Citation Failure in LLMs: 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 does not hold for *citation failure*, when a model generates a helpful response, but fails to cite 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. 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 analyze failure modes. Experiments show that failures increase with relational complexity and suggest that combining citation methods could improve performance, motivating step 2. To improve LLM citation efficiently, 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 can enhance the usefulness of LLMs in RAG (Lewis et al., 2020) by allowing users to quickly verify responses through cited evidence. Yet, when a model generates a valid response but fails to output complete citations, this benefit diminishes: users must either search all sources or discard an otherwise helpful answer. To address this, we take a two-step approach: (1) we analyze when citation failure occurs, and (2) propose efficient methods to reduce it.

To illustrate, 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 combination of source documents (e.g. "1960"), and the cited evidence necessarily does not 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]).

Step 1: analyzing citation failure Recently, Hu et al. (2025) showed that the performance of citation evaluation models is strongly influenced by the "reasoning complexity" of inferring a response from evidence, but did not investigate the task of citation itself. Some works have analyzed properties of the source documents as factors influencing citation failure (e.g. Koo et al. 2024; Tang et al. 2024b; Sorodoc et al. 2025), but how the relation between response and evidence impacts citation failure has not been investigated. As the dataset by Hu et al. (2025) can't be used for evaluating LLM citation, a suitable analysis framework is missing.

Prior analyses have further limitations: First, they do not distinguish response and citation failure, which confounds analysis: retrieving supporting citations is not possible in the case of response failure. Second, they rely on LLM-based evaluators (Tang et al., 2024a; Honovich et al., 2022), whose accuracy can drop to $\sim 50\%$ in complex cases (Hu et al., 2025), making results unreliable.

To address these gaps, we ask:

RQ1: How does the response-evidence relation affect LLM Citation? Building on prior work in citation evaluation (Hu et al., 2025) and
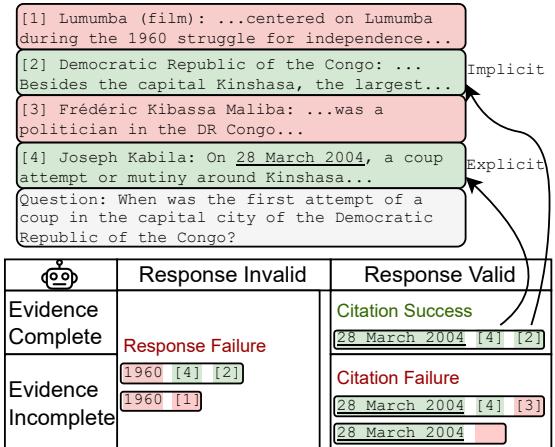

Figure 1: Citation Example: An LLM receives multiple documents and a question. The confusion matrix shows the possible outcomes for generated response and evidence. The response-evidence relation has reasoning type `multi-hop`. It is `explicit` for the response and [4], and `implicit` for the response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

intertextuality (Kuznetsov et al., 2022), we define key properties of this relation and propose CITECONTROL, a benchmark that varies these properties in a controlled fashion. All instances come with verifiable answers and known evidence, which allows controlling for response failure and makes our analysis independent of error-prone citation evaluation models.

Our experiments with generative citation from LLMs and retrieval-based baselines on CITECONTROL show that small LLMs struggle even for straightforward relations, while all models struggle in complex cases, such as multi-document reasoning. Analysis reveals that different citation methods suit different relation types, suggesting that their combination can improve performance.

Step 2: Mitigating citation failure Most existing approaches for improving LLM citation come with high resource demands either for additional training (e.g. Huang et al. 2024; Zhang et al. 2024; Li et al. 2024) or multiple LLM calls (e.g. Hirsch et al. 2025; Slobodkin et al. 2024; Chuang et al. 2025; Qi et al. 2024), calling for more efficient methods. In this direction, existing work uses retrievers for *post-hoc* citation (Sancheti et al., 2024; Ramu et al., 2024), but these are constrained by the limited model capacity.

In related settings, several works have successfully used attention values to efficiently exploit

the capabilities of LLMs: Zhang et al. (2025) and Chen et al. (2025) apply this paradigm in reranking, while Cohen-Wang et al. (2025) retrieve parts of the context that most influenced an LLM's output. This is efficient, as attention values are available "for free" during generation,[2] but the potential of these approaches to obtain more complete LLM citations has not been explored. Therefore, we aim to answer the research question

RQ2: How can LLM citation failure be mitigated efficiently? We contribute CITENTION, a framework that unifies generative citation with efficient retrieval-based and attention-based methods. Using CITECONTROL as a testbed and building on our results from step 1, we show that the efficient citation methods in CITENTION effectively mitigate generation-based citation failure, and that their combination often results in improved performance. Experiments on two transfer datasets show that the CITENTION methods are effective in absence of in-domain training data.

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, especially in cases with complex statement-evidence relations (§4). (2) CITENTION, a framework that integrates generation-based citation with efficient retrieval-based and attention-based methods (§5), and experiments showing that these methods can mitigate citation failure effectively on CITECONTROL and in a transfer setting (§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 LLM Citation A range of datasets (Malaviya et al., 2024; Kamalloo et al., 2023; Golany et al., 2024) and benchmarks (Gao et al.,

---

[2]In practice, attention implementations such as Flash Attention (Dao, 2023) do not give access to attention values, making an additional forward pass the most time-efficient implementation to date (Cohen-Wang et al., 2025).

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., 2024b; Buchmann et al., 2024; Wu et al., 2025), authorship information (Abolghasemi et al., 2024), and time-dependence and semantic properties of their contents (Sorodoc et al., 2025) as factors influencing citation failure. While Hu et al. (2025) have studied the effect of varying "reasoning complexity" between response and evidence on citation evaluation, how the response-evidence relation affects the task of citation itself has not been investigated.

Improving Corroborative Citation Prior works on improving corroborative citation fall into four classes: (1) Training-based approaches collect data and design training regimes (Huang et al., 2024; Penzkofer and Baumann, 2024; Zhang et al., 2024; Li et al., 2024). (2) Multi-step methods split attribution across multiple LLM calls (Hirsch et al., 2025; Slobodkin et al., 2024; 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., 2025; Chen et al., 2025) and contributive attribution (Cohen-Wang et al., 2025; Phukan et al., 2024; Ding et al., 2024). 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.

We make important contributions to the described research areas: In analyzing LLM citation, we are the first to provide methodology and ex-periments 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., 2024b; 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 $S = \{s_1, ..., s_{|S|}\}$. The task is to generate a *response* r based on S (e.g. an answer), and to retrieve corroborative *evidence* $E \subset S$ (see §2, Rashkin et al. 2023).

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`: Chain-like reasoning over multiple facts (as in Fig. 1, named "concatenation" in Hu et al. 2025).
- `intersection`: Reasoning over multiple facts in an aggregative manner (e.g. computing the time between two events).

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

	# train/dev/test	$|S|$	$|s|$	$|r|$	$|E|$	Reasoning	Overtness
Squad	68272/18549/5928	20.0	119.4	3.1	1.0	single	explicit
BoolQ	7541/1886/3270	20.0	96.2	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.
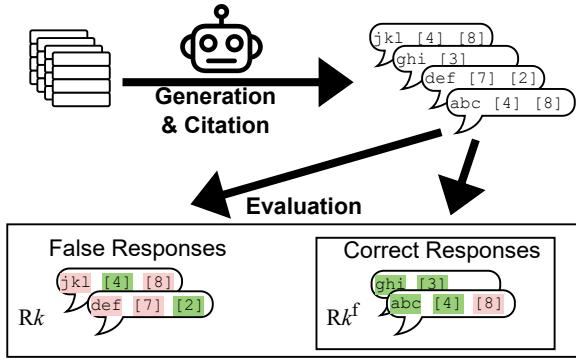

Figure 2: Evaluation strategy on CITECONTROL: For Rk, all predictions are evaluated for evidence recall @k, while for Rk^{f}, only predictions with correct responses are evaluated.

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.

SQuAD (Rajpurkar et al., 2018) and **BoolQ** (Clark et al., 2019) consist of tuples of a Wikipedia paragraph, question and answer. While SQuAD answers are extractive, BoolQ answers are "yes" or "no". **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 SQuAD, BoolQ and NeoQA, we combine evidence documents with distractors to obtain 20 source documents per instance (see §A.2 for details on data processing).

The reasoning type in SQuAD and BoolQ 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 8. It is explicit for SQuAD 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 and NeoQA intersection instances.

3.4 Evaluation

As in related work, we perform recall-focused evaluation of citations (Gao et al., 2023; Buchmann et al., 2024). Using annotated evidence as ground truth, we evaluate recall @k to avoid rewarding over-generation of evidence. We evaluate on all instances (Rk) and on the subset of correctly answered instances to disentangle response failure and citation failure (Rk^{f}, see Fig. 2 and §A.2).

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.3.3).

4.2 Results and Discussion

We ran 10 instruction-tuned LLMs[3] between 0.6B and 32B from the Llama (Grattafiori et al., 2025), Mistral (Jiang et al., 2023) and Qwen (Yang et al., 2025) families on CITECONTROL, showing results in Tab. 2. After analyzing the effect of our proposed filtered evaluation, we interpret the results with regard to our main research question.

Response failure has an impact on citation failure We observe that for most combinations of model and task, filtered recall @k ($\mathrm{R}k^{\mathrm{f}}$) is higher than or equal to unfiltered evaluation ($\mathrm{R}k$). While the magnitude of this difference depends on the model and dataset, on MuSiQue and NeoQA we observe of up to ∼15 points difference between $\mathrm{R}k^{\mathrm{f}}$ and $\mathrm{R}k$,[4] showing that controlling for response failure is important. In the following, we therefore focus on $\mathrm{R}k^{\mathrm{f}}$ scores.

Small models fail even for `single` reasoning, while all models fail in more complex reasoning For SQuAD and BoolQ, where `single` reasoning is required, models with 3B parameters or more achieve almost perfect $\mathrm{R}k^{\mathrm{f}}$ (>95), while the smaller models obtain lower $\mathrm{R}k^{\mathrm{f}}$ scores. On MuSiQue and NeoQA, that require more complex `multi-hop` and `intersection` reasoning over multiple documents, we observe reduced $\mathrm{R}k^{\mathrm{f}}$ scores for all models.

It seems that model size affects citation more than it affects response generation: Qwen3-1.7B has a higher proportion of correct responses than Qwen3-4B on SQuAD, MuSiQue and NeoQA, but substantially lower $\mathrm{R}k^{\mathrm{f}}$. Similarly, the difference

---

[3] We omit "-Instruct" specifiers for brevity.

[4] e.g. Ministral-8B, Qwen3-8B on MuSiQue, Qwen3-4B on NeoQA

---

between Llama-3.2-1B and Llama-3.2-3B is much more pronounced in $\mathrm{R}k^{\mathrm{f}}$ evaluation than in the proportion of correct responses.

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, suggesting that LLMs rank citations by confidence.

Evidence recall decreases with moving up in the reasoning chain Figure 3(B) shows $\mathrm{R}k^{\mathrm{f}}$ by the position of evidence in the reasoning chain for MuSiQue and NeoQA (`multi-hop` instances only). Compared to hop 0 (`explicit` overtness) $\mathrm{R}k^{\mathrm{f}}$ is strongly reduced for earlier hops (`implicit` overtness), showing that models struggle to trace the reasoning chain when performing citation. The retrieval-based baselines DRAG and BM25 are notable exceptions: Their recall is higher for hop -3 than hop -2. DRAG exhibits the highest hop -3 recall overall, while its hop 0 recall is lower than several LLMs. The retrieval-based models are forced to use the information from both question and response to find evidence. This means reduced focus on the response, which explains their sub-optimal performance for hop 0, which contains the response verbatim. At the same time, the question contains helpful information to find evidence for earlier hops (see the ROUGE scores in Table 8), which explains the elevated scores for these hops.

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

Explicitness can bias citation In Fig. 6, we show average $\mathrm{R}k^{\mathrm{f}}$ on `multi-hop` and `intersection` NeoQA instances, as well as $\mathrm{R}k^{\mathrm{f}}$ per hop. As expected, for most LLMs, $\mathrm{R}k^{\mathrm{f}}$ is highest on hop 0 evidence in `multi-hop` instances, likely due to the `explicit` response-evidence relation. Unexpectedly, $\mathrm{R}k^{\mathrm{f}}$ on hop -1 and average $\mathrm{R}k^{\mathrm{f}}$ on `multi-hop` instances is lower than average $\mathrm{R}k^{\mathrm{f}}$ on `intersection` instances for most models. This suggests that

	Squad			BoolQ			Musique			NeoQA		
	$\mathbf{R}k$	$\frac{|\mathcal{R}^{\mathrm{f}}|}{|\mathcal{R}|}$	$\mathbf{R}k^{\mathrm{f}}$	$\mathbf{R}k$	$\frac{|\mathcal{R}^{\mathrm{f}}|}{|\mathcal{R}|}$	$\mathbf{R}k^{\mathrm{f}}$	$\mathbf{R}k$	$\frac{|\mathcal{R}^{\mathrm{f}}|}{|\mathcal{R}|}$	$\mathbf{R}k^{\mathrm{f}}$	$\mathbf{R}k$	$\frac{|\mathcal{R}^{\mathrm{f}}|}{|\mathcal{R}|}$	$\mathbf{R}k^{\mathrm{f}}$
Ministral-8B	90.5	51.7	96.7	99.2	82.3	99.3	31.3	26.7	44.0	51.0	36.2	53.2
Llama-3.2-1B	28.8	43.0	32.4	40.0	55.4	41.1	10.0	11.5	14.1	8.0	14.8	9.6
Llama-3.2-3B	95.7	67.3	96.7	98.1	74.4	98.0	39.1	24.7	46.0	47.9	32.2	50.3
Llama-3.1-8B	96.6	78.8	97.1	99.4	81.3	99.5	44.7	36.4	54.3	63.6	41.4	63.7
Qwen3-0.6B	71.2	12.2	83.7	86.1	42.8	86.4	27.0	4.9	39.4	24.2	16.1	23.4
Qwen3-1.7B	88.4	50.9	92.0	95.5	71.1	95.6	33.5	15.6	45.0	37.0	42.0	38.4
Qwen3-4B	93.8	45.3	98.8	97.7	75.8	99.5	47.4	14.3	59.1	54.9	20.6	76.3
Qwen3-8B	97.1	67.8	98.9	99.7	80.7	99.7	55.4	30.2	68.8	68.4	47.0	72.0
Qwen3-14B	97.7	75.4	99.3	99.8	87.3	99.8	63.0	35.4	75.6	69.6	50.8	70.5
Qwen3-32B	97.0	70.6	99.8	99.5	80.4	99.7	60.0	33.5	73.8	69.2	47.9	74.6
Oracle-BM25	96.3	100.0	96.3	62.6	100.0	62.6	48.7	100.0	48.7	61.5	100.0	61.5
Oracle-DRAG	99.5	100.0	99.5	99.6	100.0	99.6	70.4	100.0	70.4	58.3	100.0	58.3

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). $\mathbf{R}k$ / $\mathbf{R}k^{\mathrm{f}}$: Recall @k on all instances / instances answered correctly. $\frac{|\mathcal{R}^{\mathrm{f}}|}{|\mathcal{R}|}$: Proportion of correctly answered instances.

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 $\mathbf{R}k^{\mathrm{f}}$ is higher on `multi-hop` instances than on `intersection` instances.

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

To mitigate the citation failures found in §4, we introduce CITENTION, our framework for enhancing generative LLM citation efficiently. 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 $\mathrm{M}(r, s) \to \mathbb{R}$, 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 combination methods $\mathrm{M}^{\Omega} = \mathrm{Agg}(\mathrm{M}^{1}, ..., \mathrm{M}^{|\mathrm{M}^{\Omega}|})$, 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 $\delta(\mathrm{M}(r, s)) : \mathbb{R} \to \{0, 1\}$ that maps the citation score to a decision `no-cite` or `cite`.

5.1 Citation Methods

Besides generative citation, we consider: Attention-based methods, for which the potential in reranking (Chen et al., 2025; Zhang et al., 2025)) and contributive attribution (Cohen-Wang et al., 2025) has recently been shown; Retrieval-based methods, which have been successfully applied for post-hoc citation (e.g. Bohnet et al. 2022; Sancheti et al. 2024). 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 $\mathrm{M}^{\mathrm{Gen}}$ as the length-normalized (Murray and Chiang, 2018) probability for generating the citation (see §A.3.1). 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. 2025) 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:
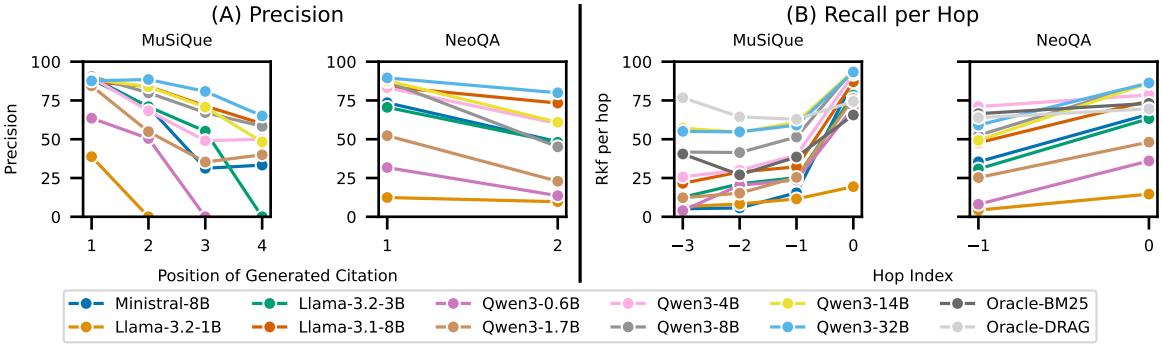
Figure 3: (A) Citation precision decreases with the order of appearance in generation. (B) Rk^f is highest for the evidence document that contains the response (hop 0) and is reduced when going to earlier hops, DRAG and BM25 are notable exceptions (see §4.2). Plots show data for correctly answered instances.

1. Compute a single score $M_d(r, s)$ per attention head h_d, where $d \in \{1, ..., |H|\}$ and $|H|$ is the number of attention heads (see §A.3.2).

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

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,[5] so $\forall d : \theta_d^{ICR} = \frac{1}{|H|}$

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

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

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

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:

$$M^\Omega = \sum_{i=1}^{|M^\Omega|} w_i M^i + b \qquad (1)$$

This retains efficiency and avoids introducing confounders into our analysis. To learn w and b, we fit a linear model[6] 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.2 and §6.2.1

6 How can LLM Citation Failure be Mitigated Efficiently?

In this section, we test the potential of the efficient citation methods in CITENTION. While previous work has studied generation-based citation (e.g. Gao et al. 2023; Tang et al. 2024b) and retrieval-based citation (e.g. Bohnet et al. 2022; Ramu et al. 2024) in isolation, we are the first to

---

[5] 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.

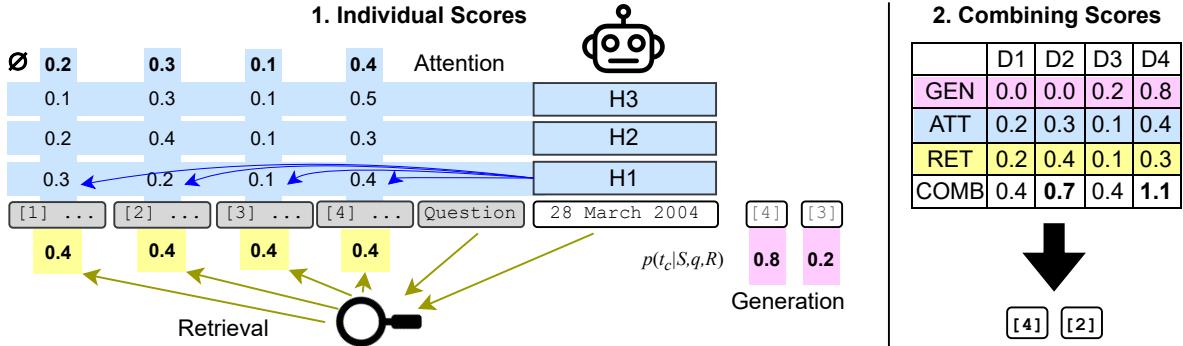[6] Using a LinearModel from scikit-learn (Pedregosa et al., 2011)

Figure 4: Overview of CITENTION. 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 method weights w, b are omitted.

investigate attention-based citation, and the combination of generation-, attention- and retrieval-based citation. We describe experiments and results on CITECONTROL (§6.1) and in a transfer setting (§6.2).

6.1 Experiments on CITECONTROL

We run the CITENTION methods on CITECONTROL to test if they can mitigate the citation failures found in §4, using a subset of models due to limited resources.

6.1.1 Experimental Setup

Training We use the train splits of the CITECONTROL datasets except for NeoQA: as it does not have a train split, we train on its dev split. 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 $|H^{\mathrm{QR}}|$ to 16 as in Zhang et al. (2025). 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.

6.1.2 Results

We evaluated the performance of the CITENTION methods on CITECONTROL, using Llama-3.2-1B, Llama-3.1-8B, Qwen3-1.7B and Qwen3-8B. Fig. 5(A) visualizes $\mathrm{R}k^{\mathrm{f}}$ scores, while Tab. 6 shows numbers and averages across datasets.

Efficient citation methods improve generative citation on CITECONTROL For Llama-3.2-1B-Instruct, the retrieval-based and attention-based methods improve $\mathrm{R}k^{\mathrm{f}}$ by up to 50 points on all

datasets. For Llama-3.1-8B, Qwen3-1.7B and Qwen3-8B, the improvements are small on the `single` reasoning datasets (SQuAD and BoolQ), as the generative citation performance is already high. On the datasets with more complex reasoning (MuSiQue and NeoQA), we observe increases in $\mathrm{R}k^{\mathrm{f}}$ of more than 10 points for these models.

Combining citation methods improves over individual methods For Llama-3.2-1B, Llama-3.1-8B and Qwen3-1.7B, combining all methods (COMB) results in the highest average $\mathrm{R}k^{\mathrm{f}}$ scores, while for Qwen3-8B, it is the combination of generative and retrieval-based citation COMB-R (Table 6). Notably, while none of the individual attention-based methods has higher average $\mathrm{R}k^{\mathrm{f}}$ than generative citation (GEN) for Qwen3-8B, their combination with GEN (COMB-A) improves average $\mathrm{R}k^{\mathrm{f}}$ by 5 points. This confirms that combined citation methods can complement each other in identifying relevant source documents.

Retrieval-based citation mostly performs better than attention-based citation. For all models, average $\mathrm{R}k^{\mathrm{f}}$ is higher when combining retrieval-based methods (COMB-R) than when combining attention-based methods (COMB-A). We attribute this to the fact that only the retrieval-based citation methods have access to the question, which is helpful in finding evidence (§4.2). Among the individual citation methods, we observe the highest average $\mathrm{R}k^{\mathrm{f}}$ for the Dragon retriever (DRAG) except for Llama-3.1-8B, where it is QR.

AT2 and QR are the best attention-based methods. Comparing average $\mathrm{R}k^{\mathrm{f}}$ of attention-based citation methods, we can observe that ICR is not
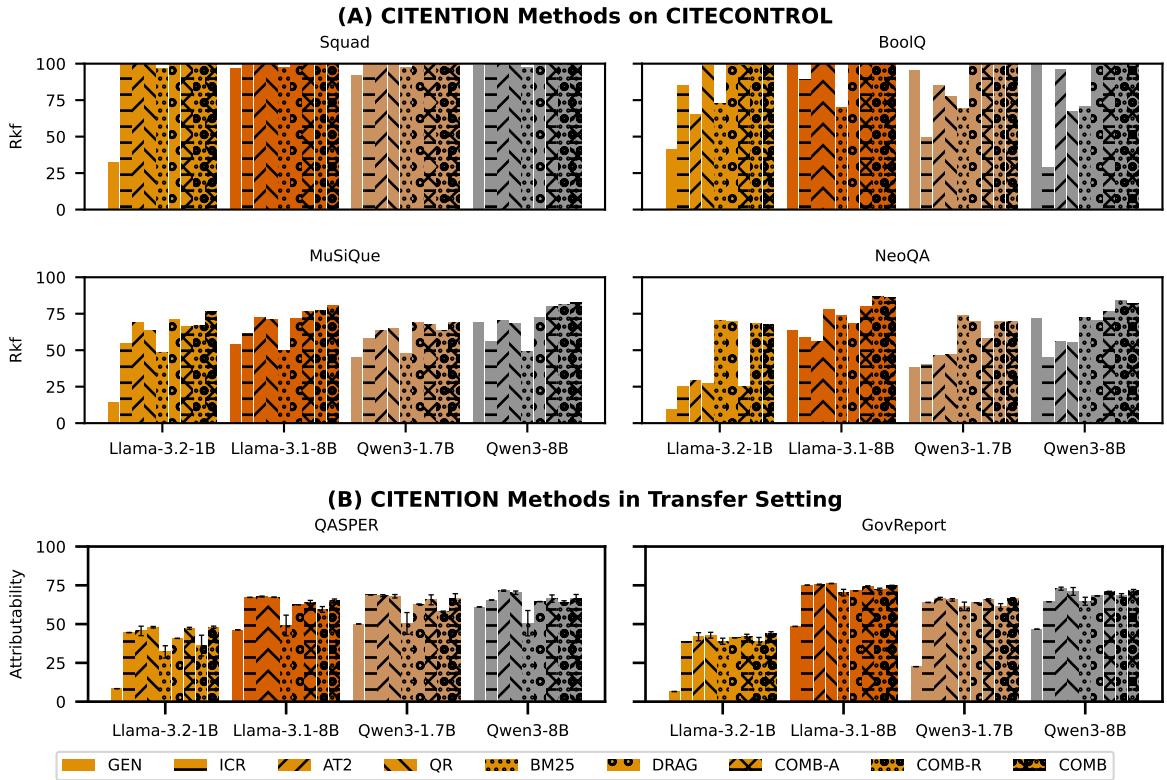
Figure 5: (A) CITENTION methods improve citation $\mathrm{R}k^{\mathrm{f}}$ scores on CITECONTROL (§6.1). (B) CITENTION methods trained on CITECONTROL improve $\mathrm{R}k^{\mathrm{f}}$ scores on unseen tasks (§6.2). Bars show proportion of answer statements that are attributable to the evidence (averaged over train datasets). Whiskers show standard deviation. For the unaggregated data see Tab. 7.

able to improve over generative citation (GEN), with the exception of Llama-3.2-1B. In contrast, QR and AT2 mostly improve over GEN and ICR. We explain this with the fact that they were optimized on the respective datasets in CITECONTROL, while ICR was not. Further, we find that QR works better for the Llama models, while AT2 works better for the Qwen models.

Attention-based methods and `BM25` are sensitive to overtness　While $\mathrm{R}k^{\mathrm{f}}$ scores on SQuAD (`explicit` statement-evidence relation) are consistently high for attention-based and retrieval-based citation methods, they are reduced on BoolQ (`implicit`) for attention-based methods and `BM25` (Fig. 5A) showing that these methods are sensitive to changes in overtness. This is an interesting contrast to GEN and DRAG, where the scores are higher on BoolQ than on SQuAD.

Efficient methods can improve citation in complex cases　Fig. 7 shows the recall of evidence by its position in the reasoning chain for MuSiQue and NeoQA. We can observe improvements on

all hops from efficient citation methods. 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.

6.2 Transfer

We showed that generation-based citation can be improved efficiently with retrieval-based and attention-based methods in the controlled setup of CITECONTROL. Next, we extend our investigation to a transfer setting, without task-specific training data and longer response statements: We evaluate the CITENTION methods on two challenging datasets from a recent long-document attribution benchmark (Buchmann et al., 2024).

6.2.1 Experimental Setup

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 $\rightarrow$ coherent document) and for QASPER in domain (Wikipedia / General $\rightarrow$ 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., 2024a) 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 sentence. We report the proportion of response statements evaluated as attributable to the 2 highest-scoring source documents ($k = 2$).

6.2.2 Results

We evaluated the parameters trained on each of the CITECONTROL tasks (§6.1) on QASPER and GovReport. Fig. 5(B) shows the proportion of attributable response statements averaged over the 4 train tasks, while Tab. 7 shows complete results.

CITENTION methods are effective in transfer settings. Attributability is higher for the attention-based and retrieval-based methods and their combination than for generation-based citation alone. This means that LLM citation can be efficiently enhanced without requiring additional training. Our findings on retrieval-based methods agree with those from Ramu et al. (2024) and Sancheti et al. (2024), who found these can improve over LLMs in post-hoc citation, but did not investigate attention-based methods.

Attention-based methods improve over retrieval-based methods. In contrast to the results on CITECONTROL, citations from attention-based methods result in higher attributability scores than citations from retrieval-based methods. 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. As in §6.1, the performance of AT2 and QR is roughly similar, while ICR lags behind in several cases (e.g. Qwen3-8B on both transfer datasets).

Individual attention-based methods can improve over method combination. For Llama-3.1-8B, Qwen3-1.7B and Qwen3-8B, average attributability is higher for AT2 and QR than for their combination with generation-based and retrieval-based citation (COMB-A, COMB). This suggests that while the attention head weights θ transfer well, the score combination weights w, b do not transfer well in these cases. Only for Llama-3.2-1B, combining all citation methods results in best average attributability.

The train task has a small effect on attention-based methods, but a larger effect on BM25. The standard deviations of in Fig. 5(C) show that the fluctuations in attributability are small for AT2 and QR, suggesting that the train data has a small effect on transfer performance for these methods. This reflects the findings from Cohen-Wang et al. (2025) and Zhang et al. (2025), who found similar results in contributive attribution and retrieval, respectively. In contrast, the data used for obtaining the token statistics for BM25 has a stronger effect on its performance, visible in the high standard deviation on QASPER.

7 Conclusion

In this work, we have defined citation failure and presented key findings for understanding and mitigating it: First, using our controlled citation evaluation framework CITECONTROL, we showed that citation failure occurs frequently, especially in cases with complex statement-evidence relations. Building on this, we proposed CITENTION, a framework that unifies generation-based citation with efficient retrieval-based and and attention-based methods, and showed that it can mitigate citation failure effectively, with promising results from combining multiple citation methods.

We hope that our research inspires further investigation into mitigating citation failure. We are particularly excited about research into combining multiple citation methods in transfer settings, and in enabling native access to attention values in efficient attention implementations, which could increase the efficiency of CITENTION even further.

8 Acknowledgements

References

Amin Abolghasemi, Leif Azzopardi, Seyyed Hadi Hashemi, Maarten de Rijke, and Suzan Verberne. 2024. Evaluation of attribution bias in retrieval-augmented large language models. *ArXiv Preprint*, abs/2410.12380.

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, Shannon Zejiang Shen, Zhaofeng Wu, Hu Xu, Xi Victoria Lin, James Glass, Shang-Wen Li, and Wen tau Yih. 2025. Selfcite: Self-supervised alignment for context attribution in large language models. *ArXiv Preprint*, abs/2502.09604.

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 Mądry. 2024. Contextcite: Attributing model generation to context. In *Advances in Neural Information Processing Systems*, volume 37, pages 95764–95807. Curran Associates, Inc.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *ArXiv preprint*, abs/2307.08691.

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. 2024. Attention with dependency parsing augmentation for fine-grained attribution. *ArXiv Preprint*, abs/2412.11404.

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. *ArXiv Preprint*, abs/2505.05949.

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. 2025. 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. *ArXiv Preprint*, abs/2506.01187.

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. *ArXiv Preprint*, abs/2401.14640.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chap-

lot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv Preprint*, abs/2505.05949.

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*.

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

Dongfang Li, Zetian Sun, Baotian Hu, Zhenyu Liu, Xinshuo Hu, Xuebo Liu, and Min Zhang. 2024. 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.

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.

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.

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.

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

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. 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. Association for Computational Linguistics.

Marco Tulio 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*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

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

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. *ArXiv Preprint*, abs/2506.07671.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024a. 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. 2024b. L-citeeval: Do long-context models truly leverage context for responding? *ArXiv preprint*, abs/2410.02115.

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, Rémi 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 M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv Preprint*, abs/1910.03771.

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. 2024. Longcite: Enabling llms to generate fine-grained citations in long-context qa. *ArXiv Preprint*, abs/2409.02897.

Wuwei Zhang, Fangcong Yin, Howard Yen, Danqi Chen, and Xi Ye. 2025. Query-focused retrieval heads improve long-context reasoning and re-ranking. *ArXiv Preprint*, abs/2506.09944.

A Replication Information

A.1 Prompts

For each dataset, we use a 3-shot prompt, taking care to maximize diversity between in-context examples and shortening the example source documents for efficiency reasons. As in related work, each source document $s_i \in S$ is prepended by its index in square brackets (Gao et al., 2023), and questions are always put after the source documents (Buchmann et al., 2024). Below, we show an example prompt. For task explanations and format explanations used, see 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 CITECONTROL Details

Data processing SQuAD and BoolQ come with a single context paragraph. For each instance, we combine it with 19 randomly selected distractor paragraphs from other instances. For NeoQA, we select 20 articles as source documents per instance, such that 1 or 2 of them are required as evidence. Models receive a list of 1 true and 5 distractor answer options as proposed by the NeoQA authors. For all datasets in CITECONTROL, we remove unanswerable instances.

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 SQuAD and MuSiQue, and exact match for BoolQ 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.3 Details of Citation Methods in CITENTION

A.3.1 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]").

Dataset	task_explanation
SQuAD	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. If the question is a yes/no question, 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.
BoolQ	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	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.

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

$$\mathrm{M}^{\mathrm{Gen}}(r, s) =$$

$$\exp\left(\frac{1}{|c|}\sum_{i=1}^{|c|}\log\left(\frac{\exp(\ell_t[t_i^c])}{\sum_{v=1}^{V}\exp(\ell_t[v])}\right)\right) \quad (2)$$

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

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

Dataset	format_explanation
SQuAD	Retrieved Paragraphs: <omitted> Question: When did Beyonce start becoming popular? Answer: in the late 1990s [7]
BoolQ	Retrieved Paragraphs: <omitted> Question: Can alcohol cause depression? Answer: yes [7]
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[7]
Dense retrieval	Sentence Transformers (Reimers and Gurevych, 2019)
Aggregation weight fitting	Scikit-Learn (Pedregosa et al., 2011)
ROUGE score computation	Rouge-Score[8]

Table 5: Python packages used in experiments.

A.3.2 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.3.3 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[9] and set hyperparameters to common values k1= 1.5; b= 0.75 (Robertson and Zaragoza, 2009).

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.4 Technical Details

Dataset splits We use the development splits of SQuAD, BoolQ and MuSiQue for evaluation, as their test splits are hidden. We use the test splits of NeoQA, QASPER and GovReport (CRS subset).

Generation All text was generated at temperature 0 for maximum reproducibility. For the Qwen models, opening and closing thinking tokens ("<think></think>") were added to the prompt to ensure comparability with non-reasoning models.

Mapping evidence to hops in NeoQA To perform the analysis of recall per hop for NeoQA (Figs. 3 and 6), 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.

Used packages See Tab. 5

---

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

B Additional Results

		SQuAD	BoolQ	MuSiQue	NeoQA	Avg
Llama-3.2-1B	GEN	32.4	41.1	14.1	9.6	24.3
	ICR	99.2	85.2	54.4	25.1	66.0
	AT2	99.9	65.3	68.9	29.5	65.9
	QR	99.9	99.0	63.8	27.2	72.5
	BM25	97.0	72.8	48.3	**70.2**	72.1
	DRAG	**100.0**	**99.9**	71.3	69.9	85.3
	COMB-A	**100.0**	98.8	65.9	25.4	72.5
	COMB-R	99.7	99.4	67.1	68.4	83.7
	COMB	**100.0**	99.7	**76.8**	67.8	**86.1**
Llama-3.1-8B	GEN	97.1	99.5	54.3	63.7	78.7
	ICR	**100.0**	88.8	61.7	58.6	77.3
	AT2	**100.0**	**100.0**	72.7	55.6	82.1
	QR	**100.0**	99.9	71.1	78.1	87.3
	BM25	97.4	69.8	49.6	73.6	72.6
	DRAG	**100.0**	**100.0**	71.8	68.1	85.0
	COMB-A	**100.0**	**100.0**	76.8	79.7	89.1
	COMB-R	**100.0**	**100.0**	77.2	**86.6**	90.9
	COMB	**100.0**	**100.0**	**80.9**	86.3	**91.8**
Qwen3-1.7B	GEN	92.0	95.6	45.0	38.4	67.7
	ICR	99.5	49.4	58.3	40.0	61.8
	AT2	99.9	85.0	63.4	46.3	73.7
	QR	99.9	77.6	65.2	46.8	72.4
	BM25	97.5	69.5	47.6	**74.1**	72.2
	DRAG	**100.0**	**100.0**	68.9	69.4	84.6
	COMB-A	99.9	99.2	67.6	57.8	81.1
	COMB-R	99.6	98.6	63.4	69.8	82.8
	COMB	99.9	99.5	**69.2**	70.0	**84.7**
Qwen3-8B	GEN	98.9	99.7	68.8	72.0	84.9
	ICR	99.1	29.0	55.7	45.0	57.2
	AT2	**100.0**	96.0	70.7	56.0	80.7
	QR	**100.0**	67.3	68.5	55.1	72.7
	BM25	97.4	70.4	49.4	72.7	72.5
	DRAG	**100.0**	**100.0**	72.4	70.3	85.7
	COMB-A	**100.0**	**100.0**	80.0	76.7	89.2
	COMB-R	**100.0**	**100.0**	81.1	**83.9**	**91.2**
	COMB	**100.0**	**100.0**	**82.5**	82.2	91.1

Table 6: Results from CITENTION methods on CITECONTROL. All numbers show Rk^{f} values. See §6.2 for analysis and discussion.

	Eval Task	QASPER				GovReport				
	Train Task	SQ	BO	MU	NE	SQ	BO	MU	NE	Avg
Llama-3.2-1B	GEN	8.4	8.4	NaN	8.4	6.5	6.5	6.5	6.5	7.3
	ICR	44.5	44.5	44.5	44.5	38.5	38.5	38.5	38.5	41.5
	AT2	47.4	41.0	47.0	47.0	44.5	38.2	42.6	42.1	43.7
	QR	48.0	47.2	48.3	**48.3**	44.1	43.1	44.0	40.1	45.4
	BM25	35.1	32.5	34.4	26.3	40.6	39.5	39.5	36.0	35.5
	DRAG	40.8	40.8	40.8	40.8	41.2	41.2	41.2	41.2	41.0
	COMB-A	**48.2**	47.0	46.8	47.0	43.8	40.8	42.0	42.2	44.7
	COMB-R	39.8	39.3	39.3	26.1	40.2	40.3	40.0	35.3	37.5
	COMB	48.1	**47.9**	**48.8**	47.1	**45.1**	**43.7**	**44.3**	**42.4**	**45.9**
Llama-3.1-8B	GEN	46.3	46.3	46.3	46.3	48.5	48.5	48.5	48.5	47.4
	ICR	67.2	67.2	67.2	67.2	75.1	75.1	75.1	75.1	71.1
	AT2	**67.7**	**68.2**	**67.5**	**67.5**	75.9	75.3	75.3	75.5	71.6
	QR	67.4	67.1	67.4	67.4	**76.2**	**76.0**	**76.2**	**76.3**	**71.7**
	BM25	54.1	50.3	52.6	39.4	71.8	71.1	71.1	67.1	59.7
	DRAG	62.5	62.5	62.5	62.5	71.4	71.4	71.4	71.4	66.9
	COMB-A	64.3	63.1	64.0	65.6	74.6	74.0	74.0	74.7	69.3
	COMB-R	59.9	59.9	61.2	56.8	72.6	72.7	72.9	71.1	65.9
	COMB	64.3	63.7	65.5	66.4	74.7	74.2	74.5	75.1	69.8
Qwen3-1.7B	GEN	50.1	50.1	50.1	50.1	22.6	22.6	22.6	22.6	36.3
	ICR	68.9	**68.9**	**68.9**	68.9	63.9	63.9	63.9	63.9	66.4
	AT2	**69.0**	67.4	68.5	68.4	**67.1**	66.3	**66.9**	65.4	**67.4**
	QR	68.5	66.3	68.4	68.6	66.4	65.3	66.0	64.6	66.8
	BM25	55.3	51.5	54.7	40.2	63.7	62.3	62.3	57.5	55.9
	DRAG	62.8	62.8	62.8	62.8	63.8	63.8	63.8	63.8	63.3
	COMB-A	66.2	61.7	65.0	69.5	65.9	64.9	65.3	66.4	65.6
	COMB-R	57.9	57.6	57.8	58.6	62.9	62.1	62.0	59.1	59.7
	COMB	66.3	62.9	66.6	**70.4**	66.2	**66.5**	66.4	**67.2**	66.6
Qwen3-8B	GEN	61.1	61.1	61.1	61.1	46.7	46.7	46.7	46.7	53.9
	ICR	65.5	65.5	65.5	65.5	64.4	64.4	64.4	64.4	65.0
	AT2	**71.2**	**71.1**	**71.9**	**71.9**	**73.6**	**73.2**	**73.2**	71.4	**72.2**
	QR	71.1	68.6	70.6	70.6	73.1	71.0	72.6	67.5	70.6
	BM25	56.4	51.6	55.6	38.8	66.6	65.7	65.7	60.9	57.7
	DRAG	64.5	64.5	64.5	64.5	68.2	68.2	68.2	68.2	66.3
	COMB-A	65.4	64.8	65.5	70.0	71.2	70.7	71.1	70.9	68.7
	COMB-R	64.3	64.0	64.8	61.7	69.1	68.6	69.4	66.7	66.1
	COMB	65.4	65.0	66.4	70.1	71.5	70.5	72.0	**72.3**	69.2

Table 7: Results from CITENTION methods on QASPER and GovReport. All numbers show proportion of attributable predictions according to attributability evaluation models. For analysis and discussion see §6.2. SQ: SQuAD, BO: BoolQ, MU: MuSiQue, NE: NeoQA

		Answer		Question	
		Evi	No Evi	Evi	No Evi
Squad	-	1.00	0.10	0.68	0.24
BoolQ	-	0.04	0.03	0.69	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
	Aggregative	0.17	0.14	0.67	0.54

Table 8: 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.
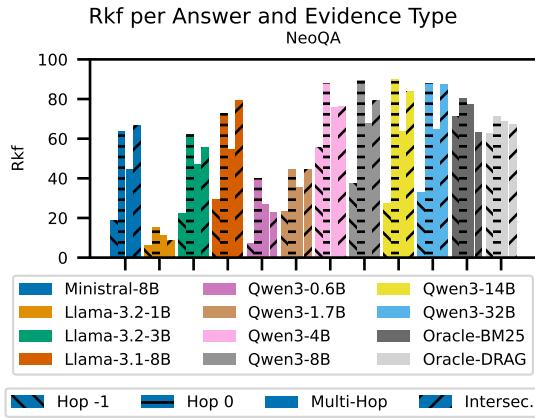

Figure 6: Detailed results for NeoQA. Hop -1/0: Rk^{f} per hop on `multi-hop` instances. `multi-hop` / `intersection`: average Rk^{f} for the respective instance type. For analysis and discussion see §4.
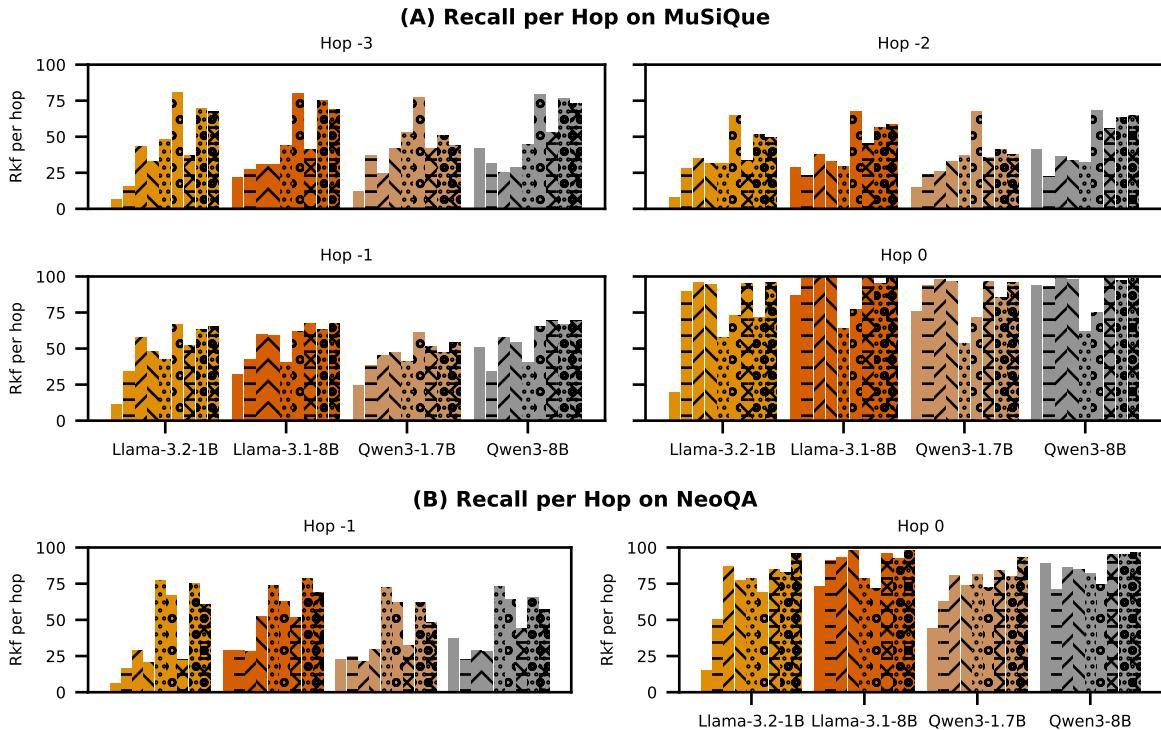

Figure 7: Recall per hop on MuSiQue and NeoQA `multi-hop` instances for models and citation approaches from §6.1.