Citation Failure: Definition, Analysis and Efficient Mitigation

Jan Buchmann and **Iryna Gurevych**

Ubiquitous Knowledge Processing Lab (UKP Lab)
Department of Computer Science and Hessian Center for AI (hessian.AI)
Technical University of Darmstadt
www.ukp.tu-darmstadt.de

Abstract

Citations from LLM-based RAG systems are supposed to simplify response verification. However, this goal is undermined in cases of *citation failure*, where a model generates a helpful response, but fails to generate citations to complete evidence. In contrast to previous work, we propose to disentangle this from *response failure*, where the response itself is flawed, and citing complete evidence is impossible. To address citation failure, this work follows a two-step approach: (1) We study when citation failure occurs and (2) how it can be mitigated efficiently. For step 1, we extend prior work by investigating how the relation between response and evidence affects citation quality. We introduce CITECONTROL, a benchmark that systematically varies this relation to enable the analysis of failure modes. Experiments show that failures increase with relational complexity and suggest that combining citation methods could improve performance, motivating step 2. To study the efficient improvement of LLM citation, we propose CITENTION, a framework integrating generative, attention-based, and retrieval-based methods. Results demonstrate substantial citation improvements on CITECONTROL and in transfer settings. We make our data and code publicly available.[1]

1 Introduction

Citations for LLM-generated responses can improve usefulness and accountability in RAG (Lewis et al., 2020) by enabling users to verify responses through evidence. However, this advantage weakens when a model produces a valid response without sufficient cited evidence: users must then either search additional sources or discard an otherwise helpful answer. To address this important issue, we adopt a two-step approach: We (1) analyze when models generate insufficient citations, and (2) investigate efficient mitigation strategies.

Step 1: Analyzing insufficient citations Approaches to obtaining citations for LLM-generated responses (also known as attributions; Rashkin et al. 2023) operate in three distinct paradigms: (1) In *generative citation*, models generate responses and citations in a single step (Huang et al., 2024; Droste et al., 2026). (2) *Attribution-first* approaches prompt LLMs to select evidence pieces before generation, thereby enforcing a strict mapping between evidence and response (Slobodkin et al., 2024; Wright et al., 2025). (3) *Post-hoc*-approaches use LLMs (Balasubramanian et al., 2026; Hirsch et al., 2025) or retrievers (Ramu et al., 2024; Sancheti et al., 2024) to retrieve citations for responses generated by a different model. LLM-based attribution-first and post-hoc approaches suffer from high complexity and long inference times, as they require multiple LLM calls. Therefore, we focus on generative citation, which is simpler and faster, and use retrieval-based post-hoc approaches as efficient baselines.

Prior work that studies generative citation does not differentiate between *citation failure* and *response failure* (e.g. Gao et al. 2023; Zhang et al. 2025a; Wu et al. 2025). We believe that this is an important gap. To illustrate this distinction, consider a question about the time of a coup in the capital of the DR Congo and source documents from a web search (Fig. 1). Multi-hop reasoning is required: Document [2] states that Kinshasa is the capital, while [4] reports a coup there on 28 March 2004. In this situation, an LLM may exhibit: (1) *Response failure*: where the generated response is invalid, i.e. not supported by any com-

[1] https://github.com/UKPLab/tacl2026-citation-failure

Transactions of the Association for Computational Linguistics, vol. 14, pp. 1453–1481, 2026. https://doi.org/10.1162/TACL.a.736
Action Editor: Emily Pitler. Submission batch: 9/2025; Revision batch: 3/2026; Published 7/2026.

bination of source documents (e.g. "1960"), and any cited evidence cannot support it. (2) *Citation success*: where the response is valid ("28 March 2004") and the evidence is complete ("[2] [4]"). (3) *Citation failure*: where the response is valid, but the evidence is incomplete (e.g. by citing [3] instead of [2]).

Most work analyzing insufficient citations focuses on source document properties and composition (e.g., Wu et al. 2025; Sorodoc et al. 2025). In LLM-based post-hoc citation, the relationship between evidence and statements—such as varying reasoning complexity (Zhu et al., 2025; Balasubramanian et al., 2026)—strongly affects citation performance. However, its impact on generative citation remains largely unexplored and requires further study.

Two aspects of existing datasets and benchmarks make them unsuitable for this analysis: First, they do not allow rigorous verification of response correctness, making it hard to separate response from citation failure. Second, they rely on LLM-based evaluators (Tang et al., 2024; Honovich et al., 2022), whose accuracy can drop to ∼50% in complex cases (Hu et al., 2025). As a result, a dedicated benchmark and analysis of how the response–evidence relationship affects citation failure is needed.

To address these gaps, we ask:

RQ1: How does the response–evidence relation affect citation failure? Building on prior work in citation evaluation (Hu et al., 2025) and intertextuality (Kuznetsov et al., 2022), we define *reasoning type* and *overtness* as core properties of this relation and introduce CITECONTROL, a benchmark that systematically varies them. All instances include verifiable answers and known evidence, enabling us to control for response failure and avoid reliance on error-prone citation evaluators.

Experiments on CITECONTROL with generative citation from 18 LLMs across five families, along with retrieval-based post-hoc baselines, validate the need to separate citation from response failure: citation performance is consistently higher for correctly answered instances. The results also reveal distinct failure modes: smaller models struggle even with simple 1-to-1 relations, while all models face difficulties in multi-document reasoning and tend to under-generate citations. Comparing generative and retrieval-based methods shows that different approaches suit different relation types, indicating that combining them can improve performance.

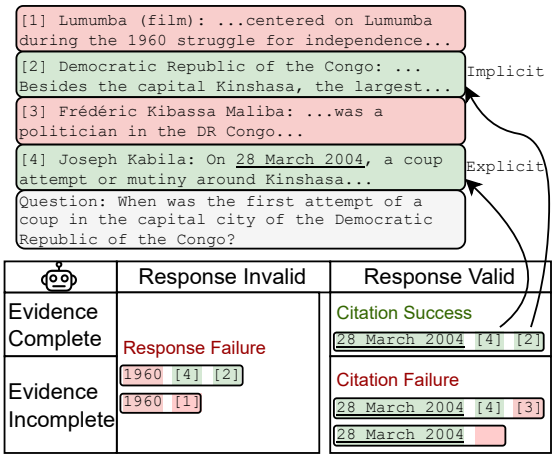

Figure 1: Citation Example:[2] An LLM receives multiple documents and a question. The confusion matrix shows the possible outcomes for generated response and evidence. The response-evidence relation has reasoning type `multi-hop`. It is `explicit` for the response and [4], and `implicit` for the response and [2].

Step 2: Mitigating citation failure efficiently Most approaches to improving generative citation rely on fine-tuning (e.g., Huang et al. 2024; Zhang et al. 2025a; Li et al. 2024a), which involves complex data collection and training setups, risks catastrophic forgetting (Biderman et al., 2024; Dou et al., 2024), and demands substantial resources. Similarly, attribution-first and post-hoc methods often require multiple additional LLM calls, making them costly. In contrast, retrieval-based approaches (Ramu et al., 2024; Sancheti et al., 2024) are efficient but limited by their small parameter capacity. This highlights the need for methods that are both efficient and powerful in mitigating citation failure.

In related settings, several works have successfully used attention values for reranking (Zhang et al., 2025b; Chen et al., 2025) and retrieving parts of the context that most influenced an LLM's output (Cohen-Wang et al., 2025). This efficiently exploits the capabilities of LLMs, as attention values are available "for free" during generation,[3] but

[2] Robot icon created by Good Ware - Flaticon

[3] In practice, attention implementations such as Flash Attention (Dao, 2024) do not give access to attention values, making an additional forward pass the most time-efficient im-

the potential of these approaches to mitigate citation failure has not been studied. Therefore, we aim to answer the research question:

RQ2: How can attention values be used to efficiently mitigate LLM citation failure? We contribute CITENTION, a framework designed for studying this problem that integrates generative, attention-based, and retrieval-based citation. We perform experiments with CITENTION across 6 LLMs (1.7–70B parameters) on CITECONTROL and two transfer datasets.

Our key findings are: First, attention-based citation is effective for extractive and abstractive responses (at least +10% average relative improvement over generative citation on transfer datasets), but partially fails with complex reasoning and short abstractive responses. Second, combining generation-, attention-, and retrieval-based citation reliably improves over generative citation, including complex reasoning cases (at least +5% relative improvement averaged over all datasets). Third, masking reasoning tokens during attention computation improves attention-based citation. These findings provide clear guidance for future research on efficient LLM citation.

In summary, we make two main contributions: (1) CITECONTROL, a novel benchmark for LLM citation (§3), on which experiments reveal that there is ample room for improvement in cases with complex statement-evidence relations (§4). (2) CITENTION, a framework for studying the efficient mitigation of citation failure with attention-based and retrieval-based methods (§5), and experiments on CITECONTROL and in a transfer setting showing that failures from generation-based citation can be mitigated effectively (§6).

2 Related Work

Citations The goal of citations is to provide *corroborative* attribution, i.e. retrieving evidence E for a statement s, such that "according to E, s" is true (Rashkin et al., 2023). *Contributive* attribution is a related, but distinct paradigm: Here, the goal is to estimate the effect of parts of the context on the LLM output, often measured by changes in output probability when removing these parts (Cohen-Wang et al., 2024; Ribeiro et al., 2016).

Analyzing Generative LLM Citation A range of datasets (Malaviya et al., 2024; Kamalloo et al., 2023; Golany et al., 2024) and benchmarks (Gao et al., 2023; Liu et al., 2023) for comparing LLMs in their citation abilities have been proposed. Existing analyses focus on the properties of the source documents such as their number and combination (Koo et al., 2024; Tang et al., 2025; Buchmann et al., 2024; Wu et al., 2025), authorship information (Abolghasemi et al., 2025), and time-dependence and semantic properties of their contents (Sorodoc et al., 2025) as factors influencing citation failure. While the effect of the response-evidence relation has been studied in post-hoc citation (Zhu et al., 2025; Balasubramanian et al., 2026), this has not been studied for generative citation.

Improving Citations To improve citations for LLM responses, (1) training-based approaches collect data and design training regimes to improve generative citation capabilities (Huang et al., 2024; Penzkofer and Baumann, 2024; Zhang et al., 2025a; Li et al., 2024a) (2) LLM-based attribution-first (Slobodkin et al., 2024; Wright et al., 2025) and post-hoc approaches split attribution across multiple LLM calls (Hirsch et al., 2025; Qian et al., 2024). (3) Contributive-attribution based methods ablate context across multiple forward passes to isolate relevant sources (Chuang et al., 2025; Qi et al., 2024). (4) Retrieval-based post-hoc approaches use sparse or dense retrievers post-generation (Sancheti et al., 2024; Ramu et al., 2024). Categories (1–3) are resource-intensive at training (1) or inference (2-3), while (4) is constrained by the retriever's capacity, typically smaller than the LLM's.

Using LLM Internals for Efficient Retrieval Several works have proposed directly using LLM internals such as attention values or hidden states on related problems such as reranking (Zhang et al., 2025b; Chen et al., 2025) and contributive attribution (Cohen-Wang et al., 2025; Phukan et al., 2024; Ding et al., 2025). This exploits LLM capabilities efficiently, as no training of the LLM itself or additional LLM calls are needed. Hirsch et al. (2025) recently showed that the hidden-states-based method from Phukan et al. (2024) shows mediocre performance in a fine-grained setting, so we do not consider it here. To our knowledge, the use of attention-based methods for citation has not been investigated.

plementation to date (Cohen-Wang et al., 2025).

We make important contributions to the described research areas: In analyzing LLM citation, we are the first to provide response-failure-aware methodology and experiments investigating the effect of the response-evidence relation on citation failure. The results inform our research on improving LLM citation, where we investigate the potential of attention-based methods for corroborative citation and its combination with generation-based and retrieval-based citation for the first time.

3 CITECONTROL

To study how the response-evidence relation affects citation, we introduce CITECONTROL, a framework for evaluating and analyzing LLM citation. Unlike prior benchmarks (Gao et al., 2023; Tang et al., 2025; Buchmann et al., 2024), it separates response failure from citation failure, and avoids reliance on error-prone attribution models (Hu et al., 2025). We first formalize the citation task (§3.1), then detail how we vary response–evidence relations (§3.2), followed by datasets (§3.3) and evaluation (§3.4).

3.1 Task Formalization

An instance in CITECONTROL consists of an instruction q (e.g. a question) and a set of source documents $S = \{s_1, ..., s_{|S|}\}$ (these could come from a previous retrieval step). The task is to generate a *response* r based on S (e.g. an answer), and to cite corroborative *evidence* $E \subset S$ (see §2, Rashkin et al. 2023). The evidence should enable response verification, so models need to cite relevant source documents even if the contained information is present in their parametric knowledge.

3.2 Varying the Relation between Response and Evidence

We build on related work to define two key properties of the response-evidence relation and study their effect on LLM citation:

Reasoning Type Following Hu et al. (2025), we distinguish the types of reasoning required to infer the response from the evidence, omitting "union" due to lack of suitable data:

- `single`: Reasoning over a single evidence document.
- `multi-hop`:[4] Reasoning over a chain of facts, the final fact being the correct response (as in Fig. 1).
- `intersection`: Combining multiple facts to a new "fact" (e.g. computing the time between two events).

[4]Named "concatenation" in Hu et al. (2025)

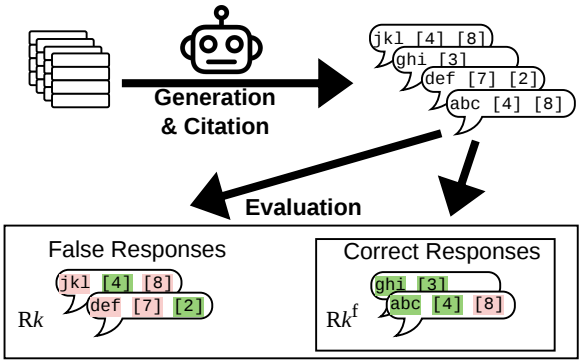

Figure 2: Evaluation strategy on CITECONTROL:[2] For Rk, all predictions are evaluated for evidence recall @k, while for Rk^{f}, the subset of predictions with correct responses are evaluated.

Overtness Hu et al. (2025) assume verbatim extraction from evidence documents, and do not differentiate their analysis between individual evidence documents. This misses the *overtness* of the response-evidence relation, which is recognized by more general research on intertextuality (Kuznetsov et al., 2022), and distinguishes:

- `implicit`: the response does not appear verbatim, but the evidence document is relevant (e.g document [2] in Fig. 1)
- `explicit`: the response appears verbatim in the evidence document (e.g. [4] in Fig. 1)

3.3 Datasets

Datasets in CITECONTROL must (1) provide known reasoning types and overtness (§3.2), (2) include verifiable responses to separate response from citation failure, and (3) specify complete ground-truth evidence to avoid reliance on error-prone evaluation models (§1). Based on these criteria, we select four datasets. See Tables 1 for a dataset overview and 4 for examples.

RepliQA (Monteiro et al., 2024) and **BoolQ-M** consist of tuples of a context paragraph, question and answer. While RepliQA answers are extractive, BoolQ-M answers are "yes" or "no". We created BoolQ-M to control for data contamination in BoolQ (Clark et al., 2019) by replacing enti-

	# train/dev/test	$\|S\|$	$\|s\|$	$\|r\|$	$\|E\|$	Reasoning	Overtness
RepliQA	6084/1933/2087	20.0	84.3	5.1	1.0	`single`	`explicit`
BoolQ-M	3394/849/2173	20.0	96.4	1.0	1.0	`single`	`implicit`
MuSiQue	15950/3988/2417	20.0	82.4	2.3	2.4	`multi-hop`	`exp` / `imp`
NeoQA	0/264/1157	19.7	283.4	6.3	1.9	`multi-hop` / `intersec.`	`exp` / `imp`

Table 1: The datasets in CITECONTROL. $|S|$ / $|E|$: Number of source / evidence documents per instance. $|s|$ / $|r|$: Number of words per source document / response.

ties with fictional entities using GPT-OSS-120B[5]. **MuSiQue** (Trivedi et al., 2022) is a dataset of 2- to 4-hop questions and answers combined with 2 to 4 evidence paragraphs and 16 to 18 distractor paragraphs from Wikipedia. **NeoQA** (Glockner et al., 2025) is a dataset of time-span and 2-hop questions on synthetic news articles. For time-span questions, models are given two events and need to compute the time span between them. For RepliQA, BoolQ-M and NeoQA, we combine evidence documents with distractors to obtain 20 source documents per instance. For details on data processing, see §A.2.

The reasoning type in RepliQA and BoolQ-M is `single`. All MuSiQue instances and NeoQA 2-hop instances are `multi-hop`, while NeoQA time-span instances are `intersection`.

The overtness of the response-evidence relation can be seen in the ROUGE-1 scores (Lin, 2004) between response and evidence shown in Table 7. It is `explicit` for RepliQA and between the response and the evidence document that contains the final answer in MuSiQue and NeoQA `multi-hop` instances. It is `implicit` for evidence documents upstream in the `multi-hop` reasoning chain, as well as for BoolQ-M and NeoQA `intersection` instances.

3.4 Evaluation

We propose *filtered* recall @k (Rk^{f}) as our primary evaluation metric (Fig. 2), which improves over existing recall-focused evaluation of citations (Gao et al., 2023; Buchmann et al., 2024) by disentangling response failure from citation failure: For a specific model, we only evaluate the subset of instances that were answered correctly. @k means that for any instance, we only evaluate the first k citations to avoid rewarding over-generation of evidence. For reference, we report recall @k on all instances (Rk) and the proportion of correctly answered instances ($\frac{|\mathcal{R}^{\text{f}}|}{|\mathcal{R}|}$). For details, see §A.3.

4 How Does the Relation Between Response Statement and Evidence Affect LLM Citation?

In this section, we use CITECONTROL to analyze the effect of the response-evidence relation on citation performance. We describe experimental details in §4.1 and results in §4.2.

4.1 Experimental Setup

Prompts We instruct models to cite by appending document indices to response statements (Fig. 1) with 3-shot prompts. For details and examples, see §A.1.

Retrieval-based oracle baselines We include `BM25`, a lexical-matching based retriever (Robertson and Zaragoza, 2009), and Dragon (`DRAG`, Lin et al. 2023), a dense retriever, which have been shown to perform well on a recent benchmark (Buchmann et al., 2024). We use the concatenation of question and ground truth answer as the query (for details see §A.6.3).

4.2 Results and Discussion

We ran 18 LLMs[6] (5 families, 0.6B – 120B) on CITECONTROL, showing results in Table 2. After analyzing the effect of our proposed filtered evaluation, we interpret the results with regard to our main research question.

Citation performance is higher on correct responses We observe that for most combinations of model and task, filtered recall @k (Rk^{f}) is higher than or equal to unfiltered evaluation (Rk). While the magnitude of this difference depends on the model and dataset, on MuSiQue and NeoQA we observe up to ∼14 and ∼8 points difference between Rk^{f} and Rk, respectively,[7] showing that dif-

[5] `https://huggingface.co/openai/gpt-oss-120b`

[6] We omit "-Instruct" specifiers for brevity. See Table 6 for model details

[7] e.g. Qwen3-32B on MuSiQue, Qwen3-1.7B on NeoQA

	RepliQA			BoolQ-M			MuSiQue			NeoQA		
	Rk	$\frac{\lvert\mathcal{R}^f\rvert}{\lvert\mathcal{R}\rvert}$	**Rk^f**	**Rk**	$\frac{\lvert\mathcal{R}^f\rvert}{\lvert\mathcal{R}\rvert}$	**Rk^f**	**Rk**	$\frac{\lvert\mathcal{R}^f\rvert}{\lvert\mathcal{R}\rvert}$	**Rk^f**	**Rk**	$\frac{\lvert\mathcal{R}^f\rvert}{\lvert\mathcal{R}\rvert}$	**Rk^f**
	No Reasoning											
Ministr-8B	96.9	73.2	98.3	99.7	81.9	99.8	31.3	26.7	44.0	51.0	47.6	52.5
Mistr-S-3.2-24B	99.2	85.2	99.9	100	87.3	99.9	61.5	50.0	70.1	75.6	71.7	73.2
Llama-3.2-1B	39.1	63.1	41.4	43.7	58.4	44.6	10.0	11.5	14.1	8.0	19.1	8.6
Llama-3.2-3B	96.0	73.0	97.1	98.5	70.8	98.6	39.1	24.7	46.0	47.9	40.5	48.6
Llama-3.1-8B	98.9	78.2	99.6	99.0	80.1	99.3	44.7	36.4	54.3	63.6	55.4	61.6
Llama-3.3-70B	97.0	79.4	98.2	98.4	78.4	98.8	41.1	29.6	49.1	61.5	52.8	61.4
Phi-4-Mini	93.7	68.9	95.4	92.8	63.2	97.6	41.2	21.8	53.5	61.8	42.5	63.0
Phi-4	97.8	80.6	99.2	98.5	79.0	99.4	64.7	36.3	75.4	73.7	49.4	72.2
Oracle-BM25	98.5	100	98.5	68.3	100	68.3	48.7	100	48.7	70.8	100	70.8
Oracle-DRAG	99.4	100	99.4	100	100	100	70.4	100	70.4	68.0	100	68.0
	Reasoning											
Qwen3-0.6B	75.9	14.2	83.5	88.2	70.0	89.9	30.5	1.1	50.0	27.6	37.0	35.7
Qwen3-1.7B	91.5	70.5	94.5	95.5	81.4	96.1	31.8	29.7	40.9	45.5	48.3	53.0
Qwen3-4B	97.0	71.2	99.0	99.2	87.5	99.6	44.0	34.7	63.2	70.1	67.2	75.2
Qwen3-8B	96.9	80.2	98.6	98.8	87.0	99.8	54.9	50.4	68.0	74.5	77.4	80.8
Qwen3-14B	98.1	81.2	99.6	99.0	88.4	99.7	57.8	51.3	69.7	75.4	81.4	79.7
Qwen3-32B	98.7	80.5	99.8	99.8	90.0	99.9	62.5	54.4	76.1	75.9	84.8	81.4
GPT-OSS-20B	96.1	77.0	97.6	99.3	84.1	99.6	51.1	48.3	63.3	77.1	57.2	82.2
GPT-OSS-120B	98.6	78.2	99.6	99.9	90.0	100	66.2	63.8	74.2	88.5	86.0	91.1
GPT-5-Mini	99.0	81.1	99.6	99.9	89.2	99.9	73.2	64.5	80.5	89.0	88.1	91.0
GPT-5.2	99.5	82.8	100	100	90.6	100	80.2	68.6	87.0	92.8	96.1	93.8

Table 2: Results on CITECONTROL: Small models ($\leq$ 3B parameters) show citation failure even in simple cases, while all models fail in more complex cases (see §4.2). Rk / Rk^f: Recall @k on all instances / instances answered correctly. $\frac{|\mathcal{R}^f|}{|\mathcal{R}|}$: Proportion of correctly answered instances.

ferentiating between response failure and citation failure is important. In the following, we therefore focus on Rk^f scores.

Small models fail for `single` reasoning; all models fail for complex reasoning On the `single` reasoning datasets RepliQA and BoolQ-M, models with $\geq$3B parameters achieve near-perfect Rk^f ($>$ 95), while smaller models score lower. On MuSiQue and NeoQA, requiring `multi-hop` and `intersection` reasoning over multiple documents, all models show reduced Rk^f scores, indicating failure to identify evidence documents even though models correctly identify the necessary information for generating the response. Across models and tasks, there is a strong positive correlation between response scores[8] and Rk^f (r =0.745, p =0.000), showing that the response-evidence relation affects both response generation and citation.

[8] Answer F1 / Exact Match, data not shown.

LLM citations are (imperfectly) ordered by confidence Fig. 3(A) shows the precision of citations by their order of appearance in generation. It is visible that precision decreases from the first to the last citation for most models, suggesting that LLMs rank citations by confidence. The GPT models are notable exceptions, showing high precision on all generated citations.

Evidence recall decreases with moving up in the reasoning chain Figure 3(B) shows Rk^f by evidence position in the reasoning chain for MuSiQue and NeoQA (`multi-hop` instances only). Compared to hop 0 (`explicit` overtness), Rk^f drops strongly for earlier hops (`implicit` overtness), showing that models struggle to trace the reasoning chain during citation. The retrieval-based baselines `DRAG` and `BM25` are notable exceptions, with higher recall at hop -3 than hop -2. As retrieval-based models rely on both question and response to find evidence, their focus on the response is reduced. This explains their sub-optimal performance at hop 0 (which contains the re-

sponse verbatim) but elevated scores at earlier hops, where the question provides helpful signals (visible in elevated ROUGE scores between question and evidence in Table 7).

Implicitness alone does not entail citation failure The response-evidence relation is `implicit` in BoolQ-M, and `explicit` in RepliQA, which explains the reduced Rk^{f} for `BM25` on BoolQ-M. In contrast, most LLMs exhibit Rk^{f} scores close to 100 on BoolQ-M, showing that they are able to cite evidence in the absence of an `explicit` statement-evidence relation.

Explicitness can bias citation In Fig. 8, we show average Rk^{f} on `multi-hop` and `intersection` NeoQA instances, as well as Rk^{f} per hop. As expected, for most LLMs, Rk^{f} is highest on hop 0 evidence in `multi-hop` instances, likely due to the `explicit` response-evidence relation. Unexpectedly, Rk^{f} on hop -1 and average Rk^{f} on `multi-hop` instances is lower than average Rk^{f} on `intersection` instances for most models. This suggests that the `explicit` relation between the response and hop 0 evidence in `multi-hop` instances biases models to cite only hop 0 correctly, while the absence of this bias allows for better average citation performance on `intersection` instances. Again, `DRAG` and `BM25` are exceptions, as Rk^{f} is higher on `multi-hop` instances than on `intersection` instances.

Error analysis The two main citation failure modes are (a) generating too few citations and (b) generating wrong citations. Table 9 shows the number of instances with citation failure (Rk^{f} <1), along with: cases without any citations ($n_\emptyset$), the average difference between generated and correct citation counts (Diff), and precision @k (P@k). On the `single` reasoning datasets RepliQA and BoolQ-M, failure is mostly due to wrong citations, with $n_\emptyset$ and Diff near 0. On MuSiQue and NeoQA (Fig. 3 C), both failure modes appear: most models generate some citations ($n_\emptyset$ =0) but too few (Diff < 0), and precision decreases as the number of citations grows.

Data contamination has diverse effects on citation Data contamination—the presence of relevant information in pretraining data (Sainz et al., 2023; Li et al., 2024b)—could reduce reliance on contextual information and thus hurt citation performance, though this has not been studied. Fig. 3 (D) shows the difference in Rk^{f} between contaminated and uncontaminated instances for RepliQA, BoolQ-M, and MuSiQue.[9] On RepliQA and BoolQ-M, contamination has both beneficial and detrimental effects across models. On MuSiQue, the effect is clearly beneficial, contradicting our initial expectation.

Our analysis revealed that retrieval- and generation-based methods excel under different conditions. This suggests that combining citation methods can improve performance, which we will investigate in the following sections.

5 CITENTION: A Framework for Investigating Efficient LLM Citation

All models showed citation failure in our experiments, calling for mitigation. Attention values are a promising direction for this purpose: They have been exploited successfully in related tasks (Chen et al., 2025; Zhang et al., 2025b) and can be used efficiently, requiring the training of only a small number of parameters (or none at all) and not requiring additional generation. However, their potential for citation has not been studied.

To close this gap, we introduce CITENTION, our framework for studying the attention-based enhancement of generative citation. After giving an overview, we describe the used citation methods in §5.1 and their combination in §5.2.

Overview To enhance generative citation with other efficient citation methods, we assume individual citation methods $\mathrm{M}(r, s) \rightarrow \mathbb{R}$, that predict a *citation score* reflecting the relevance of document s as evidence for response r. Our results in §4.2 and research on model ensembles (Dietterich, 2000) suggest that combining scores can improve performance, so we experiment with method combination $\mathrm{M}^{\Omega} = \mathrm{Agg}(\mathrm{M}^1, ..., \mathrm{M}^{|\mathrm{M}^{\Omega}|})$, where $\mathrm{Agg}(\cdot)$ is an aggregation function and $|\mathrm{M}^{\Omega}|$ is the number of individual citation functions in M^{Ω}. Finally, we require a decision function $\delta(\mathrm{M}(r, s)) : \mathbb{R} \rightarrow \{0, 1\}$ that maps the citation score to a decision `no-cite` or `cite`.

[9] For details, see §A.4. We exclude NeoQA as contamination analysis is difficult given its multiple choice design.

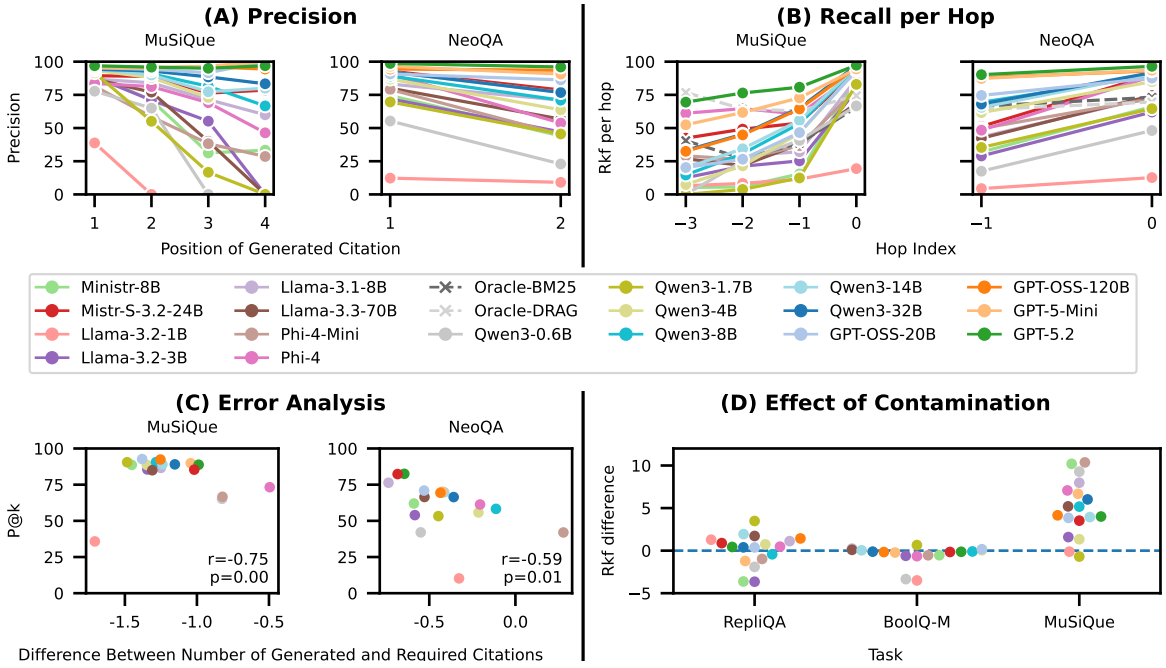

Figure 3: (A) Citation precision decreases with the order of appearance in generation, except for GPT models. (B) Rk^{f} is highest for the evidence document that contains the response (hop 0) and is reduced when going to earlier hops, DRAG and BM25 are notable exceptions. (C) Models tend to under-generate citations on MuSiQue and NeoQA. Approaching the correct number of citations is correlated with lower precision (r and p exclude Llama-3.2-1B). (D) Contamination has diverse effects on citation performance. Plot shows difference in Rk^{f} on contaminated / uncontaminated instances. See §4.2.

5.1 Citation Methods

To enhance generation-based citation, we consider the attention-based methods ICR (Chen et al., 2025), AT2 (Cohen-Wang et al., 2024) and QR (Zhang et al., 2025b). Retrieval-based methods, which have been applied successfully in related work on post-hoc citation (e.g. Bohnet et al. 2022; Sancheti et al. 2024), serve as efficient baselines. We introduce these methods below and refer the reader to the respective publications for details.

5.1.1 Generation-Based Citation

We obtain the citation score M^{Gen} as the length-normalized (Murray and Chiang, 2018) probability for generating the citation (see §A.6.1). For source documents without citations, the score is 0.

5.1.2 Attention-Based Citation

We focus on three recent attention-based methods: ICR, (Chen et al., 2025), QRHEAD (QR, Zhang et al. 2025b) and AT2 (Cohen-Wang et al., 2025). We first describe their general approach and then introduce the individual methods, adapting the notation from Cohen-Wang et al. (2025).

General approach To obtain citation scores, the attention-based methods work in two steps:

1. Compute a single score $\text{M}_d(r, s)$ per attention head h_d, where $d \in \{1, ..., |H|\}$ and $|H|$ is the number of attention heads (see §A.6.2).

2. Compute a weighted average over per-head scores: $\text{M}(r, s) = \sum_{d=1}^{|H|} \theta_d \text{M}_d(r, s)$.

The difference between the attention-based methods is in the way the weight vector θ is obtained:

ICR (Chen et al., 2025) puts equal weights on all attention heads,[10] so $\forall d : \theta_d^{\text{ICR}} = \frac{1}{|H|}$

QR (Zhang et al., 2025b) selects a subset of "query-focused retrieval heads" $H^{\text{QR}} \subset H$:

$$\theta_d^{\text{QR}} = \begin{cases} \frac{1}{|H^{\text{QR}}|} & \text{if } h_d \in H^{\text{QR}} \\ 0 & \text{otherwise} \end{cases}$$

H^{QR} is obtained as the heads that give the highest scores to relevant documents on a training set, where $|H^{\text{QR}}|$ is set based on model size.

[10] Chen et al. (2025) propose a calibration method for ICR that is also used in QR. Our preliminary experiments showed that it leads to decreased performance, so we are not using it.

AT2 (Cohen-Wang et al., 2025) learns a soft weighting θ^{AT2} such that the score for a source document s reflects the effect of removing s from the context: For a given training example, an LLM generates a continuation. Source documents are removed randomly from the context, the change in probability of the original generation is recorded, and θ^{AT2} is optimized with a correlation loss.

5.1.3 Retrieval-Based Citation

As in §4, we employ `BM25` (Robertson and Zaragoza, 2009) and `DRAG` (Lin et al., 2023).

5.2 Aggregation and Decision Functions

Aggregation To aggregate scores from different citation methods, we use a weighted average:

$$\mathrm{M}^{\Omega} = \sum_{i=1}^{|\mathrm{M}^{\Omega}|} w_i \mathrm{M}^{i} + b \tag{1}$$

This retains efficiency and avoids introducing confounders into our analysis. To learn w and b, we fit a linear model[11] on the train set scores from individual attention-based methods. We experiment with 3 combinations of scores:

- `COMB-A`: Generative and attention-based citation (`GEN`, `ICR`, `AT2` and `QR`)
- `COMB-R`: Generative and retrieval-based citation (`GEN`, `BM25` and `DRAG`)
- `COMB`: All CITENTION methods (`GEN`, `ICR`, `AT2`, `QR`, `BM25` and `DRAG`)

Decision Function As done in previous work, we predict evidence by selecting the k highest scoring source documents for a given response statement (Bohnet et al., 2022; Ramu et al., 2024). We choose k depending on dataset and instance as described in §A.3 and §6.1.2

6 How Can Attention Values be Used to Efficiently Mitigate LLM Citation Failure?

We study how to mitigate citation failures in generative citation using CITENTION. Prior work has examined generation-based citation (e.g., Gao et al. 2023; Tang et al. 2025) and retrieval-based citation (e.g., Bohnet et al. 2022; Ramu et al. 2024) separately, but we are the first to explore attention-based citation and to combine generation-, attention-, and retrieval-based approaches. To broaden the evaluation of CITENTION, we complement experiments on CITECONTROL with a transfer setting using two challenging datasets from a recent long-document attribution benchmark (Buchmann et al., 2024).

6.1 Experimental Setup

6.1.1 Methods

Training Where available, we train on the train splits of the CITECONTROL datasets, and train on the dev split of NeoQA. We train one set of parameters per combination of LLM and task. For θ^{QR}, we randomly choose 150 examples per dataset for selecting heads, and set $|H^{\text{QR}}|$ to 16 for models $\leq$8B parameters and else to 32 as in Zhang et al. (2025b). For θ^{AT2}, we train on all available examples and use the same hyperparameters as in Cohen-Wang et al. (2025). We train w and b for the combination methods on the train (dev) set scores of the individual CITENTION methods.

Masking Reasoning Tokens for Attention-Based Methods Except where otherwise noted, we evaluate attention-based methods with masked reasoning tokens, as this improved their performance in preliminary experiments (see 6.2.1).

6.1.2 Transfer

Datasets QASPER is a question-answering dataset on scientific articles proposed by Dasigi et al. (2021). We exclude unanswerable instances. GovReport is a summarization dataset on reports from US government agencies proposed by Huang et al. (2021). Besides requiring longer response statements than the datasets in CITECONTROL, these datasets represent a shift in source document type (incoherent collection of paragraphs → coherent document) and for QASPER in domain (Wikipedia / General → scientific).

Evaluation QASPER and GovReport are datasets with free-form responses and incomplete evidence annotations, requiring more flexible evaluation than Rk/Rk^{f}. Therefore, we employ the attributability evaluation models Minicheck (Tang et al., 2024) for QASPER and TRUE (Honovich et al., 2022) for GovReport, which have been shown to obtain >75% accuracy on these datasets (Buchmann et al., 2024). For QASPER, we treat LLM responses as a single statement, while we split responses for GovReport by

[11] Using a LinearModel from scikit-learn (Pedregosa et al., 2011)

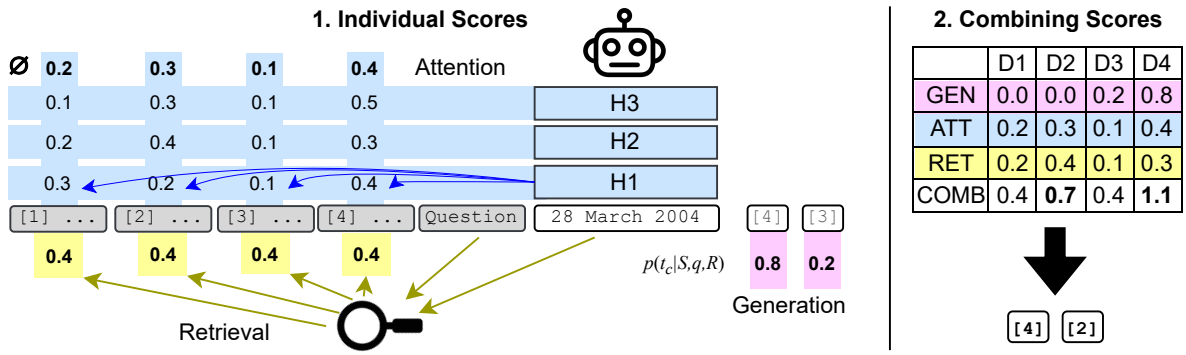

Figure 4: Overview of CITENTION (§5). Left: Individual scores for each document are obtained from generation-based, attention-based and retrieval-based methods. Attention scores are averaged over individual heads. Right: Scores from individual methods are summed to obtain a final citation prediction. Attention head weights θ and combination weights w, b are omitted.

sentence using spacy.[12] We report the proportion of response statements evaluated as attributable to the 2 highest-scoring source documents ($k = 2$).

CITENTION Parameters Based on preliminary experiments, we use AT2 and QR parameters trained on MuSiQue for QASPER, and parameters trained on RepliQA for GovReport. We retrain w and b score combination parameters based on attributability scores.

6.2 Results

We evaluated the performance of the CITENTION methods on CITECONTROL and the transfer datasets, using Llama-3.2-1B, Llama-3.1-8B, Llama-3.3-70B, Qwen3-1.7B, Qwen3-8B and Qwen3-32B, showing results in Fig. 5 and Tab. 10. We first analyze the performance of attention-based citation in comparison to generation-based citation and the retrieval baselines, and then the performance of method combination.

6.2.1 Results of Isolated Citation Methods

Attention-based and retrieval-based citation can mitigate citation failure Attention-based and retrieval-based methods achieve higher average scores than generation-based citation (GEN) for the Llama models and Qwen3-1.7B on CITECONTROL (DRAG improves over GEN by at least 6 points), and for all models on the transfer datasets (QR improves over GEN by at least 5 points). This agrees with Ramu et al. (2024) and Sancheti et al. (2024), who found that retrieval-based methods can improve over LLMs in post-hoc citation, but did not investigate attention-based methods. In contrast, for Qwen3-8B and Qwen3-32B, GEN is stronger, as the average Rk^{f} scores are more than 10 points above the average scores from GEN on the other models, and consistently higher than the scores from the other isolated citation methods.

Retrieval improves over citation on CITECONTROL, attention-based methods are better on transfer datasets On CITECONTROL, the average Rk^{f} for DRAG is mostly higher than for attention-based methods. We attribute this to the fact that only the retrieval-based citation methods have access to the question, which is helpful in finding evidence with implicit statement-evidence relation. This is visible in the ROUGE scores in Table 7: There is consistent lexical overlap between the question and evidence for all datasets, while this is not the case for the response and the evidence. On the other hand, QR and AT2 exhibit the highest scores on the transfer datasets. This suggests that the more long-form responses required for QASPER and GovReport enable the attention methods to use the LLM-internal information more effectively, giving them an advantage over retrieval-based methods that do not have access to this information.

Attention-based methods work better for Llama models than for Qwen models We observe higher Rk^{f} and attributability scores for attention-based methods on Llama models than on Qwen models. This indicates that attention-based citation is sensitive to model architecture and / or pretraining regime.

Attention-based methods are sensitive to overtness The sensitivity of attention-based meth-

[12] https://github.com/explosion/spaCy

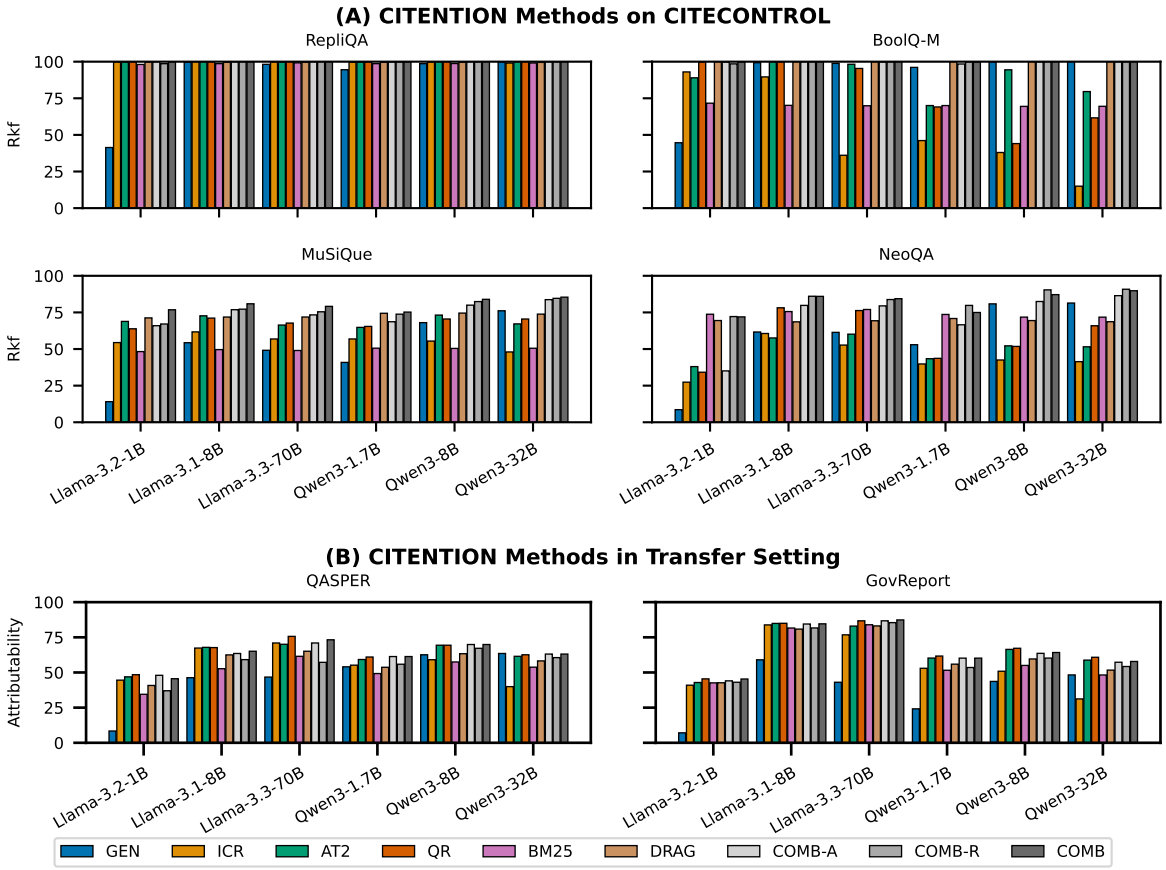

Figure 5: (A) CITENTION methods improve citation Rk^{f} scores on CITECONTROL. (B) CITENTION methods trained on CITECONTROL improve Rk^{f} scores on transfer tasks. Bars show proportion of answer statements that are attributable to the evidence. For numerical data see Table 10.

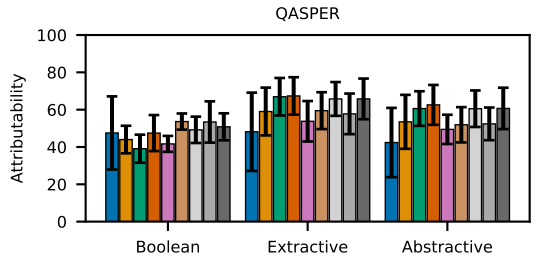

Figure 6: Rk^{f} by question type on QASPER, averaged over models. Whiskers show standard deviation. See Fig. 5 for color legend.

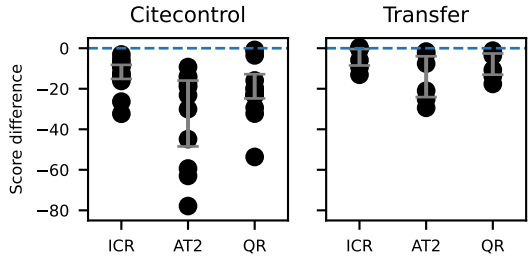

Figure 7: Attention-based citation suffers from reasoning tokens in context. Plots show difference between Rk^{f} (CITECONTROL) / attributability (transfer) for attention based methods on Qwen models evaluated with and without reasoning tokens in context. Whiskers show interquartile range

ods to overtness can be seen in multiple places. First, while Rk^{f} scores on RepliQA (`explicit` statement-evidence relation) are consistently high for attention-based methods, they are reduced on BoolQ-M (`implicit`, Fig. 5A). Similarly, attributability for extractive and abstractive questions is higher than for boolean questions on QASPER (Fig. 6). Finally the performance of attention-based methods decreases when going from the evidence document that contains the response to the ones for earlier hops on MuSiQue and NeoQA (Fig. 10).

`AT2` and `QR` provide better citations than `ICR` In almost all combinations of model and task, `QR` and `AT2` achieve higher scores than `ICR`. We ex-

plain this with the fact that they were optimized on the respective datasets in CITECONTROL, while ICR was not.

Attention-based citation is distracted by reasoning tokens Fig. 7 shows the difference in Rk^{f} / attributability between Qwen models evaluated with reasoning tokens in context and the same model evaluated with masked reasoning tokens. The negative values show that reasoning tokens hurt performance. All other plots and tables therefore only show scores from evaluation with masked reasoning tokens.

6.2.2 Results of Combined Methods

Combining citation methods reliably mitigates citation failure. The average citation score on CITECONTROL and transfer datasets is highest for all models when combining generation-based, attention-based and retrieval-based citation (COMB). For all combinations of models and datasets except Qwen3-32B on GovReport, Rk^{f} / attributability is higher for COMB than for generation-based citation. For all models on CITECONTROL, average Rk^{f} is higher when combining retrieval-based methods (COMB−R) than when combining attention-based methods (COMB−A), while it reversed on the transfer datasets. This reflects the performance of the isolated citation methods discussed above.

Attention-based and retrieval-based citation can complement each other The fact that the combination of citation methods improves over isolated methods suggests that they complement each other. This is especially visible in the recall of evidence by its position in the reasoning chain for Qwen3-8B and Qwen3-32B on MuSiQue (Fig. 10). On hop 0, where the response-evidence relation is explicit, we can observe the strongest improvements from attention-based citation. Going towards earlier hops (hops -1, -2, -3) with implicit statement-evidence relation, we can observe that DRAG exhibits the strongest performance, exemplifying how the different citation methods complement each other.

Score combination requires complementary methods and is responsible to the response-evidence relation While the combination of all citation methods (COMB) exhibits the highest average scores overall, it does not improve over individual citation methods on transfer datasets. To get more insights into the potential benefits of score combination, we propose to compute *Oracle Improvement Bound* (OIB), which reflects the proportion of instances where the predicted citations from a weaker citation method can complement those from a stronger citation method (§A.7). As visible in Fig. 9, on GovReport, the highest observable OIB is considerably lower than on the other datasets except for RepliQA, where all methods obtain almost perfect scores. This indicates that there is not much complementary potential between citation methods on GovReport, and thus it is difficult for the combined methods to improve over isolated methods.

While OIB values on QASPER are higher than on GovReport, score combination does not always improve over isolated methods. As noted in §6.2.1, the performance of the isolated citation methods varies between different response-evidence relation types on QASPER. As the score combination weights are constant, they cannot make ideal use of the citation methods depending on the relation type. This opens an exciting direction for research on improving the combination of citation methods.

7 Conclusion

We defined citation failure and presented key findings for understanding and mitigating it. Using CITECONTROL, we showed that citation failure is frequent, especially with complex statement-evidence relations. With CITENTION, our results show that attention-based citation can mitigate failure from generative citation in some settings, suggesting that LLMs encode more than they generate, and making attention-based citation a promising research direction. More broadly, our analysis of attention-based citation can inform research on attention-based reranking.

Further, we demonstrated that citation failure can be reliably mitigated by combining generative, retrieval-based, and attention-based citation methods. Our work paves the way for practical, efficient citation methods, particularly in resource-constrained settings, and motivates future research on more sophisticated approaches to combine citation methods.

Acknowledgements

This work has been funded by the LOEWE Distinguished Chair "Ubiquitous Knowledge Process-

ing", LOEWE initiative, Hesse, Germany (Grant Number: LOEWE/4a//519/05/00.002(0002)/81) and the European Union (ERC, InterText, 101054961). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. We express our sincere gratitude to the reviewers and action editor at TACL for their valuable and constructive feedback.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 technical report. *ArXiv preprint*, abs/2412.08905.

Amin Abolghasemi, Leif Azzopardi, Seyyed Hadi Hashemi, Maarten de Rijke, and Suzan Verberne. 2025. Evaluation of attribution bias in generator-aware retrieval-augmented large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21105–21124, Vienna, Austria. Association for Computational Linguistics.

Sriram Balasubramanian, Samyadeep Basu, Koustava Goswami, Ryan A. Rossi, Varun Manjunatha, Roshan Santhosh, Ruiyi Zhang, Soheil Feizi, and Nedim Lipka. 2026. Decomposition-enhanced training for post-hoc attributions in language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5070–5084, Rabat, Morocco. Association for Computational Linguistics.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John Patrick Cunningham. 2024. LoRA learns less and forgets less. *Transactions on Machine Learning Research*. Featured Certification.

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, et al. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *ArXiv preprint*, abs/2212.08037.

Jan Buchmann, Xiao Liu, and Iryna Gurevych. 2024. Attribute or abstain: Large language models as long document assistants. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8113–8140, Miami, Florida, USA. Association for Computational Linguistics.

Shijie Chen, Bernal Jimenez Gutierrez, and Yu Su. 2025. Attention in large language models yields efficient zero-shot re-rankers. In *The Thirteenth International Conference on Learning Representations*.

Yung-Sung Chuang, Benjamin Cohen-Wang, Zejiang Shen, Zhaofeng Wu, Hu Xu, Xi Victoria Lin, James R. Glass, Shang-Wen Li, and Wen-Tau Yih. 2025. SelfCite: Self-supervised alignment for context attribution in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 10839–10858. PMLR.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Benjamin Cohen-Wang, Yung-Sung Chuang, and Aleksander Madry. 2025. Learning to attribute with attention. *ArXiv preprint*, abs/2504.13752.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. 2024. Con-

textcite: Attributing model generation to context. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Tri Dao. 2024. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics.

Thomas G. Dietterich. 2000. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Qiang Ding, Lvzhou Luo, Yixuan Cao, and Ping Luo. 2025. Attention with dependency parsing augmentation for fine-grained attribution. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 372–387, Vienna, Austria. Association for Computational Linguistics.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhiheng Xi, Xiaoran Fan, Shiliang Pu, Jiang Zhu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. LoRAMoE: Alleviating world knowledge forgetting in large language models via MoE-style plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1932–1945, Bangkok, Thailand. Association for Computational Linguistics.

Benedikt Droste, Jan Philipp Harries, Maximilian Idahl, and Björn Plüster. 2026. sui-1: Grounded and verifiable long-form summarization. *ArXiv Preprint*, abs/2601.08472.

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

Max Glockner, Xiang Jiang, Leonardo F. R. Ribeiro, Iryna Gurevych, and Markus Dreyer. 2025. NeoQA: Evidence-based question answering with generated news events. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11842–11926, Vienna, Austria. Association for Computational Linguistics.

Lotem Golany, Filippo Galgani, Maya Mamo, Nimrod Parasol, Omer Vandsburger, Nadav Bar, and Ido Dagan. 2024. Efficient data generation for source-grounded information-seeking dialogs: A use case for meeting transcripts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1908–1925, Miami, Florida, USA. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo

Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya,

Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *ArXiv preprint*, abs/2407.21783.

Eran Hirsch, Aviv Slobodkin, David Wan, Elias Stengel-Eskin, Mohit Bansal, and Ido Dagan. 2025. LAQuer: Localized attribution queries in content-grounded generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15355–15370, Vienna, Austria. Association for Computational Linguistics.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, Dublin, Ireland. Association for Computational Linguistics.

Nan Hu, Jiaoyan Chen, Yike Wu, Guilin Qi, Hongru Wang, Sheng Bi, Yongrui Chen, Tongtong Wu, and Jeff Z. Pan. 2025. Can LLMs evaluate complex attribution in QA? automatic benchmarking using knowledge graphs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17096–17118, Vienna, Austria. Association for Computational Linguistics.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuxuan Gu, Weihong Zhong, Xiachong Feng, Weijiang Yu, Weihua Peng, Duyu Tang, Dandan

Tu, and Bing Qin. 2024. Learning fine-grained grounded citations for attributed large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14095–14113, Bangkok, Thailand. Association for Computational Linguistics.

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

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution. *ArXiv preprint*, abs/2307.16883.

Seonmin Koo, Jinsung Kim, YoungJoon Jang, Chanjun Park, and Heuiseok Lim. 2024. Where am I? large language models wandering between semantics and structures in long contexts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14144–14160, Miami, Florida, USA. Association for Computational Linguistics.

Ilia Kuznetsov, Jan Buchmann, Max Eichler, and Iryna Gurevych. 2022. Revise and resubmit: An intertextual model of text-based collaboration in peer review. *Computational Linguistics*, 48(4):949–986.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wentau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Dongfang Li, Zetian Sun, Baotian Hu, Zhenyu Liu, Xinshuo Hu, Xuebo Liu, and Min Zhang. 2024a. Improving attributed text generation of large language models via preference learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5079–5101, Bangkok, Thailand. Association for Computational Linguistics.

Yucheng Li, Yunhao Guo, Frank Guerin, and Chenghua Lin. 2024b. An open-source data contamination report for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 528–541, Miami, Florida, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023. How to train your dragon: Diverse augmentation towards generalizable dense retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6385–6400, Singapore. Association for Computational Linguistics.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Pa-

tricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. 2026. Ministral 3. *ArXiv Preprint*, abs/2601.08584.

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

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3025–3045, Mexico City, Mexico. Association for Computational Linguistics.

João Monteiro, Pierre-André Noël, Étienne Marcotte, Sai Rajeswar Mudumba, Valentina Zantedeschi, David Vázquez, Nicolas Chapados, Chris Pal, and Perouz Taslakian. 2024. Repliqa: A question-answering dataset for benchmarking llms on unseen reference content. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223, Brussels, Belgium. Association for Computational Linguistics.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. 2025. gpt-oss-120b & gpt-oss-20b model card. *ArXiv preprint*, abs/2508.10925.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Van-

derplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Vinzent Penzkofer and Timo Baumann. 2024. Evaluating and fine-tuning retrieval-augmented language models to generate text with accurate citations. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 57–64, Vienna, Austria. Association for Computational Linguistics.

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

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

Hongjin Qian, Zheng Liu, Kelong Mao, Yujia Zhou, and Zhicheng Dou. 2024. Grounding language model with chunking-free in-context retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1298–1311, Bangkok, Thailand. Association for Computational Linguistics.

Pritika Ramu, Koustava Goswami, Apoorv Saxena, and Balaji Vasan Srinivasan. 2024. Enhancing post-hoc attributions in long document comprehension via coarse grained answer decomposition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17790–17806, Miami, Florida, USA. Association for Computational Linguistics.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Abhilasha Sancheti, Koustava Goswami, and Balaji Srinivasan. 2024. Post-hoc answer attribution for grounded and trustworthy long document comprehension: Task, insights, and challenges. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 49–57, Mexico City, Mexico. Association for Computational Linguistics.

Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. 2024. Attribute first, then generate: Locally-attributable grounded text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3344, Bangkok, Thailand. Association for Computational Linguistics.

Ionut Teodor Sorodoc, Leonardo F. R. Ribeiro, Rexhina Blloshmi, Christopher Davis, and Adrià de Gispert. 2025. GaRAGe: A benchmark with grounding annotations for RAG evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17030–17049, Vienna, Austria. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Zecheng Tang, Keyan Zhou, Juntao Li, Baibei Ji, Jianye Hou, and Min Zhang. 2025. *L-CiteEval*: A suite for evaluating fidelity of long-context models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5254–5277, Vienna, Austria. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Dustin Wright, Zain Muhammad Mujahid, Lu Wang, Isabelle Augenstein, and David Jurgens. 2025. Unstructured evidence attribution for long context query focused summarization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1839–1867, Suzhou, China. Association for Computational Linguistics.

Junjie Wu, Gefei Gu, Yanan Zheng, Dit-Yan Yeung, and Arman Cohan. 2025. Ref-long: Benchmarking the long-context referencing capability of long-context language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23861–23880, Vienna, Austria. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *ArXiv preprint*, abs/2505.09388.

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

Wuwei Zhang, Fangcong Yin, Howard Yen, Danqi Chen, and Xi Ye. 2025b. Query-focused retrieval heads improve long-context reasoning and re-ranking. In *Proceedings of the 2025*

Conference on Empirical Methods in Natural Language Processing, pages 23791–23805, Suzhou, China. Association for Computational Linguistics.

Junnan Zhu, Min Xiao, Yining Wang, Feifei Zhai, Yu Zhou, and Chengqing Zong. 2025. TROVE: A challenge for fine-grained text provenance via source sentence tracing and relationship classification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11755–11771, Vienna, Austria. Association for Computational Linguistics.

A Replication Information

A.1 Prompts

We use a 3-shot prompt per dataset with diverse in-context examples and shortened source documents. Each source document $s_i \in S$ is prepended by its index in square brackets (Gao et al., 2023), with questions placed after the source documents (Buchmann et al., 2024). Task and format explanations are in Tables 3 and 4.

```
<user_input_start>
# Task Explanation
Task: {task_explanation}

# Format Explanation
Follow this example for answer formatting:
{format_explanation}

# 3-shot examples omitted

<user_input_start>
Retrieved Paragraphs: [0] {document_0}
[1] {document_1}
...

Question: {question}

<assistant_input_start>
Answer:_
```

A.2 Data Processing

RepliQA (Monteiro et al., 2024) is a dataset of human-written multi-paragraph documents, questions, long answers and short answers. We only use questions where both answers appear verbatim in the context document. We split the context documents split into paragraphs and sampled one answer-containing paragraph as evidence. We sampled 19 distractor paragraphs from the same document (non-answer paragraphs) and supplemented them with paragraphs from other documents if needed.

BoolQ-M was created in a two-step modificiation process of BoolQ (Clark et al., 2019), which is a dataset of context passages, questions and yes/no answers: (1) **Filtering**: GPT-OSS-120B selected instances where the question concerns a concrete, identifiable entity that appears in the passage (excluding questions about abstract concepts, laws, or definitions). (2) **Entity replacement**: GPT-OSS-120B replaced entities in suitable instances with fictional ones, being instructed to preserve the question and passage structure and style, logic, and original truth value, and to not use replacements from existing fictional universes. Distractor source documents for BoolQ / BoolQ-M are 19 randomly sampled paragraphs from other instances in the same split.

MuSiQue (Trivedi et al., 2022) is a dataset of questions, answers, 2 – 4 relevant context paragraphs and 16 – 18 distractor paragraphs, questions and answers, which we used as is.

NeoQA (Glockner et al., 2025) is a dataset of fictional news articles about one of several event timelines, questions, and 6 answer options (one of them true). To assemble source documents for a specific question, we sampled 2 news articles that contain the relevant information pieces as evidence, and sampled 18 news articles from the same timeline that do not contain the relevant information pieces as distractors.

Dataset	task_explanation
RepliQA	You are given a question and a list of retrieved paragraphs, which might contain relevant information to the question. Answer the Question using only the information from the retrieved paragraphs. Your answer should be concise and not more than a single phrase. Do not provide any explanation. Provide the paragraph that can be used to verify the answer by writing the integer id in square brackets.
BoolQ-M	You are given a yes/no question and a list of retrieved paragraphs, which might contain relevant information to the question. Answer the Question using only the information from the retrieved paragraphs. Your answer should be "yes" or "no". Do not provide any explanation. Provide the paragraph that can be used to verify the answer by writing the integer id in square brackets.
MuSiQue	You are given a question and a list of retrieved paragraphs, which might contain relevant information to the question. Answer the Question using only the information from the retrieved paragraphs. Your answer should be concise and not more than a single phrase. Do not provide any explanation. Provide all paragraphs that are needed to verify the answer by writing the integer id in square brackets.
NeoQA	You are given a list of retrieved news articles, a question and 6 answer options. The news articles might contain relevant information to the question. Answer the Question by responding with one of the answer options. Your answer should be exactly the same as one of the answer options. Do not provide any explanation. Provide the ids of the news articles that information needed to answer the question by writing the integer id in square brackets.
QASPER	You are given a Scientific Article and a Question. Answer the Question as concisely as you can, using a single phrase or sentence. If the question is a yes/no question, your answer should be "yes" or "no". Do not provide any explanation. Provide the paragraphs that can be used to verify the answer by writing their integer ids in square brackets. Put each id in separate brackets. Always provide ids of content paragraphs, not section headlines.
GovReport	Task: You are given a government report document. Write a one page summary (max. 15 sentences) of the document. Each sentence in your summary should reference the source paragraphs from the document that can be used to verify the summary sentence. There should be no sentences without references. Always put the references after the summary sentence.

Table 3: Task explanations used in the prompts in our experiments.

Dataset splits RepliQA was sequentially released in 5 splits between June 2024 and June 2025[15]. We use splits 0–2 for training, split 3 for development and the latest split 4 for testing. We use the development splits of, BoolQ / BoolQ-M and MuSiQue for evaluation, as their test splits are hidden. We use the test splits of NeoQA, QASPER and GovReport (CRS subset).

A.3 Filtered evaluation ($\mathbf{R}k^{\mathrm{f}}$)

To perform filtered evaluation, we consider responses with a response evaluation score >0.7. To evaluate response correctness, we use token F1 score for RepliQA and MuSiQue, and exact match for BoolQ, BoolQ-M and NeoQA, as done in the respective original dataset papers. We set $k = |E^*| + 1$, i.e. one larger than the size of the ground truth evidence set for a particular instance. We assume the order of generated citations as their ranking from highest to lowest.

A.4 Comparison between contaminated and uncontaminated instances

For RepliQA and MuSiQue, we determine the contamination status of test set questions for a specific model by prompting without source documents. If the answer F1 score is (a) above 0.7 (b) below 0.3 (c)

[15] https://github.com/ServiceNow/repliqa

Dataset	format_explanation
RepliQA	Retrieved Paragraphs: <omitted> Question: When did Beyonce start becoming popular? Answer: in the late 1990s [7]
BoolQ /BoolQ-M	Retrieved Paragraphs: <omitted> Question: is axl rose the new lead singer of acdc Answer: yes
MuSiQue	Retrieved Paragraphs: <omitted> Question: When was the institute that owned The Collegian founded? Answer: 1960 [5] [9]
NeoQA	News Articles: <omitted> Question: What is the duration between the date when Crestfield Property Holdings shared the preliminary findings of the ZentroTek Solutions review (assumed to be shared by the end of January) and the date when Everstead Technical Systems discovered the calibration issue affecting the surveillance cameras? Answer options: a) 21 days b) 35 days c) 30 days d) 31 days e) 28 days f) 14 days Answer: 28 days [4] [7]
QASPER	Scientific Article: <omitted> Question: Which baselines were used? Answer: BERT, RoBERTa [7] [8]
GovReport	Report: <omitted> Under the Arms Export Control Act and its implementing regulations, DOD is required to recover nonrecurring costs—unique one-time program-wide expenditures—for certain major defense equipment sold under the FMS program. [2] [4] These costs include research, development, and one-time production costs, such as expenses for testing equipment. [3]

Table 4: Format explanations used in the prompts in our experiments.

Purpose	Package
Base for CITENTION	AT2 (Cohen-Wang et al., 2025)
Generation	Huggingface Transformers (Wolf et al., 2020)
BM25 retrieval	Rank-BM25[13]
Dense retrieval	Sentence Transformers (Reimers and Gurevych, 2019)
Aggregation weight fitting	Scikit-Learn (Pedregosa et al., 2011)
ROUGE score computation	Rouge-Score[14]

Table 5: Python packages used in experiments.

in between, the question is (a) seen as contaminated (b) uncontaminated (c) discarded for this analysis. As the chance for randomly answering a BoolQ-M question correctly is 50%, determining the contamination status of individual questions is difficult. Instead, we assume all original BoolQ questions as contaminated, and our newly created BoolQ-M as uncontaminated. This is verified in Table 8, which shows that performance on BoolQ-M without context is strongly reduced and below random chance, while it is above random chance and higher for BoolQ without context.

Model	#Params(Active)	Reasoning Type	Citation	Huggingface Tag
Ministr-8B	8B	No Reasoning	Liu et al. 2026	`mistralai/Ministral-8B-Instruct-2410`
Mistr-S-3.2-24B	24B	No Reasoning	Huggingface	`mistralai/Mistral-Small-3.2-24B-Instruct-2506`
Llama-3.2-1B	1B	No Reasoning	Grattafiori et al. 2024	`meta-llama/Llama-3.2-1B-Instruct`
Llama-3.2-3B	3B	No Reasoning	Grattafiori et al. 2024	`meta-llama/Llama-3.2-3B-Instruct`
Llama-3.1-8B	8B	No Reasoning	Grattafiori et al. 2024	`meta-llama/Llama-3.1-8B-Instruct`
Llama-3.3-70B	70B	No Reasoning	Grattafiori et al. 2024	`meta-llama/Llama-3.3-70B-Instruct`
Phi-4-Mini	4B	No Reasoning	Abdin et al. 2024	`microsoft/phi-4-mini-instruct`
Phi-4	14B	No Reasoning	Abdin et al. 2024	`microsoft/phi-4`
Qwen3-0.6B	0.6B	Reasoning	Yang et al. 2025	`Qwen/Qwen3-0.6B`
Qwen3-1.7B	1.7B	Reasoning	Yang et al. 2025	`Qwen/Qwen3-1.7B`
Qwen3-4B	4B	Reasoning	Yang et al. 2025	`Qwen/Qwen3-4B`
Qwen3-8B	8B	Reasoning	Yang et al. 2025	`Qwen/Qwen3-8B`
Qwen3-14B	14B	Reasoning	Yang et al. 2025	`Qwen/Qwen3-14B`
Qwen3-32B	32B	Reasoning	Yang et al. 2025	`Qwen/Qwen3-32B`
GPT-OSS-20B	20B (3.6B)	Reasoning	OpenAI et al. 2025	`openai/gpt-oss-20b`
GPT-OSS-120B	120B (3.6B)	Reasoning	OpenAI et al. 2025	`openai/gpt-oss-120b`
GPT-5-mini	N/A	Reasoning	OpenAI Website	N/A
GPT-5.2	N/A	Reasoning	OpenAI Website	N/A

Table 6: The LLMs employed in this work.

A.5 Employed LLMs

A.6 Details of Citation Methods in CITENTION

A.6.1 Generation-Based Citation

For generation-based citation, we obtain the citation score for source document s_j as the length-normalized (Murray and Chiang, 2018) probability for generating the citation tokens $c = \{t_1^c ... t_{|c|}^c\}$ that point to s_j (e.g. "`[4]`").

$$\mathrm{M}^{\mathrm{Gen}}(r, s) = \exp\left(\frac{1}{|c|}\sum_{i=1}^{|c|}\log\left(\frac{\exp(\ell_t[t_i^c])}{\sum_{v=1}^{V}\exp(\ell_t[v])}\right)\right) \tag{2}$$

which is equivalent to the geometric mean of the token probabilities:

$$\mathrm{M}^{\mathrm{Gen}}(r, s) = \left(\prod_{i=1}^{|c|} p(t_i^c \mid x, t_{<i}^c)\right)^{1/|c|} \tag{3}$$

A.6.2 Attention-Based Citation

Computing per-head attention scores The per-head attention score $\mathrm{M}_d(r, s)$, for query r consisting of tokens $t_1^r ... t_{|r|}^r$ and source document s consisting of tokens $t_1^s ... t_{|s|}^s$ is obtained as

$$\mathrm{M}_d(r, s) = \frac{1}{|r|}\sum_{i=1}^{|r|}\sum_{j=1}^{|s|}\mathrm{ATT}_d(t_i^r, t_j^s) \tag{4}$$

$\mathrm{ATT}_d(t_i, t_j)$ is the softmax-normalized attention score from token i to token j in head h_d.

A.6.3 Retrieval-Based Citation

`BM25` `BM25` (Robertson and Zaragoza, 2009) computes relevance scores by computing the token overlap between query and document, and weighting overlapping tokens according to their frequency of occurence in a training corpus. While computationally simple, it is still considered a competitive retrieval baseline (Thakur et al., 2021). We compute token frequency statistics on the train sets of the CITECONTROL tasks[16] and set hyperparameters to common values `k1`$= 1.5$; `b`$= 0.75$ (Robertson and Zaragoza, 2009).

[16] We use the dev set of NeoQA as it does not have a train set.

DRAG Dragon (Lin et al., 2023) is based on a dual transformer-encoder architecture and was trained with a mixture of data augmentation techniques. Relevance scores are computed as the dot product of the query and document vector representations. We leave the parameters unchanged, as it has been optimized for zero-shot retrieval.

A.7 Technical Details

Mapping evidence to hops in NeoQA To perform the analysis of recall per hop for NeoQA (Figs. 3 and 8), a mapping between the evidence documents and the hop index is needed. To obtain this mapping, we ordered the evidence documents by lexical overlap (ROUGE-1) with the ground truth answer. The document with the higher ROUGE-1 was used as the evidence for hop 0, while the document with the lower ROUGE-1 was used as the evidence for hop -1.

Oracle Improvement Bound On the recall-focused CITECONTROL tasks, the highest possible recall from a combination of two citation methods M^a and M^b with predicted citation sets E_i^a and E_i^b on instance i can be obtained by replacing incorrect predictions of one method with correct predictions from the other, assuming oracle knowledge of the ground truth and a fixed maximum number of citations. Let $E_i^{b\to a}$ denote the set of additional correct citations that method a could obtain from b, and let E_i^* be the set of ground-truth citations.

The recall-based directional potential improvement ratio for improving a using b is defined as

$$I_{b\to a}^R = \frac{\sum_i |E_i^{b\to a}|}{\sum_i |E_i^*|}.$$

Analogously, $I_{a\to b}$ is defined by exchanging the roles of a and b.

For the transfer tasks, we do not have knowledge of ground truth citations. Thus, we set

$$E_i^{b\to a}$$

to 1 if and only if the response statement ri is attributable to E_i^b, but not to E_i^a, and vice versa. Then we compute

$$I_{b\to a}^{\mathrm{Att}} = \frac{\sum_i |E_i^{b\to a}|}{|\mathcal{R}|}.$$

Since the two methods may differ in their baseline recall, these directional improvement ratios are generally asymmetric. When estimating the maximal achievable improvement of the better-performing method under a fixed prediction budget, only the smaller of the two ratios constitutes a valid upper bound. The larger ratio corresponds to improving the weaker method and therefore overestimates the achievable gain when starting from the stronger system. We therefore define the Oracle Improvement Bound (OIB) as

$$\mathrm{OIB} = \min\left(I_{b\to a},\, I_{a\to b}\right).$$

Used packages See Tab. 5

A.8 Limitations

Analysis of Source document composition Several works analyze the effect of source document composition on citation performance (§2), which we view as complementary to our response-evidence analysis. Due to the large number of experiments, we fixed source documents to one set per instance. Future work could investigate interactions between both dimensions using CITECONTROL as a starting point.

Score combination method We use weighted averages to combine citation scores, which is efficient and avoids confounders. While we showed this improves over isolated citation methods, it represents a lower bound for score combination. More sophisticated combination methods could further improve performance, and we hope future work builds on CITENTION to explore this direction.

Closed source models Attention-based citation requires access to model internals, precluding closed-source models. While our extensive experiments demonstrate that attention-based citation can efficiently enhance citation performance, future analysis on closed-source models would be valuable.

B Additional Results

		Answer		Question	
		Evi	No Evi	Evi	No Evi
RepliQA	-	1.00	0.18	0.75	0.31
BoolQ-M	-	0.03	0.02	0.72	0.27
MuSiQue	Multi-Hop	0.07/0.08/0.12/0.97	0.10	0.36/0.40/0.46/0.44	0.43
NeoQA	Multi-Hop	0.48/0.96	0.43	0.71/0.65	0.50
	Intersection	0.17	0.14	0.67	0.54
	Boolean	0.01	0.00	0.41	0.18
QASPER	Extractive	0.85	0.14	0.44	0.20
	Abstractive	0.63	0.17	0.43	0.20

Table 7: Lexical overlap between ground truth answers / questions and evidence / no evidence source documents. Numbers show ROUGE-1 recall for tokens from answers / questions in evidence documents. For `multi-hop` instances, we show values per evidence document for hop .../-1/0.

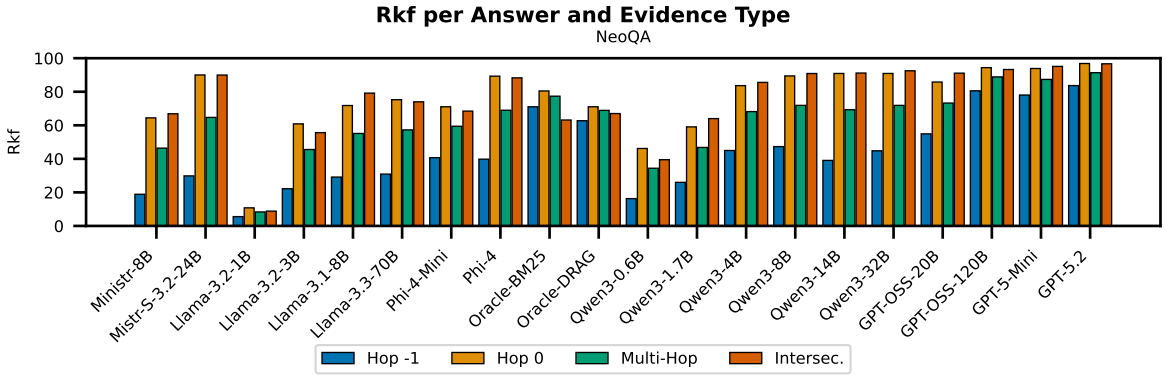

Figure 8: Detailed results for NeoQA. Hop -1/0: Rk^{f} per hop on `multi-hop` instances. `multi-hop` / `intersection`: average Rk^{f} for the respective instance type. For analysis and discussion see §4.

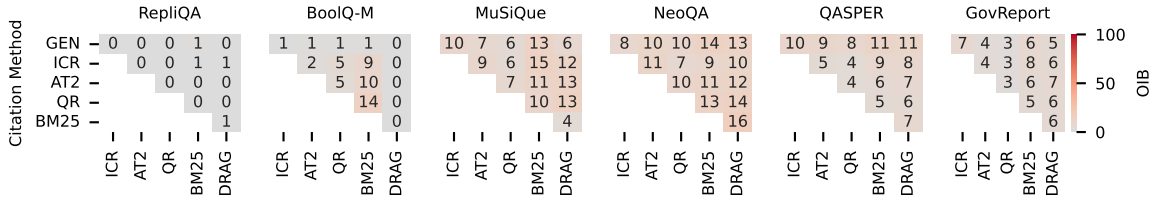

Figure 9: Oracle improvement bounds (OIB) averaged over models. See §6.2.2 and §A.7 for details.

	RepliQA		**BoolQ**		**BoolQ-M**		**MuSiQue**		**NeoQA**	
	C	NC	C	NC	C	NC	C	NC	C	NC
Llama-3.3-70B	86.2	15.3	76.5	62.6	78.4	43.4	37.9	12.0	52.8	33.7
Qwen3-32B	87.0	17.7	90.5	75.2	90.0	41.0	61.8	17.9	84.8	10.4
GPT-OSS-120B	85.3	18.5	91.6	55.4	90.0	30.4	70.2	24.9	86.0	38.5
GPT-5-Mini	87.6	14.4	89.4	75.3	89.2	44.1	71.8	34.9	88.1	47.0
GPT-5.2	88.2	16.0	91.3	83.3	90.6	45.2	75.8	36.0	96.1	46.7

Table 8: Data contamination on CITECONTROL: Response performance with (C) and without context (NC). Metrics: RepliQA, MuSiQue: Answer F1. BoolQ, BoolQ-M, NeoQA: Accuracy

	RepliQA			BoolQ-M			MuSiQue				NeoQA			
	n	$n_\emptyset$	Diff	n	$n_\emptyset$	Diff	n	$n_\emptyset$	Diff	P@k	n	$n_\emptyset$	Diff	P@k
Ministr-8B	26	0	0.0	3	0	0.0	597	1	-1.5	88.7	421	2	-0.6	62.0
Mistr-S-3.2-24B	1	0	0.0	1	1	-1.0	749	0	-1.0	85.4	427	0	-0.7	82.3
Llama-3.2-1B	772	98	-0.1	703	0	0.0	278	20	-1.7	35.8	220	9	-0.3	10.2
Llama-3.2-3B	44	9	-0.2	21	0	0.0	555	0	-1.3	85.5	370	13	-0.6	54.0
Llama-3.1-8B	7	0	0.0	12	0	0.0	734	0	-1.3	86.6	433	0	-0.7	76.3
Llama-3.3-70B	30	0	0.0	20	0	0.0	633	2	-1.3	85.0	413	0	-0.5	66.5
Phi-4-Mini	66	3	-0.0	33	0	0.0	429	1	-0.8	66.6	316	0	0.3	42.0
Phi-4	13	0	0.0	11	0	0.0	481	0	-0.5	73.2	302	0	-0.2	61.3
Qwen3-0.6B	49	0	0.0	153	1	-0.0	23	0	-0.8	65.2	373	12	-0.5	42.0
Qwen3-1.7B	81	0	0.0	69	0	0.0	698	0	-1.5	90.5	407	2	-0.4	53.3
Qwen3-4B	15	0	0.0	7	0	0.0	562	0	-1.3	88.6	350	0	-0.2	55.8
Qwen3-8B	23	0	0.0	4	0	0.0	761	0	-1.3	90.6	328	0	-0.1	58.3
Qwen3-14B	6	0	0.0	6	0	0.0	744	0	-1.2	89.0	368	0	-0.4	69.7
Qwen3-32B	4	0	0.0	1	0	0.0	679	0	-1.2	89.0	346	0	-0.4	66.4
GPT-OSS-20B	39	0	0.0	7	0	0.0	805	1	-1.4	92.7	215	0	-0.5	70.9
GPT-OSS-120B	7	0	0.0	0	0	0.0	825	1	-1.3	92.4	170	0	-0.4	69.4
GPT-5-Mini	7	0	0.0	1	0	0.0	718	0	-1.0	89.8	179	0	-0.4	70.1
GPT-5.2	0	0	0.0	0	0	0.0	504	0	-1.0	88.8	138	0	-0.6	82.5

Table 9: Citation error analysis: Table shows number of instances where Rk^{f} is lower than 100 (n) as well as the average difference to the required number of citations (Diff) and precision@k on these instances (P@k, omitted for RepliQA and BoolQ-M as it is always 0). See §4 for details

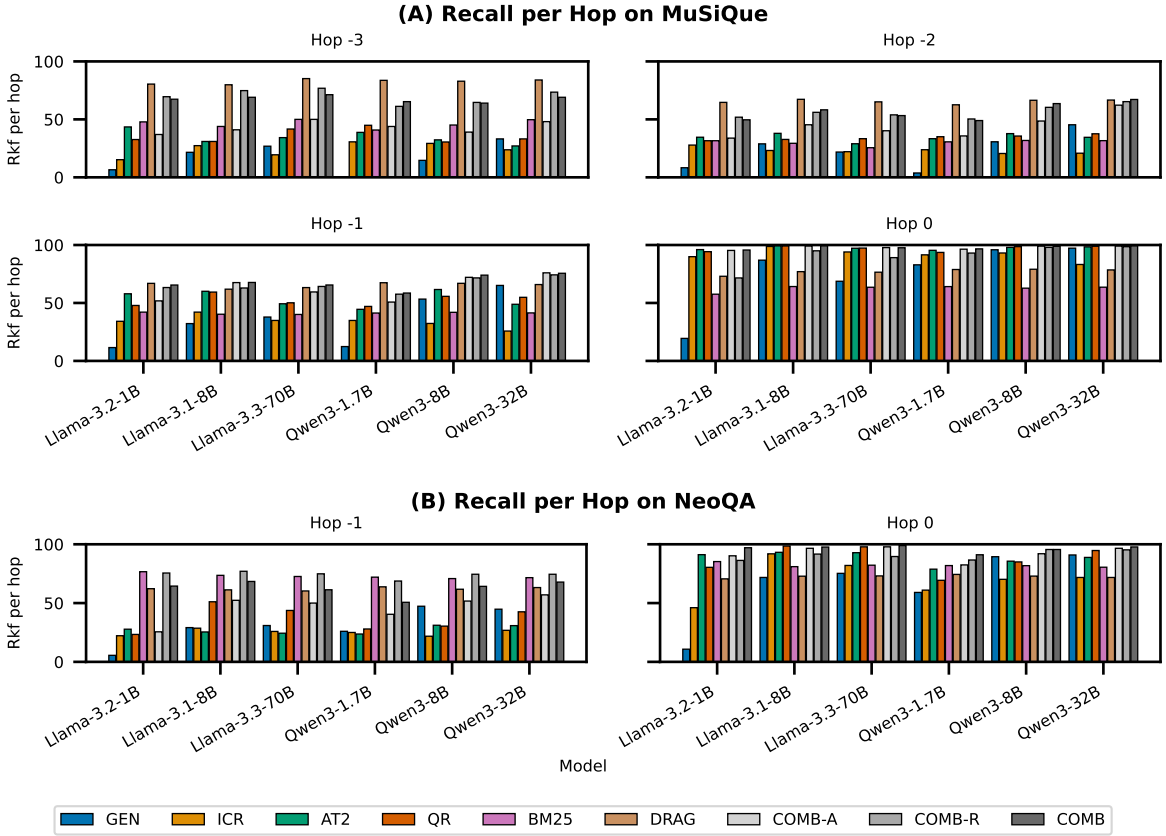

Figure 10: Recall per hop on MuSiQue and NeoQA `multi-hop` instances for models and citation approaches from §6.2.

		CITECONTROL					Transfer			
		Re	**Bo**	**Mu**	**Ne**	**Avg**	**QA**	**GO**	**Avg**	**Avg**
Llama-3.2-1B	GEN	41.4	44.6	14.1	8.6	27.2	8.4	7.1	7.7	20.7
	ICR	99.8	93.0	54.4	27.4	68.6	44.5	40.9	42.7	60.0
	AT2	99.8	89.0	68.9	38.0	73.9	46.9	42.9	44.9	64.2
	QR	**99.9**	99.8	63.8	34.2	74.4	**48.4**	**45.5**	**46.9**	65.3
	BM25	98.1	71.7	48.3	**73.8**	72.9	34.5	42.6	38.5	61.5
	DRAG	99.8	**100.0**	71.3	69.5	85.1	40.8	42.7	41.7	70.7
	COMB-A	99.9	99.8	65.9	35.1	75.2	48.0	44.0	46.0	65.5
	COMB-R	98.6	98.5	67.1	72.2	84.1	37.0	43.1	40.0	69.4
	COMB	99.9	99.8	**76.8**	71.9	**87.1**	45.6	45.4	45.5	**73.2**
Llama-3.1-8B	GEN	99.6	99.3	54.3	61.6	78.7	46.3	59.0	52.7	70.0
	ICR	**100.0**	89.6	61.7	60.6	78.0	67.4	83.8	75.6	77.2
	AT2	99.9	99.9	72.7	57.6	82.5	**67.9**	84.9	**76.4**	80.5
	QR	99.9	99.9	71.1	78.2	87.3	67.7	**85.0**	76.3	83.6
	BM25	98.7	70.2	49.6	75.6	73.5	52.7	81.6	67.2	71.4
	DRAG	99.6	**100.0**	71.8	68.6	85.0	62.5	80.9	71.7	80.6
	COMB-A	99.9	100.0	76.8	79.8	89.1	63.5	84.5	74.0	84.1
	COMB-R	100.0	100.0	77.2	**86.0**	90.8	59.1	81.7	70.4	84.0
	COMB	99.9	100.0	**80.9**	86.0	**91.7**	65.1	84.6	74.8	**86.1**
Llama-3.3-70B	GEN	98.2	98.8	49.1	61.4	76.9	46.7	43.1	44.9	66.2
	ICR	99.9	36.1	56.9	52.7	61.4	71.0	76.8	73.9	65.5
	AT2	99.8	98.2	66.4	60.1	81.1	70.1	83.0	76.5	79.6
	QR	**100.0**	95.4	67.7	76.3	84.8	**75.7**	86.7	**81.2**	83.6
	BM25	99.2	69.9	49.0	77.0	73.8	61.5	83.9	72.7	73.4
	DRAG	99.6	**100.0**	71.9	69.3	85.2	65.1	83.1	74.1	81.5
	COMB-A	100.0	99.8	73.3	79.5	88.1	71.0	86.7	78.9	85.1
	COMB-R	100.0	99.9	75.4	83.8	89.8	57.2	85.4	71.3	83.6
	COMB	100.0	100.0	**79.1**	**84.4**	**90.9**	73.2	**87.4**	80.3	**87.4**
Qwen3-1.7B	GEN	94.5	96.1	40.9	53.0	71.1	54.0	24.1	39.1	60.4
	ICR	99.7	46.1	56.9	39.8	60.6	55.2	53.0	54.1	58.4
	AT2	99.8	70.0	64.8	43.4	69.5	59.2	60.3	59.7	66.2
	QR	99.8	69.1	65.4	43.6	69.5	61.0	**61.7**	**61.3**	66.8
	BM25	98.6	70.0	50.6	73.6	73.2	49.2	51.6	50.4	65.6
	DRAG	99.6	**100.0**	74.4	70.8	86.2	53.7	55.9	54.8	75.7
	COMB-A	99.8	98.4	68.6	66.5	83.3	**61.3**	60.2	60.8	75.8
	COMB-R	99.8	99.9	73.8	**79.8**	**88.3**	55.9	53.5	54.7	77.1
	COMB	**99.9**	99.7	**75.2**	75.0	87.4	61.3	60.2	60.8	**78.5**
Qwen3-8B	GEN	98.6	99.8	68.0	80.8	86.8	62.6	43.6	53.1	75.6
	ICR	99.8	38.0	55.4	42.5	58.9	59.1	50.9	55.0	57.6
	AT2	**100.0**	94.4	73.2	52.2	79.9	69.4	66.4	67.9	75.9
	QR	99.9	44.1	70.5	51.8	66.6	69.4	**67.2**	**68.3**	67.1
	BM25	98.7	69.5	50.4	71.8	72.6	57.5	55.0	56.2	67.2
	DRAG	99.6	**100.0**	74.5	69.4	85.9	63.3	59.6	61.5	77.8
	COMB-A	100.0	99.9	80.0	82.5	90.6	**69.9**	63.6	66.7	82.6
	COMB-R	99.9	100.0	82.4	**90.4**	**93.2**	67.2	60.2	63.7	83.4
	COMB	100.0	100.0	**83.9**	87.2	92.8	69.9	64.2	67.0	**84.2**
Qwen3-32B	GEN	99.8	99.9	76.1	81.4	89.3	**63.5**	48.3	55.9	78.2
	ICR	99.0	15.0	48.0	41.4	50.9	39.9	31.2	35.6	45.8
	AT2	**100.0**	79.6	67.2	51.5	74.6	61.5	58.8	60.1	69.8
	QR	100.0	61.6	70.5	65.9	74.5	62.6	**60.8**	**61.7**	70.2
	BM25	99.1	69.5	50.5	71.8	72.7	53.8	48.2	51.0	65.5
	DRAG	99.5	**100.0**	73.8	68.7	85.5	58.2	51.7	54.9	75.3
	COMB-A	100.0	99.9	83.7	86.4	92.5	63.1	57.3	60.2	81.7
	COMB-R	99.9	99.9	84.6	**90.8**	**93.8**	60.6	54.2	57.4	81.7
	COMB	100.0	99.9	**85.4**	89.8	93.8	63.1	57.8	60.5	**82.7**

Table 10: Numerical values for Fig. 5. Metrics: CITECONTROL: Rk^{f}, Transfer: Attributability.