

CiteEval: Principle-Driven Citation Evaluation for Source Attribution

Yumo Xu¹ Peng Qi^{2◊*} Jifan Chen^{1*} Kunlun Liu¹ Rujun Han^{3◊}
 Lan Liu¹ Bonan Min¹ Vittorio Castelli¹ Arshit Gupta¹ Zhiguo Wang¹

¹AWS AI Labs ²Orby.ai ³Google

{yumomxu, chenjf, k11, liuall, bonanmin, vittorca, arshig, zhiguow}@amazon.com
 peng@orby.ai rujunh@google.com

Abstract

Citation quality is crucial in information-seeking systems, directly influencing trust and the effectiveness of information access. Current evaluation frameworks, both human and automatic, mainly rely on Natural Language Inference (NLI) to assess binary or ternary supportiveness from cited sources, which we argue is a suboptimal proxy for citation evaluation. In this work we introduce CiteEval, a citation evaluation framework driven by principles focusing on fine-grained citation assessment within a broad context, encompassing not only the cited sources but the full retrieval context, user query, and generated text. Guided by the proposed framework, we construct CiteBench, a multi-domain benchmark with high-quality human annotations on citation quality. To enable efficient evaluation, we further develop CITEEVAL-AUTO, a suite of model-based metrics that exhibit strong correlation with human judgments. Experiments across diverse systems demonstrate CITEEVAL-AUTO’s superior ability to capture the multifaceted nature of citations compared to existing metrics, offering a principled and scalable approach to evaluate model-generated citations.¹

1 Introduction

Information-seeking systems, such as retrieval-augmented generation (RAG; Lewis et al. 2020) for question answering, play a vital role in how we access and understand knowledge. A key aspect of these systems is their ability to provide accurate source attribution, typically in the form of citations (Gao et al., 2023b). Accurate citations establish user trust and enable verification of generated content (Liu et al., 2023; Malaviya et al., 2024). Evaluating the quality of citations, however, remains a significant and under-addressed challenge. Pioneered by Attributable to Identified Sources (AIS;

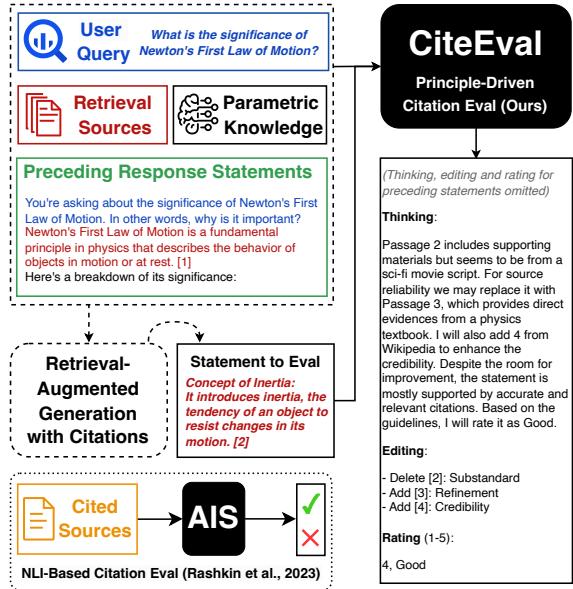


Figure 1: CiteEval considers all contexts used in the generation phase (with dashed lines) for fine-grained citation assessment. In contrast, AIS (Rashkin et al. 2023; bottom left) judges citation quality with NLI, solely based on **cited sources** (e.g., Passage 2 in the shown example), a subset of **retrieval sources**. Statements in the response are in the same color as the contexts to which they should be attributed.

Rashkin et al. 2023), existing work has largely focused on measuring the degree of *supportiveness*, using frameworks based on Natural Language Inference (NLI; Williams et al. 2018), for both human (Gao et al., 2023b; Yue et al., 2023) and automatic evaluation (Zhang et al., 2024c; Fierro et al., 2024).

In this work we critically examine the received wisdom and rethink whether NLI is truly the optimal lens through which to evaluate citation quality. As shown in Figure 1, NLI-based metrics determine the quality of citations solely based on the cited passages. On the other hand, RAG systems often consume a wide range of contexts, and content in a model-generated response often come from para-

*Equal contribution. [◊]Work done at AWS AI Labs.

¹Our code and datasets can be found at <https://github.com/amazon-science/CiteEval>

phrasing the user query (Katsis et al., 2025), fusing parametric knowledge from pre-training (Jiang et al., 2023), or reasoning over preceding statements in the response (Trivedi et al., 2023). Not all can and should be attributable to the retrieved passages, and doing so, as a result of *context insufficiency* in evaluation, leads to inaccurate estimation of citation quality. Additionally, equating citation quality to binary or ternary supportiveness often falls short in capturing the nuances of citation utility (Malaviya et al., 2024). For instance, in Figure 1, the original citation that supports the “*Concept of Inertia*” is from a sci-fi movie script, which can be supportive but less informative and credible compared to a physics book or Wikipedia page, as shown in the “*Thinking*” and “*Editing*” sections.

To address these limitations, we introduce CiteEval, a citation evaluation framework driven by a set of design principles, including:

1. *Evaluating citations against full retrieval sources* (Section 2.2),
2. *Evaluating citations beyond the retrieval context* (Section 2.3), and
3. *Evaluating citations with fine-grained criteria and scenarios* (Section 2.4).

Particularly, CiteEval mitigates the context insufficiency in NLI-based approaches with Principles 1 and 2 jointly, and moves beyond supportiveness and entailment in NLI by considering what constitutes good citations to human with Principle 3. Guided by the proposed framework, we develop CiteBench, a high-quality citation evaluation benchmark to support research in this field. CiteBench includes statement-level human judgments on RAG responses and citations, constructed with a multi-stage annotation process comprising context attribution, critical editing, and citation rating. To enable efficient assessment of citation quality, we further propose CITEEVAL-AUTO, a suite of model-based metrics that demonstrate a strong correlation with human judgment. We benchmark the citation quality of a wide array of existing systems, providing insights on the impact of critical RAG components, and the potential of CiteEval in citation improvement.

2 Citation Evaluation Principles

2.1 Background

Problem Formulation We first briefly introduce the task formulation for RAG. Given a corpus consisting of documents \mathcal{D} , a retriever fetches relevant sources \mathcal{S} (e.g., fixed-length passages) for user query Q : $\mathcal{S} = \text{Retriever}_\psi(\mathcal{D}, Q)$. A generator then reads the retrieved sources \mathcal{S} to produce a response. Following Gao et al. (2023b), we employ an LLM-based generator which generates a text response R , *jointly* with fine-grained citations \mathcal{C} : $R, \mathcal{C} = \text{Generator}_\chi(\mathcal{S}, Q)$.² As shown in Figure 1, when applicable, citations are provided at the end of each *statement*: $\mathcal{C} = \{\mathcal{C}_i\}_{i=1}^{|R|}$ where $|R|$ denotes the number of statements in the response, and statement-level citations $\mathcal{C}_i = \{c_{ij}\}$ is a subset of retrieved passages: $\mathcal{C}_i \subseteq \mathcal{S}$.

Attributable to Identified Sources (AIS) For the i th sentence in the response R_i , AIS (Rashkin et al., 2023) evaluates its citations \mathcal{C}_i as:

$$r_i \stackrel{\text{def}}{=} \text{NLI}_\phi(\text{concat}(\mathcal{C}_i), R_i) \quad (1)$$

where the rating $r_i \in \{0, 1\}$ denotes whether statement R_i (i.e., the hypothesis) can be inferred from the concatenation of *cited* sources \mathcal{C}_i (i.e., the premise). As the rating approximates whether all required sources are cited, it has been adopted as citation *recall* in AIS and its automatic version, Auto-AIS (Gao et al., 2023a) based on an NLI model ϕ . Auto-AIS further extends the notion to citation *precision*, by applying ϕ to assess the relevance of each individual citation $c_{ij} \in \mathcal{C}_i$ (Liu et al., 2023; Gao et al., 2023b).

CiteEval: A Principle-Driven Framework Departing from existing approaches based on NLI, such as Auto-AIS, we propose CiteEval which formulates the problem of citation evaluation as:

$$r_i \stackrel{\text{def}}{=} f_\theta(\mathcal{C}_i; \mathcal{S}, R, Q). \quad (2)$$

Particularly, f_θ directly estimates fine-grained ratings r_i based on the full retrieval sources \mathcal{S} and beyond (e.g., R and Q). We then derive the response-level rating based on $\{r_i\}_{i=1}^{|R|}$. We next state the principles mentioned in Section 1 that we follow to drive the modeling of f_θ .

²Approaches to generate citations are many and varied (see Section 6 for details). We opt for joint generation considering its effectiveness and conceptual simplicity, but nothing prohibits this work from being applied to other approaches.

User Context
• Affirmation: <i>Sometimes, when falling asleep, individuals may experience a heightened perception of sounds, which can feel extremely loud when Q asks about the reason behind sound sensitivity when falling asleep.</i>
Response Context
• Mathematical reasoning: ... <i>Therefore, an Oreo without the filling has 21.6 calories</i>)
• Logical reasoning: ... <i>Therefore, it is not possible for some planets to orbit the Sun in the opposite direction</i>
• Planning: ... <i>However, I need to find who played Zordon in the original Power Rangers series</i>
• Abstention: <i>I cannot find an answer from your documents.</i>
Parametric Context
• Factual: <i>The molecular formula of phenol is C6H5OH</i> when the retrieval context only mentions its molar mass.
• Formatting: <i>The Washington Redskins went to the Super Bowl in the following years</i> for response coherence.

Table 1: Examples of statements attributable to contexts beyond retrieval.

2.2 Evaluating Citations against Full Sources

One underlying assumption of AIS is that *self-evaluation* of \mathcal{C}_i is sufficient, i.e., its rating is independent of the cite-worthiness of uncited sources $\mathcal{S}_i^- = \mathcal{S} \setminus \mathcal{C}_i$. While source quality can be partially estimated based on its content, we note that its cite-worthiness should be determined in relation to other citation candidates. For instance, citations will always receive the maximum AIS recall, as long as they semantically entail the target response statement. This leads to over-estimation of the quality, when more reliable sources exist in the retrieval context and should be cited instead (Malaviya et al., 2024). On the other hand, a citation will receive the lowest AIS precision score when it is partially supportive, even if there exists no better source that can fully support the statement. In this case, the score is underestimated as the citation can still be helpful in source verification, compared to the case where no citation is provided.

To address these issues, we follow the principle that citation quality should be estimated against *full* retrieval sources, i.e., all documents or passages retrieved for response and citation generation. This allows the model to leverage the quality of uncited sources for more accurate quality estimation for cited sources. Additionally, full retrieval sources are also necessary in determining the boundaries with other contexts that should be considered in citation evaluation, which we will discuss next.

2.3 Evaluating Citations Beyond Retrieval Context

Statements in the model response are often attributable to contexts beyond the retrieval context. For instance, in Figure 1, “*You’re asking about the significance of Newton’s First Law of Motion*” is simply a repetition of the user query and should not be attributed to the retrieved sources. We consider three additional types of contexts beyond the retrieval sources: the *user*, the *response*, and the *parametric knowledge*.

Table 1 (upper block) presents a common case where responses start with a leading statement that repeats or paraphrases the query. These statements do not introduce new claims or facts beyond the context supplied by the *user*, and are therefore not applicable for citation evaluation. Autoregressive language models also condition token generation on their preceding statements in the *response*, enabling local reasoning of varied types, from mathematical reasoning to planning. Despite being topically related, statements shown in Table 1 (middle block) are not directly conditioned on specific evidence in the retrieval context.

The profound inherent knowledge of LLMs further allows *parametric* facts to be incorporated in their outputs. These facts can be identified based on whether they are provided as part of the retrieval, user, or response context. Formatting statements, such as transitional expressions in long-form answers (Xia et al., 2024), also fall into this category, considering the procedural knowledge that LLMs encapsulate. When only parametric knowledge is involved in a statement, i.e., it is fully parametric, no citation should be possible given \mathcal{S} .³

As contexts beyond retrieval are not citable, the above-mentioned statement types are *not applicable* (N/A) for citation evaluation. In prior work, they are either implicitly penalized (Gao et al., 2023a,b) or promoted (Zhang et al., 2024b) in existing citation evaluation frameworks. In this paper, we emphasize that all these contexts should be explicitly considered to enable accurate citation evaluation.

2.4 Evaluating Citations with Fine-Grained Criteria and Scenarios

Criteria Citations for the same statement can be inter-dependent. To address this, NLI-based recall metrics evaluate citations for each statement as a whole, and project the combinatorial citation space

³We discuss partially parametric cases in Appendix B.

Subset	Source	#Instances
ASQA	Wikipedia [W]	948 queries
ELI5	Sphere [S]	1,000 queries
MS MARCO	Bing [B]	1,000 queries
LFRQA	LoTTE [L]	1,000 queries
Total	[W], [S], [B], [L]	3,948 queries
Full Development	[W], [S], [B]	948 queries
Full Test	[W], [S], [B], [L]	3,000 queries
Metric Development	[W], [S], [B]	200 responses
Metric Test	[W], [S], [B]	1,000 responses

Annot. Ratio

Contexts	retrieval (87.0%), user (0.6%), response (9.3%), parametric (3.1%)
Ratings	1-5: 10.3%, 2.1%, 8.6%, 16.9%, 62.0%
Edits	delete: misleading (6.9%), substandard (1.3%), redundancy (4.5%); add: evidence (11.3%), refinement (1.4%), credibility (6.7%); keep: 67.8%

Table 2: CiteBench summary. We report the distributions of queries and responses (above) and human annotations for citation evaluation (below).

into binary (Gao et al., 2023b) or ternary (Zhang et al., 2024b; Yue et al., 2023) supportiveness. This can potentially be too coarse when there are multiple citations and multiple evaluation aspects to consider. Inspired by text generation where human evaluation is typically performed on a Likert scale (Xu and Lapata, 2022; Zheng et al., 2023), we argue that citation evaluation can benefit from a more fine-grained evaluation schema, with rating guidelines incorporating relevant dimensions, such as citation redundancy and source credibility.

Scenarios Accessing the full contexts described in Section 2.3 helps evaluate the citation quality for all statements in the full response. However, we note that the full contexts may not always be observable to end users: commercial search engines such as Perplexity AI⁴ and Microsoft Copilot⁵ only present *cited* sources to users, rather than the full retrieval results. In this scenario, users can only possibly focus on the statements that are cited, and highlighting the citation quality for these statements may better align with the user experience. To this end, citation evaluation frameworks should cover the following two scenarios: 1) **Full**, which evaluates all statements requiring citations, amongst which the uncited ones are penalized, and

2) **Cited**, where statements without citations, no matter whether citable, are treated as N/A.

3 Building A Citation Benchmark with Principled Human Annotation

Driven by the outlined principles, we create CiteBench, a multi-domain citation evaluation dataset. Particularly, CiteBench factorizes citation evaluation into three steps: context attribution, citation editing, and citation rating. Next we will introduce the human annotation process, followed by the benchmark details.

3.1 Human Evaluation

Process and Guidelines For context attribution, we provided human annotators a query, source passages retrieved by the query, and a model response consisting of one or more statements. Annotators were asked to first read all source passages and then attribute each statement to one of the major context types: retrieval, user, response, or parametric. See Appendix A.2 for the complete guidelines.

For statements attributed to the retrieval context, annotators were then asked to provide critical edits, to serve as rating evidences. Particularly, we provided three deletion actions for different reasons: delete-misleading, delete-substandard, and delete-redundant, as well as three addition actions: add-evidence, add-refinement, and add-credibility. Descriptions for these actions and detailed instructions were provided in Appendix A.3 as part of the annotation guidelines. This allows edits to be performed with fine-grained reasons to impose varied impacts on citation ratings. For instance, deleting a misleading citation is likely to harm the rating more than a redundant one. Each action operates on one target citation, which is either from existing citations (for delete actions), or other retrieved sources (for add actions). Citations that are not associated with an action are considered as keep.

As the final step, we further asked annotators to rate the overall citation quality for each statement on a 1-5 Likert scale. We present the rating guidelines in Appendix A.4, which emphasize the fine-grained citation issues identified in the previous step, such as citation redundancy and missing evidence.

Critical editing and rating were skipped for statements attributed to contexts other than retrieval, which were labeled as not applicable (N/A). We ob-

⁴<https://www.perplexity.ai>

⁵<https://copilot.microsoft.com>

Metric	Evaluator	Model	CiteBench-Statement			CiteBench-Response		
			Pearson	Spearman	Kendall-Tau	Pearson	Spearman	Kendall-Tau
<i>AutoAIS-based Metrics</i>								
AUTOAIS-PRECISION		T5-XXL	—	—	—	0.170	0.058	0.057
AUTOAIS-RECALL		T5-XXL	0.409	0.264	0.237	0.223	0.075	0.073
AUTOAIS-F1		T5-XXL	—	—	—	0.219	0.105	0.097
AUTOAIS-PRECISION [†]		T5-XXL	0.416	0.315	0.278	0.256	0.113	0.106
AUTOAIS-F1 [†]		T5-XXL	0.419	0.315	0.278	0.249	0.115	0.108
<i>AttriScore-based Metrics</i>								
ATTRISCORE-STRICT [†]		GPT-4-turbo	0.459	0.281	0.254	0.196	0.079	0.097
ATTRISCORE-RELAXED [†]		GPT-4-turbo	0.447	0.274	0.249	0.098	0.066	0.092
ATTRISCORE-STRICT [†]		GPT-4o	0.449	0.297	0.269	0.221	0.094	0.108
ATTRISCORE-RELAXED [†]		GPT-4o	0.450	0.291	0.263	0.128	0.080	0.104
<i>LQAC-based Metrics</i>								
LQAC-PRECISION		GPT-4o	—	—	—	-0.043	-0.092	-0.057
LQAC-RECALL		GPT-4o	0.607	0.423	0.375	0.526	0.447	0.379
LQAC-F1		GPT-4o	—	—	—	0.118	0.071	0.090
LQAC-PRECISION [†]		GPT-4o	0.468	0.269	0.247	0.147	0.052	0.082
LQAC-F1 [†]		GPT-4o	0.468	0.284	0.255	0.182	0.074	0.096
<i>CiteEval-Auto Metrics (Ours)</i>								
CITEEVAL-AUTO (ITERCoE)		GPT-4o	0.710	0.549	0.491	0.647	0.580	0.499
CITEEVAL-AUTO (EDITDIST)		GPT-4o+MLR	0.711	0.558	0.486	0.633	0.585	0.487
CITEEVAL-AUTO		GPT-4o+MLR	0.731	0.559	0.486	0.668	0.589	0.492

Table 3: Human correlation of different citation evaluation metrics on CiteBench (metric test set). [†] denotes our adapted version described in Section 4.3 for statement-level evaluation. CITEEVAL-AUTO (last row) is an ensemble which linearly interpolates scores from the two proposed rating methods.

tain response-level citation ratings via aggregating human ratings for statements that are applicable for citation evaluation with mean pooling.

Quality Control Data annotation was performed by contracted data professionals with three blind passes. Across the three annotations for each sample, the context for each statement is determined using the majority vote, and the citation rating is determined using the average rating. We have a dedicated team of data linguists to validate the annotation quality. We performed three rounds of pilot annotation to fine-tune the taxonomy of context types and citation edits, addressing the ambiguities in the provided guidelines. The Inter-Annotator Agreement (IAA) of context attribution and citation rating are 0.980 and 0.774, respectively (The Krippendorff’s α).

3.2 Benchmark Dataset Construction

Query Sampling and Passage Retrieval We focus on Long-Form QA (LFQA) datasets for query sampling, as long answers for non-factoid queries

can lead to more diverse citation behaviors for citation evaluation. Specifically, we include ASQA (Stelmakh et al., 2022), ELI5 (Fan et al., 2019), MS MARCO (Bajaj et al., 2018), and LFRQA (Han et al., 2024). Table 2 summarizes the query distribution, featuring the coverage of multiple domains. Particularly, besides common knowledge corpora (e.g., Wikipedia and Bing), we cover queries from five emerging domains in LFRQA, including Science, Technology, Lifestyle, Recreation, and Writing. We provide 10 fixed-size passages as the retrieval context for each query, obtained with various retrievers for each dataset to ensure contextual diversity. More details can be found in Appendix C. We uniformly sample in total 948 instances from MS MARCO, ASQA and ELI5 as a development set and use the rest 3,000 instances for testing.

Response and Citation Generation for Human Annotation

To perform human evaluation described in Section 3.1, following Gao et al. (2023b) we generate responses and citations in one pass

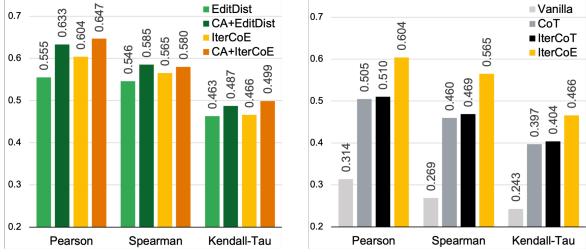


Figure 2: Effects of context attribution (left) and citation editing (right) in citation evaluation.

with LLMs. We instruct citations to be generated in brackets at sentence end, which are then extracted with regex. The detailed prompt can be found in Appendix F.1. To control the annotation size, we randomly sample from ASQA, ELI5, and MS MARCO 100 queries each, and consider outputs from 4 models, including GPT-4o and GPT-4o-mini, Llama-3-70b and Llama-3-8b (MetaAI, 2024), to balance proprietary and open-source models of varied sizes. This yields a dataset of 1,200 instances for human annotation. We randomly sampled 200 instances for metric development and the rest 1,000 instances for meta-evaluation. We present the annotation statistics in Table 2.

4 Model-Based Citation Evaluation

As human evaluation can be costly and time-consuming, we propose CITEEVAL-AUTO to automate citation evaluation with model-based metrics.

4.1 Model-Based Context Attribution

We instruct LLMs to attribute response statements to their contexts based on the context definition. Gao et al. (2023b) perform evaluation for each statement independently. For instance, the statement to be evaluated in Figure 1 “*It establishes the concept of inertia*” is handled without co-reference resolution for *It*. To avoid the ambiguity, we provide all statements in one prompt, and instruct LLMs to iteratively attribute each statement to its context in one pass. Citations are removed from responses to avoid introducing the bias of the retrieval context for cited statements (and vice versa). Detailed prompt can be found in Appendix F.2.

4.2 Model-Based Citation Rating

For statements applicable for citation evaluation, we propose and discuss two rating approaches.

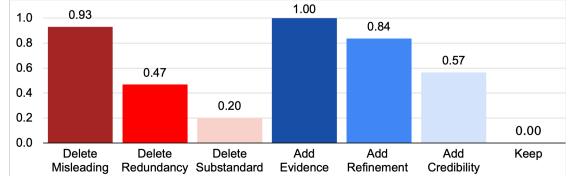


Figure 3: Edit distance for actions in EDITDIST. Estimated on the metric development set.

Iterative Chain of Edits (ITERCOE) Given the citation rating guidelines, we instruct LLMs to follow a thinking step-by-step fashion by first reasoning about each statement against the given contexts. We supply to LLMs `delete` and `add` actions introduced in Section 3.1, allowing a sequence of edits to be generated for improving the citation quality. LLMs then rate the target sentence on a 1-5 Likert scale, based on the generated edits and the rating guidelines. Detailed prompts can be found in Appendix F.3. We normalize the estimated ratings to $[0, 1]$. We mask out the ratings for statements identified as N/A in context attribution, and aggregate the rest statement-level ratings into a response-level rating via mean pooling.

Edit Distance (EDITDIST) Alternatively, we rate citations based on required edit actions and their estimated distances. Particularly, different types of citation errors (e.g., a missing essential citation vs. a redundant one) should ideally impact the quality rating differently. To this end, we learn an edit distance for each edit action type, based on how strongly the frequency of that action correlates with the overall human Likert score for a given statement. Specifically, let $\{a_k\}_{k=1}^K$ denote the edit actions defined in Section 3.1, and $|\mathcal{A}_{i,k}^*|$ the number of occurrences of action a_k in ground-truth actions \mathcal{A}_i^* . We estimate the distance function $d(a_k)$ and overall rating r_i via multiple linear regression $\min \sum_i \text{MSE}(r_i, \hat{r}_i)$, where \hat{r}_i denotes the human rating from the metric development set. The estimated rating r_i is defined as:

$$r_i = \sum_{k=1}^K d(a_k) * \frac{|\mathcal{A}_{i,k}^*|}{|\mathcal{A}_i^*|} + b \quad (3)$$

where b is a bias term. At test time, actions \mathcal{A}_i are first generated using the same instructions as in ITERCOE and then used for rating estimation.

4.3 Metric Evaluation Setup

We compare the performance of CITEEVAL-AUTO against existing automatic metrics based on NLI:

Models	CITEEVAL-AUTO		Statistics	
	Full	Cited	$ R $	M
<i>Proprietary Models</i>				
GPT-4o	0.898	0.949	1.975	0.197
GPT-4o-mini	0.848	0.925	1.759	0.217
GPT-4-turbo	0.863	0.940	1.835	0.250
GPT-3.5-turbo	0.724	0.839	1.352	0.223
<i>Open-source Instruction-tuned Models</i>				
Llama-3-70b	0.909	0.926	1.853	0.159
Llama-3-8b	0.800	0.871	2.398	0.250
Mixtral-8×22b	0.746	0.871	2.322	0.386
Mixtral-8×7b	0.755	0.827	2.554	0.363
Qwen2.5-72b	0.895	0.913	1.461	0.161
Qwen2.5-7b	0.663	0.722	8.467	0.950
<i>Open-source Fine-tuned Models</i>				
LongCite-9B	0.564	0.843	8.867	0.435
LongCite-8B	0.559	0.846	8.694	0.452

Table 4: Citation quality in **Full** and **Cited** scenarios of different LLMs responses (full test set). We also report the response length $|R|$ and missing citation ratio M .

AUTOAIS (Gao et al., 2023b) Unlike AutoAIS recall, AUTOAIS Precision and F1 do not operate at the statement level. Apart from the original results, we tailor the framework to first produce statement-level precision and F1 scores, which are then averaged to response-level ratings (similar to our proposed approach).

ATTRSCORE (Yue et al., 2023) ATTRSCORE evaluates each citation independently and classifies it as attributable, extrapolatory, or contradictory. We convert discrete categories into continuous ratings in two settings: i) strict, which assigns a rating of 1 to *attributable* and 0 to both *extrapolatory* and *contradictory*, and ii) relaxed, which assumes *extrapolatory* is to be relevant (but insufficient) which is assigned 0.5.

LQAC (Zhang et al., 2024b) LQAC (Long-Context QA with Citations) extends AutoAIS to include *partial support* in precision and employs GPT-4o as the NLI model. Similar to AUTOAIS, we adapt it to first produce statement-level ratings and then response-level ratings.

4.4 Metric Evaluation Results

Human Correlation Table 3 shows the human correlation results in the **Full** scenario.⁶

⁶Results for the **Cited** scenario can be found in Appendix D.4. Details of the two scenarios are provided in Section 2.4.

CITEEVAL-AUTO based on GPT-4o substantially outperforms state-of-the-art citation evaluators, at both the statement- and response-level.⁷ We note that with GPT-4o as the backbone, LQAC-RECALL achieves higher correlation compared to AUTOAIS-RECALL. Interestingly, LQAC-PRECISION does not yield better performance than its AUTOAIS counterpart, although it is reported to achieve higher correlation with binary human labels on supportiveness.

Ablation Study As shown in Table 7, our proposed context attribution model yields 0.957 F1 in predicting a statement’s applicability for citation evaluation (see Appendix D.2 for a detailed performance breakdown). To understand the effects of context attribution on final citation ratings, we further perform an ablation study and present the results in Figure 2 (left). We compare standalone citation rating models (e.g., EDITDIST) and their full CITEEVAL-AUTO pipelines augmented with context attribution (e.g., CA+EDITDIST). Removing context attribution causes substantial performance drops for both rating models we proposed. Figure 2 (right) further compares with the following approaches that directly rate citations without citation editing: VANILLA which directly rates all statements given the guidelines, CoT which performs a reasoning step before rating, as well as ITERCoT which interleaves reasoning with citation rating. ITERCoE outperforms these approaches by large margins, showing the effectiveness of explicitly reasoning over the editing space and aligning model-generated edits with the rating guidelines.

We further show in Figure 3 the estimated distance for each edit action. As we can see, add actions lead to higher penalties compared to their delete counterparts, demonstrating the necessity of identifying missing or better citation sources in citation evaluation.

5 Citation Benchmarking for RAG

Models For a comprehensive automatic evaluation of LLMs on CiteBench, in addition to GPT-4o and Llama-3 model families used in human evaluation (Section 3.2), we expand proprietary models to include GPT-4-turbo (2024-04-09) and GPT-3.5-turbo. For public models, we further include two Mixtral models (Jiang et al., 2024): Mixtral-

⁷We also experimented with other LLM backbones such as GPT-4-turbo and GPT-4o performs the best. Details are provided in Appendix D.1.

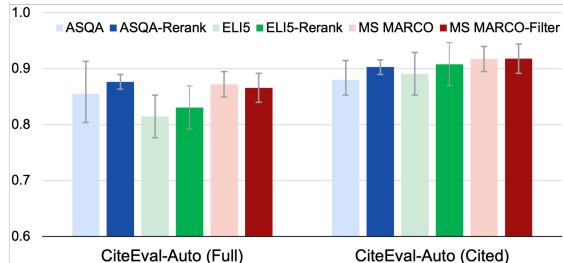


Figure 4: Performance of different retrieval settings (full development set) in both **Full** and **Cited** scenarios. Error bars denote the standard deviation over averaged ratings for different models.

$8 \times 22B$ -Instruct and Mixtral- $8 \times 7B$ -Instruct, and two Qwen models (Team, 2025): Qwen2.5-72b and Qwen2.5-7b. We also benchmark LongCite-9B and LongCite-8B (Zhang et al., 2024b) which are fine-tuned for QA with citations.

Results We show in Table 4 the benchmarking results. In the **Cited** scenario, GPT-4o is the top-performing model and Llama3-70b is on par with GPT-4o-mini. On the other hand, in the **Full** scenario, Llama3-70b outperforms GPT-4o and achieves the best performance.

To better understand the ranking differences in the two scenarios, we further examine the response and citation statistics from the benchmarked models, including the response length $|R|$ (i.e., average number of statements in a response) and missing ratio M (i.e., average ratio of statements without citation). We found that GPT-4o tends to produce longer responses than Llama3-70b. To verify the impact of response length on citation behaviors, we measured the Pearson correlation between $|R|$ and M which is 0.679 ($p < .001$), indicating a strong positive correlation, i.e., longer responses are more likely to miss citations. This reveals the challenges in jointly generating long-text generation *and* providing complete citations for all citable content. We provide more details on the correlation analysis in Appendix D.3.

We also observed high missing citation ratios and substantially longer responses from fine-tuned LongCite models, leading to a large rating gap between the two evaluation scenarios.⁸

⁸In Zhang et al. (2024b) fine-tuned models are shown to perform better than proprietary models, measured by LQAC. We acknowledge that the conclusion cannot be drawn based on CITEEVAL-AUTO. One potential reason is LQAC assigns the highest rating to all statements that do not require citations. This leads to an over-estimation of citation quality for long responses wherein many N/A statements exist.

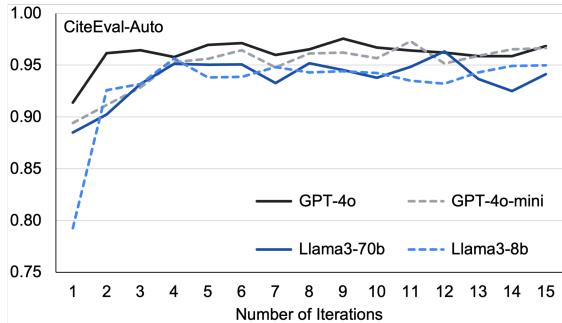


Figure 5: Performance improvement with iterative citation editing (response-level; metric development set).

Effects of Retrieval Quality To examine how retrieval quality affects citation quality, we generate responses for ASQA and ELI5 with re-ranked retrieval contexts which have substantially higher recall of relevant passages. For MS MARCO, as the original corpus is not accessible for reranking, we maximize the retrieval precision via filtering the retrieval context and keeping only passages annotated as relevant by human. Figure 4 shows that using re-ranked contexts with higher retrieval recall often leads to higher citation quality in both evaluation scenarios. On the other hand, citation quality does not benefit from better retrieval precision via filtering on MS MARCO.

Citation Improvement with Edit Actions Towards exploring the potential of CiteEval in citation improvement, we iteratively generate and execute edit actions and examine the rating dynamics. Specifically, for a set of citations $\mathcal{C}^{(t)}$, we leverage CITEEVAL-AUTO to jointly generate edit actions $\mathcal{A}^{(t)}$ and citation ratings $r^{(t)}$. We execute the actions against $\mathcal{C}^{(t)}$ to generate a set of new citations $\mathcal{A}^{(t)} : \mathcal{C}^{(t)} \mapsto \mathcal{C}^{(t+1)}$. We repeat the process T times, and report the citation rating $r^{(t)}$ at each iteration in Figure 5. We observe that CiteEval consistently improves citation quality across models, where larger models such as GPT-4o reach the performance peak quicker than smaller models. Regardless of the initial performance and model size, models within the same family converge to similar performance after a sufficient number of iterations. This opens up opportunities to improve small LLMs’ source attribution performance with the executable critique from CITEEVAL-AUTO and inference-time scaling (Snell et al., 2024).

6 Related Work

The increasing demand for the deployment of LLMs in information-seeking systems has spurred efforts in source attribution, with external evidences presented as URLs (Muller et al., 2023), snippets (Gao et al., 2023a), quotes (Menick et al., 2022), or retrieved sources (Gao et al., 2023b). Regardless of the evidence presentation, NLI is commonly adopted to judge the attribution quality (Liu et al., 2023; Zhang et al., 2024b). Yue et al. (2023) introduced two evaluation sets for source attribution, derived from existing QA data and AI search engines. The evaluation sets assume citations are provided independently and evaluate only the leading sentence of a response. In this work, we focus on improving the evaluation of *in-line* citations to retrieved sources, with a high-quality benchmark consisting of fine-grained human annotations for full responses.

7 Conclusion

We proposed CiteEval, a principle-driven framework centered on fine-grained citation ratings within comprehensive evaluation contexts. Based on CiteEval, we constructed a high-quality citation evaluation dataset CiteBench, and proposed CITEEVAL-AUTO, an automated metric for scalable evaluation. Experiments across diverse RAG systems highlight CITEEVAL-AUTO’s enhanced ability in evaluating and improving citation quality.

Directions for future work are many and varied. One research challenge is to develop distillation techniques to approximate CiteEval judgments with smaller LMs. We would also like to extend the proposed framework to RAG reward modeling and post-training, and enhance the trustworthiness of AI responses through effective knowledge grounding and attribution.

8 Limitations

Context attribution in this work focuses on typical contexts in RAG and can be expanded to cover more diverse use cases. For example, user’s demographic information such as age and location is often used for more personalized responses (Zhang et al., 2024d), which can also be considered as part of the user context in addition to queries. Also, context attribution is introduced as a sub-task for citation evaluation in this work. The task can be further applied to citation *generation*, to move beyond

attribution to the retrieval context and enable information verification from broader contexts. For instance, citations to contexts beyond retrieval could be provided through special tokens denoting the context (such as [P] for parametric knowledge), or in the form of natural language.

While our work establishes a strong correlation between CITEEVAL-AUTO and fine-grained human judgments, a comprehensive evaluation of its real-world, downstream impact is a crucial next step. Such *extrinsic* evaluation requires user studies or task-based evaluations that reliably measure constructs such as user trust and verification experience across downstream applications. We believe that CiteEval provides a necessary foundation by offering a more reliable and principled *intrinsic* evaluation of citation quality, paving the way for future studies into its *extrinsic* impact.

Additionally, in this work we treat sentences as statements, a notion that can be extended to cover finer-grained text chunks.⁹ Towards evaluating citations in a more end-to-end setting, the retrieval step in RAG can be further incorporated in the proposed framework, potentially with ground-truth annotations on the relevance of retrieval sources.

References

- Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. **MS MARCO: A human generated machine reading comprehension dataset**. Preprint, arXiv:1611.09268.
- Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. **ELI5: Long form question answering**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy.
- Constanza Fierro, Reinald Kim Amplayo, Fantine Huot, Nicola De Cao, Joshua Maynez, Shashi Narayan, and Mirella Lapata. 2024. **Learning to plan and generate text with citations**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11397–11417, Bangkok, Thailand.
- Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan,
- ⁹Chunk-level citations are now supported by Anthropic API, subsequent to this work’s completion: <https://www.anthropic.com/news/introducing-citations-api>

- and Kelvin Guu. 2023a. **RARR: Researching and revising what language models say, using language models**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada.
- Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. **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.
- Rujun Han, Yuhao Zhang, Peng Qi, Yumo Xu, Jenyuan Wang, Lan Liu, William Yang Wang, Bonan Min, and Vittorio Castelli. 2024. **RAG-QA arena: Evaluating domain robustness for long-form retrieval augmented question answering**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4354–4374, Miami, Florida, USA.
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, and et al. 2024. **Mixtral of experts**. [arXiv:2401.04088](#).
- Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. **Active retrieval augmented generation**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. **Dense passage retrieval for open-domain question answering**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781, Online.
- Yannis Katsis, Sara Rosenthal, Kshitij Fadnis, Chu-laka Gunasekara, Young-Suk Lee, Lucian Popa, Vraj Shah, Huaiyu Zhu, Danish Contractor, and Marina Danilevsky. 2025. **MTRAG: A multi-turn conversational benchmark for evaluating retrieval-augmented generation systems**. [Preprint](#), arXiv:2501.03468.
- 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 *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, BC, Canada.
- Charles Lipson. 2011. **Cite right: a quick guide to citation styles—MLA, APA, Chicago, the sciences, professions, and more**. University of Chicago Press.
- 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.
- 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.
- Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. **Teaching language models to support answers with verified quotes**. [Preprint](#), arXiv:2203.11147.
- MetaAI. 2024. **The llama 3 herd of models**. [Preprint](#), arXiv:2407.21783.
- Benjamin Muller, John Wieting, Jonathan Clark, Tom Kwiatkowski, Sebastian Ruder, Livio Soares, Roei Aharoni, Jonathan Herzig, and Xinyi Wang. 2023. **Evaluating and modeling attribution for cross-lingual question answering**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 144–157, Singapore.
- 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.
- Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. **CoBERTv2: Effective and efficient retrieval via lightweight late interaction**. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States.
- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. **Scaling llm test-time compute optimally can be more effective than scaling model parameters**. [Preprint](#), arXiv:2408.03314.
- Ivan Stelmakh, Yi Luan, Bhuvan Dhingra, and Ming-Wei Chang. 2022. **ASQA: Factoid questions meet long-form answers**. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates.
- Qwen Team. 2025. **Qwen2.5 technical report**. [Preprint](#), arXiv:2412.15115.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. **Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, USA.

Congying Xia, Chen Xing, Jiangshu Du, Xinyi Yang, Yihao Feng, Ran Xu, Wenpeng Yin, and Caiming Xiong. 2024. [FOFO: A benchmark to evaluate LLMs’ format-following capability](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 680–699, Bangkok, Thailand.

Yumo Xu and Mirella Lapata. 2022. [Document summarization with latent queries](#). *Transactions of the Association for Computational Linguistics*, 10:623–638.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. [Automatic evaluation of attribution by large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4615–4635, Singapore.

Huajian Zhang, Yumo Xu, and Laura Perez-Beltrachini. 2024a. [Fine-grained natural language inference based faithfulness evaluation for diverse summarisation tasks](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1701–1722, St. Julian’s, Malta.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. 2024b. [LongCite: Enabling LMs to generate fine-grained citations in long-context qa](#). *Preprint*, arXiv:2409.02897.

Weijia Zhang, Mohammad Aliannejadi, Yifei Yuan, Jiahuan Pei, Jia-hong Huang, and Evangelos Kanoulas. 2024c. [Towards fine-grained citation evaluation in generated text: A comparative analysis of faithfulness metrics](#). In *Proceedings of the 17th International Natural Language Generation Conference*, pages 427–439, Tokyo, Japan.

Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, Nesreen Ahmed, and Yu Wang. 2024d. [Personalization of large language models: A survey](#). *Preprint*, arXiv:2411.00027.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-bench and chatbot arena](#). In *Thirty-seventh Conference on*

Neural Information Processing Systems Datasets and Benchmarks Track, New Orleans, Louisiana, USA.

A Human Annotation Instructions

A.1 General Guidelines

Your task is to evaluate the quality of citations generated by language models based on a given query, a set of retrieved passages relevant to that query, and the model’s generated text containing the citations. Below, we define what is meant by the query, retrieved passages, and citations:

Query Usually an information-seeking question like “What is the significance of Newton’s First Law of Motion?”.

Retrieved Passages For each query, you will be given a few (maximum 10) relevant passages from the indexed corpus.

Model Generation For each query, the language model will generate a response based on the retrieved passages.

Citation The language model will generate fine-grained citations at the sentence end in responses. Citations are represented as bracketed numbers, such as [1][2][3]. For each sentence in a response, its citations link to a subset of retrieved passages, if there are any.

We break model answers into sentences for fine-grained evaluation. For each sentence in the answer, you will be asked to perform maximum three steps: context attribution, citation editing, and citation rating. Specifically, for each answer sentence:

1. **Context Attribution:** You will be asked to classify the sentence to one of four context types.
2. **Citation Editing:** Depending on the context you attribute the sentence to, you may be asked to edit the provided citations to make them better.
3. **Citation Rating:** You will be asked to rate the quality of the provided citations, based on the edits you may have performed in Step 2.

Your annotation will be used in a project aiming to develop metrics that better reflect citation quality.

A.2 Context Attribution

First read all passages and pay special attention to evidence for the given question and each answer sentence. Then classify each answer sentence into one of the context types shown below.

Query Sentences that iterate or rephrase the user query without making new claims or involving new facts.

Retrieval Sentences fully or partially supported by the retrieval context.

Response Sentences solely derived from preceding sentences within the response itself, not relying on the query context, the retrieval context, or the succeeding sentences in the response. Examples include sentences that perform mathematical and logical reasoning over preceding response sentences.

Model Sentences solely based on the inherent knowledge of the language model that generated the response. Knowledge is only inherent when it can NOT be found in, or reasonably inferred from, the query context, the retrieval context, or the response context. Examples include unsupported facts, and transitional expressions/summarization without any substantial claims.

A.3 Citation Editing

Perform a few edits improve the quality of the citations using your best judgment. Each edit operates on one citation, and can be delete or add with specific reasons (see Table 5). You can add an edit with the + button. Edit citations based on Editing Guidelines:

- If you think the citation is perfect, you don't need to do anything.
- Add 0 as the citation ID for facts that can NOT be found in, or reasonably inferred from, the user query, the retrieved passages, or the model response. This attributes the unsupported facts to inherent knowledge of the language model that generated the response.
- You should aim to achieve citations of the highest standard with minimal editing. After editing, all major claims in the statement should be cited.

- After editing, the citations should cite sources that are mostly helpful, when there are multiple related sources. The final citations for each sentence typically contain at most 3 citations, but there can be exceptions (e.g., if more than 3 citations all include direct and complementary supporting evidence, they should all be included).

A.4 Citation Rating

Review your edits if there is any. Based on the rating guidelines below, rate the quality of the original citations (NOT the citations after editing) from 1-5:

5 (Excellent) The sentence is fully supported by all relevant and accurate citations. There are no unnecessary, misleading, or missing citations. The citations (if present) enhance the credibility and informativeness of the sentence.

4 (Good) The sentence is mostly supported by accurate and relevant citations. One potentially relevant citation may be missing, or a slightly unnecessary citation may be present, but these do not significantly detract from the overall quality of the sentence.

3 (Fair) The sentence has some issues with citations. There might be one or few noticeable missing citation that somewhat weaken the sentence's support, or there might be several unnecessary or inaccurate citations that detract from the sentence's clarity or conciseness. Overall, the sentence's accuracy and credibility are somewhat compromised.

2 (Poor) The sentence has significant problems with citations. There might be multiple missing citations that leave that leave central claims unsupported, or there might be multiple unnecessary or inaccurate citations that significantly undermine the sentence's accuracy and credibility.

1 (Unacceptable) The sentence is completely unsupported by citations or is supported entirely by inaccurate, irrelevant, or misleading citations. The sentence is rendered misleading and unreliable.

Edit	Description
delete-mislead	Irrelevant citation. Removing this citation can avoid misleading users.
delete-substandard	Relevant citation, however another source is more helpful and should be cited instead.
delete-redundancy	Relevant citation, however other citations (for the same statement or a larger cited context) contain sufficient supporting evidence. Removing this citation can improve conciseness.
add-evidence	Existing citations lack certain required evidence, leaving the statement partially or fully unsupported. Adding this citation can fill the gap with the required evidence.
add-refinement	An existing citation is relevant but with suboptimal source quality. This new source is more helpful and should be cited instead (an existing citation should be deleted).
add-credibility	Existing citations cover all essential evidence from optimal sources. Adding the citation can further enhance response credibility.

Table 5: Citation edit actions in CiteEval and their applicable scenarios.

B Discussion: Partially-Parametric Statements

One typical scenario is the fusion of parametric knowledge and retrieval contexts in one statement. Consider a partially-parametric statement: *A hub simply repeats everything it hears, whereas a switch is a more intelligent device that can identify and direct traffic* [1]. The statement can be decomposed into the following two claims:

1. A hub simply repeats everything it hears, and
2. A switch is a more intelligent device that can identify and direct traffic.

In this case, [1] is the best possible citation from the retrieved sources and supports Claim 1. On the other hand, no retrieved source supports Claim 2 (and neither do the user and response contexts), attributing Claim 2 to the parametric context. We argue that the upper bound for this statement’s citation rating is always lower than the highest rating defined in the rating schema. Even with the best possible citations, the user will not be able fully verify the statement (i.e., the statement remains partially supported), and will likely require extra efforts for a complete fact checking (Liu et al., 2023). Also, providing any citations for partially supported statements may mislead users into trusting the whole statement, as citations naturally build credibility especially when users do not always check them (Lipson, 2011). Leaving this type of statements uncited does not resolve this issue, as it renders the statement appear to be completely unsupported, which is neither optimal for its verifiability nor credibility. CiteEval treats 0 as a special citation ID for parametric facts, and annotators or models can choose to add 0 as missing evidence

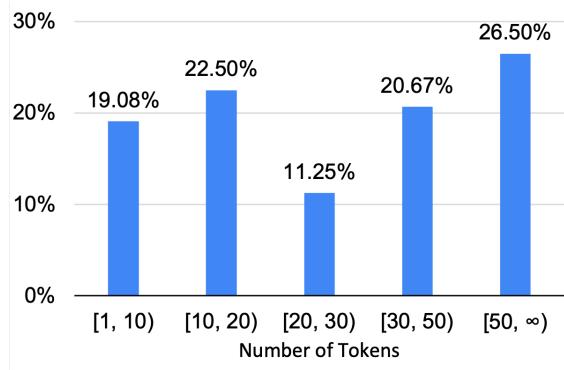


Figure 6: Length distribution of model responses in CiteBench.

when appropriate, and take it into account in the final rating (see Appendix A.3).

C CiteBench Details

C.1 Retrieval Settings

For LFRQA (Han et al., 2024), we follow the same retrieval setting which splits passages into text chunks with 100 consecutive words, and use the top 10 retrieved passages retrieved by ColBERTv2 (Santhanam et al., 2022). We randomly sample 1,000 instances from MS MARCO (Bajaj et al., 2018) which uses 10 passages from Bing logs and consists of 80% answerable queries and 20% unanswerable queries. For ASQA (Stelmakh et al., 2022) and ELI5 (Fan et al., 2019), we follow Gao et al. (2023b) and use the same subset, with top 10 passages retrieved by DPR (Karpukhin et al., 2020) and BM25, respectively.

C.2 Responses Post-processing

We remove the thinking section in model responses via matching the start token `<thinking>` and end token `</thinking>`. Figure 6 shows the response

length distribution. We split responses into statements with NLTK sentence tokenizer (version: 3.8.1).¹⁰ We extract citations from each statement with regex `\[(\d+)\]`, and keep only citations in the indices of retrieved passages [1, 10].

C.3 Dataset License

We provide license information for the datasets used in this work to construct CiteBench as follows:

- ASQA (Stelmakh et al., 2022): Apache 2.0 License, <https://github.com/google-research/language/blob/master/LICENSE>
- LFRQA (Han et al., 2024): Apache 2.0 License, <https://github.com/awslabs/rag-qa-arena?tab=Apache-2.0-1-ov-file#readme>
- ELI5 (Fan et al., 2019): BSD License, <https://github.com/facebookresearch/ELI5?tab=License-1-ov-file#readme>
- MS MARCO (Bajaj et al., 2018): CC BY 4.0 License, <https://microsoft.github.io/msmarco/LICENSE>

C.4 Annotation Details

Human annotation was performed by contracted data professionals through Summa Lingual.¹¹ The rate is \$31.5 per annotation task, which includes three blind passes for one sample, and the total cost for the official annotation batch of 1,200 samples is \$37,800. The annotation was audited and finalized by a team of full-time data linguists and scientists based in the United States.

D Further Analysis

D.1 Effects of LLM Backbones for Citation Evaluation

We show the performance of different LLMs for CITEEVAL-AUTO in Table 6.

D.2 Results on Context Attribution

Table 7 shows the precision, recall, and F1 for context attribution in predicting whether a statement is applicable for citation evaluation, where the retrieval context maps to *Applicable*, while the user, response and parametric contexts are aggregated into the same *Not Applicable* class as they

¹⁰<https://www.nltk.org>

¹¹<https://summalinguae.com/>

Model	Pearson	Spearman	Kendall-Tau
<i>CiteBench-Statement</i>			
GPT-4o	0.731	0.559	0.486
GPT-4-turbo	0.721	0.513	0.444
<i>CiteBench-Response</i>			
GPT-4o	0.668	0.589	0.492
GPT-4-turbo	0.647	0.546	0.454

Table 6: Human correlation of different LLMs for CITEEVAL-AUTO (metric test set).

Contexts	Precision	Recall	F1
Applicable	0.992	0.988	0.990
Not Applicable	0.910	0.937	0.923
Average	0.951	0.962	0.957

Table 7: Performance of model-based context attribution in CITEEVAL-AUTO (metric development set).

are treated with an identical evaluation strategy in this work.

Figure 7 further provides a breakdown of the model predictions. As can be seen, one of the major error categories for context attribution is between the parametric and retrieval contexts, which is not surprising as faithfulness evaluation and hallucination detection are challenging tasks yet to be resolved (Zhang et al., 2024a).

D.3 Benchmarking Result Analysis

We show the correlation between the response length, missing-citation ratio, and citation rating in Figure 8.

D.4 Metric Evaluation in the Cited Evaluation Scenario

We further show the performance of different evaluation metrics in the **Cited** scenario in Table 8. Consistent with the **Full** scenario, CITEEVAL-AUTO metrics achieve superior correlation with human ratings compared existing methods.

E Potential Risks

The efficiency of CITEEVAL-AUTO carries the potential risk of over-reliance on automated assessments, potentially diminishing the critical role of human judgment in fully capturing the multifaceted aspects of citation quality. To counter this, it is crucial to emphasize that CITEEVAL-AUTO is designed as a tool for efficient, scalable evaluation, not as a substitute for human expertise. CiteEval’s

Metric	Evaluator	Model	CITEBENCH-STATEMENT			CITEBENCH-RESPONSE		
			Pearson	Spearman	Kendall-Tau	Pearson	Spearman	Kendall-Tau
<i>AutoAIS-based Metrics</i>								
AUTOAIS-PRECISION		T5-XXL	—	—	—	0.187	0.065	0.062
AUTOAIS-RECALL		T5-XXL	0.227	0.136	0.122	0.119	-0.022	-0.014
AUTOAIS-F1		T5-XXL	—	—	—	0.155	0.038	0.039
AUTOAIS-PRECISION [†]		T5-XXL	0.268	0.209	0.184	0.181	0.048	0.047
AUTOAIS-F1 [†]		T5-XXL	0.253	0.202	0.178	0.153	0.032	0.034
<i>AttriScore-based Metrics</i>								
ATTRISCORE-STRICK*		GPT-4-turbo	0.459	0.281	0.254	0.196	0.079	0.097
ATTRISCORE-RELAXED*		GPT-4-turbo	0.447	0.274	0.249	0.098	0.066	0.092
ATTRISCORE-STRICK*		GPT-4o	0.449	0.297	0.269	0.221	0.094	0.108
ATTRISCORE-RELAXED*		GPT-4o	0.450	0.291	0.263	0.128	0.080	0.104
<i>LQAC-based Metrics</i>								
LQAC-PRECISION		GPT-4o	—	—	—	-0.011	-0.079	-0.046
LQAC-RECALL		GPT-4o	0.329	0.275	0.241	0.338	0.290	0.245
LQAC-F1		GPT-4o	—	—	—	0.022	-0.037	-0.011
LQAC-PRECISION [†]		GPT-4o	0.137	0.093	0.086	0.020	-0.080	-0.049
LQAC-F1 [†]		GPT-4o	0.174	0.130	0.117	0.033	-0.055	-0.027
<i>CiteEval-Auto Metrics (Ours)</i>								
CITEEVAL-AUTO (ITERCoE)		GPT-4o	0.464	0.432	0.383	0.501	0.472	0.404
CITEEVAL-AUTO (EDITDIST)		GPT-4o+MLR	0.397	0.435	0.374	0.431	0.472	0.389
CITEEVAL		GPT-4o+MLR	0.469	0.441	0.378	0.502	0.482	0.397

Table 8: Human correlation of different evaluation metrics in the **Cited** scenario (metric test set).

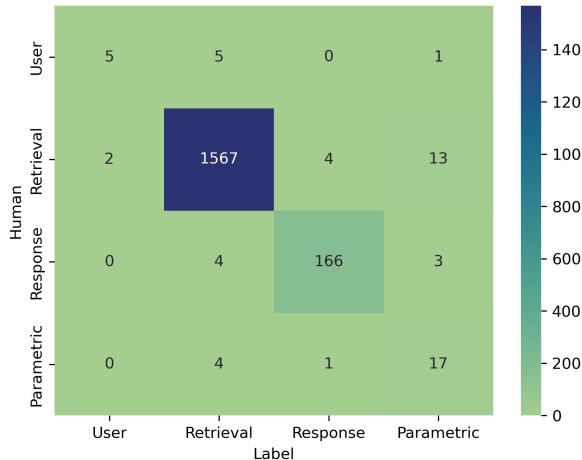


Figure 7: Confusion matrix for CITEEVAL-AUTO context attribution (metric development set).

principle-driven nature is intended to foster critical examination and iterative refinement, ultimately ensuring that human expertise remains central to the comprehensive assessment of citation quality, particularly in high-stakes applications.

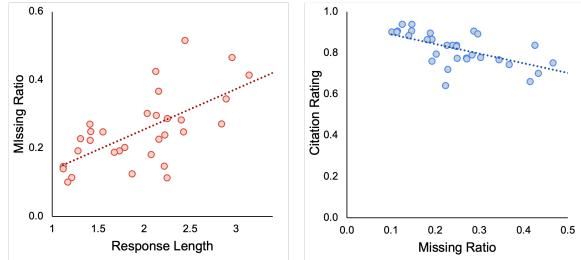


Figure 8: Correlation analysis between response length and missing citation ratio (left; Pearson correlation 0.679, $p < .001$), and missing citation ratio and citation rating (right; Pearson correlation -0.633 , $p < .001$). Each data point represents the averaged results for one model-dataset pair from the full CiteBench test set.

F Prompt Templates

F.1 Prompt Template for RAG Response Generation with Citations

We show the prompt template for retrieval-augmented response generation with citations in Table 9.

RAG Response Generation with Citations

Provide an answer to the question using information from the given passages. Passages are provided inside the <passage> </passage> XML tags. Question is provided inside the <question> </question> XML tags.

Add passage id in brackets at the end of each answer sentence to cite passages in <passage> for any factual claim. Don't use "[passage [1]]" when citing. Instead use solely passage id in brackets such as [1]. When citing several passages, use [1][2][3]. For each sentence in your answer that contains factual claims, cite at least one passage and at most three passages.

Below are the passages. Each passage has an id for citation:

[[Retrieved Passages Go Here]]

Below is the question:

<question>

[[User Query Goes Here]]

</question>

Now answer the question using only information from the passages. Think step by step first and put your thinking process into <thinking> </thinking> tags. The thinking process should not exceed 50 words. Provide the final answer after the thinking process. Remember to do citation for the final answer using bracketed numbers at sentence end. In your final answer, do not use expressions similar to "Passage 1", "Passage [1]", "according to Passage [1]" to show your thought process or justify your citations in your answer. Remember that you need to say "No answer is found" if the question cannot be answered by information in the passages.

Table 9: Prompt template for RAG response generation with citations.

F.2 Prompt Template for Context Attribution

We show the prompt template for context attribution in Table 10.

F.3 Prompt Template for Citation Editing and Rating

We show the prompt template for joint citation editing and rating in Table 11.

Context Attribution

You are an expert specializing in analyzing sentences within a given model response and classifying them based on their attribution.

Your task is to carefully examine each sentence, and attribute it to one of the following categories:

<categories>

1. Query: Sentences that iterate or rephrase the user query without making new claims or involving new facts.

2. Retrieval: Sentences fully or partially supported by the retrieval context.

3. Response: Sentences solely derived from preceding sentences within the response itself, not relying on the query context, the retrieval context, or the succeeding sentences in the response. Examples include sentences that perform mathematical and logical reasoning over preceding response sentences.

4. Model: Sentences solely based on the inherent knowledge of the language model that generated the response. Knowledge is only inherent when it can NOT be found in, or reasonably inferred from, the query context, the retrieval context, or the response context. Examples include unsupported facts, and transitional expressions/summarization without any substantial claims.

<categories>

Follow the guidelines below for ambiguous cases:

<ambiguous_cases>

- For sentences involving both the retrieval context and other types of contexts, choose 2 (Retrieval).
- For single-sentence responses indicating that no answer could be found, choose 3 (Response).
- For sentences supported by its succeeding sentences but not its preceding sentences, choose from 1 (Query), 2 (Retrieval) and 4 (Model).

</ambiguous_cases>

Below is the query:

```
<query>
[[User Query]]
</query>
```

Below is the retrieval context, consisting of documents retrieved for the query:

```
<retrieval>
[[Retrieved Passages]]
</retrieval>
```

Below is the response, consisting of the sentences to evaluate:

```
<response>
[[Response Sentences Go Here]]
</response>
```

From now on you must follow this format:

```
<thinking> Think step by step first before classifying sentence 1 </thinking>
<category sentence_id="1"> Choose the attribution of sentence 1 from 1, 2, 3, 4 </category>
<thinking> Think step by step first before classifying sentence 2 </thinking>
<category sentence_id="2"> Choose the attribution of sentence 2 from 1, 2, 3, 4 </category>
...
<thinking> Think step by step first before classifying sentence N </thinking>
<category sentence_id="N"> Choose the attribution of sentence N from 1, 2, 3, 4 </category>
```

Begin!

Table 10: Prompt template for context attribution in CITEEVAL-AUTO.

Citation Editing and Rating

You are an expert specializing in analyzing, editing, and rating citations for sentences within a given model response.

Your task is to carefully examine the citations for each sentence, provide critical editing to the citations, and rate the citation quality.

You are allowed to use a sequence of DELETE or ADD edits for critical editing. Each edit operates on one citation.

```
<edits>
DELETE: You can delete a citation due to the following reasons:
DELETE REASON 1. Misleading: the citation is irrelevant, and removing this citation avoids misleading users.
DELETE REASON 2. Substandard: the citation is relevant, however another source is more helpful and should be cited instead.
DELETE REASON 3. Redundant: the citation is relevant, however other citations contain sufficient supporting evidence. Removing this citation improves conciseness.
```

ADD: You should only add a citation due to the following reasons:

```
ADD REASON 1. Evidence: existing citations lack certain required evidence, leaving the statement partially or fully unsupported. Adding this citation fills the gap with the required evidence.
ADD REASON 2. Refinement: an existing citation is relevant but substandard. This new source is more helpful and should be cited instead (an existing citation should be deleted).
ADD REASON 3. Credibility: existing citations cover all essential evidence from optimal sources. Adding this citation further enhances response credibility.
```

```
</edits>
```

Each edit should be passed in as `<edit_name citation="{{citation}}">{{reason}}<{{edit_name}}>`, where `edit_name` is the name of the specific edit (DELETE or ADD), `{{citation}}` is a citation id to be deleted or added, and `{{reason}}` is one of the reasons from `<edits></edits>`.

You should replace `{{edit_name}}`, `{{citation}}` and `{{reason}}` with the appropriate value.

Below are the editing guidelines. Follow the guidelines when deciding whether and how to perform an edit.

```
<editing_guidelines>
- Use N/A if no editing is needed.
- Add 0 as the citation id for facts that can NOT be found in, or reasonably inferred from, the query context, the retrieval context, or the response context. This attributes the unsupported facts to inherent knowledge of the language model that generated the response.
- You should aim to achieve citations of the highest standard with minimal editing. After editing, all major claims in the statement should be cited.
- After editing, the citations should cite sources that are mostly helpful, when there are multiple related sources. The final citations for each sentence typically contain at most 3 citations, but there can be exceptions.
</editing_guidelines>
```

After providing edits, rate the original citations for each sentence, following the guidelines below:

```
<rating_guidelines>
- 5 (Excellent): The sentence is fully supported by all relevant and accurate citations. There are no unnecessary, misleading, or missing citations. The citations (if present) enhance the credibility and informativeness of the sentence.
- 4 (Good): The sentence is mostly supported by accurate and relevant citations. One potentially relevant citation may be missing, or a slightly unnecessary citation may be present, but these do not significantly detract from the overall quality of the sentence.
- 3 (Fair): The sentence has some issues with citations. There might be one or few noticeable missing citation that somewhat weaken the sentence's support, or there might be several unnecessary or inaccurate citations that detract from the sentence's clarity or conciseness. Overall, the sentence's accuracy and credibility are somewhat compromised.
- 2 (Poor): The sentence has significant problems with citations. There might be multiple missing citations that leave that leave central claims unsupported, or there might be multiple unnecessary or inaccurate citations that significantly undermine the sentence's accuracy and credibility.
- 1 (Unacceptable): The sentence is completely unsupported by citations or is supported entirely by inaccurate, irrelevant, or misleading citations. The sentence is rendered misleading and unreliable.
</rating_guidelines>
```

Table 11: Prompt template for citation editing and rating in CITEEVAL-AUTO.

Citation Editing and Rating

Below is a hypothetical example.

```
<example>
Given 10 passages related to the question "Can you explain the concept of time dilation in the context of special relativity?", and a response which has the following sentence and citations: <citation sentence_id="1", sentence="Time dilation occurs because the speed of light in a vacuum is constant for all observers, regardless of their relative motion."> 1, 6 </citation>
```

The following example shows how you should improve the citations for this sentence:

```
<thinking> This claim is directly supported by passage 1. However, passage 6 does not provide any direct evidence to the question, so I should delete it to avoid misleading users. Additionally, passage 7 clearly states that time dilation occurs due to the constant speed of light in a vacuum. It will constitute to a good citation, so I will add 7 for credibility. Based on these edits, I will rate the given citations 2 (Poor). </thinking>
<editing sentence_id="1">
<DELETE citation="6"> DELETE REASON 1 </DELETE>
<ADD citation="7"> ADD REASON 3 </ADD>
</editing>
<rating sentence_id="1"> 2 </rating>
</example>
```

Below is the query:

```
<query>
[[User Question Goes Here]]
</query>
```

Below are the retrieved sources. Each source passage <passage> </passage> has an id for citation.

```
<retrieval>
[[Retrieved Passages Go Here]]
</retrieval>
```

Below is the response:

```
<response>
[[Response Goes Here]]
</response>
```

Below are the citations to evaluate. Each <citation> has a response sentence and its sentence id that it cites for.

```
<citations>
[[Citations Go Here]]
</citations>
```

From now on you must follow this format:

```
<thinking> Think step by step first before editing citations for sentence 1. </thinking>
<editing sentence_id="1"> edits for citations in sentence 1, or N/A if no editing is needed
</editing>
<rating sentence_id="1"> rating for citations in sentence 1, from 1 - 5 </rating>
<thinking> Think step by step first before editing citations for sentence 2. </thinking>
<editing sentence_id="2"> edits for citations in sentence 2, or N/A if no editing is needed
</editing>
<rating sentence_id="2"> rating for citations in sentence 2, from 1 - 5 </rating>
...
<thinking> Think step by step first before editing citations for sentence N. </thinking>
<editing sentence_id="N"> edits for citations in sentence N, or N/A if no editing is needed
</editing>
<rating sentence_id="N"> rating for citations in sentence N, from 1 - 5 </rating>
```

Begin!

Table 11: Continued.