from incorrectly answered ones. Table 20 reports the resulting AUROC. Overall, the methods show limited ability to make this distinction, with most scores close to chance (0.5). AttnTrace performs somewhat better, reaching a mean AUROC of 0.608.

Table 20: AUROC of an example's highest sentence score for distinguishing correctly answered from incorrectly answered examples, averaged over the four held-out corpora. An AUROC of 0.5 corresponds to chance.

Method	Phi-4-mini	Ministral-3-8B	Qwen3.5-9B	Gemma-4-12B	Mean
ContextCite-32	.500	.487	.482	.505	.494
ARC-JSD	.514	.502	.428	.466	.477
AttnTrace	.590	**.632**	**.577**	**.633**	**.608**
BM25	.458	.444	.430	.460	.448
CITEPROBE (Ours)	.538	.548	.537	.549	.543

One explanation is that these methods are designed to identify supporting citations for a given answer, rather than to determine whether the answer itself is unsupported or incorrect. The available

benchmarks likewise provide evidence annotations for correct answers and are not designed to supervise this distinction. This therefore represents a limitation of both current citation-method design and available benchmark supervision, and motivates future work on jointly assessing answer correctness and citation support.