

Concise and Sufficient Sub-Sentence Citations for Retrieval-Augmented Generation

Guo Chen^{1,2}, Qiuyuan Li^{1,2}, Qiuxian Li^{1,2}, Hongliang Dai^{1,2,†}, Xiang Chen³, Piji Li^{1,2}

¹College of Artificial Intelligence,

Nanjing University of Aeronautics and Astronautics, Nanjing, 211106, China

²The Key Laboratory of Brain-Machine Intelligence Technology,

Ministry of Education, Nanjing, 211106, China

³ College of Computer Science and Technology,

Nanjing University of Aeronautics and Astronautics, Nanjing, 211106, China

{guochen, qiuyli, 162350227, hongldai, xiang_chen, pjli}@nuaa.edu.cn

Abstract

In retrieval-augmented generation (RAG) question answering systems, generating citations for large language model (LLM) outputs enhances verifiability and helps users identify potential hallucinations. Existing citation generation methods for retrieval-augmented generation usually provide evidence at the paragraph or sentence level. Although such citations improve verifiability, they may still impose unnecessary reading burden on users: a cited sentence may include information irrelevant to the answer while lacking sufficient context, such as a clear referent for the subject. In this paper, we study concise and sufficient sub-sentence citation generation, aiming to identify minimal yet self-contained evidence spans for answer verification. To this end, we first develop annotation guidelines for such citations and construct a corresponding dataset. To reduce annotation cost, we propose a data augmentation framework that generates synthetic citation annotations and filters low-quality instances using an evaluator model. Our experiments on the constructed dataset demonstrate that the proposed approach can generate high-quality and more readable citations.

1 Introduction

Large language models (LLMs) are increasingly used for question answering and information-seeking tasks, but they still suffer from hallucinations, producing plausible statements that are not fully supported by reliable evidence (Azamfirei et al., 2023; Ye et al., 2024; Ji et al., 2023). Retrieval-augmented generation (RAG) mitigates

this issue by conditioning LLMs on external documents or retrieved passages (Lewis et al., 2020). However, retrieved contexts can be long, noisy, or only partially relevant, making it difficult for users to efficiently verify generated answers. To reduce this burden, recent work has explored generating citations that explicitly link model responses to supporting evidence (Menick et al., 2022; Liu et al., 2023; Gao et al., 2023; Zhang et al., 2025; Shen et al., 2025; Chuang et al., 2025).

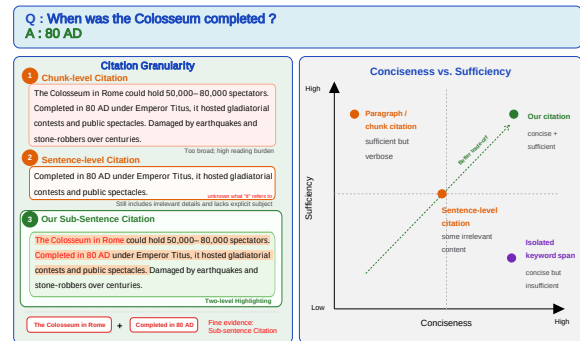


Figure 1: Motivation of concise and sufficient sub-sentence citations. Paragraph- or chunk-level citations are often sufficient but verbose, while sentence-level citations may still contain irrelevant details or miss necessary context. Our sub-sentence citation standard aims to select minimal and self-contained evidence spans for efficient answer verification.

Existing citation methods have made important progress, but their citation granularity remains limited for efficient verification. Early systems often cite entire web pages, paragraphs, or fixed-size chunks, requiring users to locate the exact evidence. More recent work, such as LongCite, improves citation granularity by generating sentence-level citations for long-context question answering (Zhang et al., 2025). However, a sentence is not always

[†]Corresponding author.

the optimal unit of reference: it may include irrelevant information, while also being insufficient on its own due to pronouns, implicit subjects, or context-dependent expressions. Therefore, a desirable citation should be fine-grained, concise, and sufficient, excluding irrelevant content while retaining the information necessary for verification. Figure 1 illustrates our motivation.

In this work, we study concise and sufficient sub-sentence citation generation for question answering. Given a question, a retrieved context, and an answer, the goal is to identify one or more evidence spans that are minimal, self-contained, and, where possible, naturally readable on their own. Compared with sentence-level citation, our task allows citations to include only the necessary clause or phrase, while adding antecedent spans when required for sufficiency. It is useful when a complete sentence contains both answer-relevant evidence and distracting background information, because users can verify the answer without reading unnecessary surrounding content.

To support this task, we develop annotation principles and construct a manually annotated benchmark covering cases where sentence-level citations are sufficient, cases with redundant sentence content, and cases requiring antecedent or multiple spans. To reduce annotation cost, we further propose a data augmentation framework that generates synthetic citation annotations, filters low-quality instances with an evaluator model, and trains the final citation generator through synthetic pre-training followed by annotated data refinement.

Experiments on our benchmark show that the proposed framework consistently improves sub-sentence citation generation across multiple open-source LLMs. Further analyses highlight the benefits of data quality control and of combining synthetic and human annotations. Our code and data are available at <https://github.com/chen-kek/cs-citation>.

Our contributions are summarized as follows:

- We formulate the task of concise and sufficient sub-sentence citation generation for RAG-based question answering, aiming to reduce verification burden beyond paragraph- and sentence-level citations.
- We develop annotation principles for sub-sentence citations and construct a manually annotated benchmark covering different citation scenarios.

- We propose a data augmentation framework that combines synthetic citation generation, evaluator-based filtering, and two-stage fine-tuning.
- Experiments and ablation studies show that the proposed framework improves sub-sentence citation quality, and that data filtering and the training strategy both contribute to the final performance.

2 Related Work

Generating responses with citations has become an important direction for improving the verifiability and trustworthiness of large language models. Early work on retrieval-augmented and web-based question answering encourages models to support generated answers with retrieved documents or external evidence (Nakano et al., 2021; Menick et al., 2022; Liu et al., 2023). More recently, ALCE (Gao et al., 2023) introduced a reproducible benchmark for automatic citation evaluation, requiring systems to retrieve supporting passages and generate answers with citations. This line of work demonstrates that citations can help users verify model outputs and diagnose unsupported claims. However, most citation-enabled generation systems still operate at a relatively coarse granularity, such as web pages, passages, paragraphs, or fixed-size chunks. Although these citations provide useful provenance, users often still need to inspect a large amount of text to locate the exact evidence.

Recent studies have explored more fine-grained citation generation. LongCite (Zhang et al., 2025) focuses on long-context question answering with sentence-level citations and constructs LongBench-Cite and LongCite-45K through a coarse-to-fine data construction pipeline. ReClaim (Xia et al., 2025) also studies fine-grained attributed generation by interleaving reference selection and claim generation, enabling models to generate long-form answers with sentence-level citations. These works substantially improve citation granularity compared with passage- or chunk-level attribution, and show that sentence-level citations can make generated answers more verifiable. Nevertheless, sentence-level citation is not always the most efficient unit for user verification. A cited sentence may contain irrelevant clauses or background descriptions, while the answer-bearing part may require additional antecedent information to be self-contained. Our work follows the same general

motivation of improving verifiability, but studies whether citations can be refined from complete sentences to concise and sufficient sub-sentence spans.

Another related direction is post-hoc attribution and explanation for language model outputs. Perturbation-based and gradient-based methods estimate which input tokens, sentences, or passages contribute to a generated statement by observing how the output changes when parts of the context are modified (Lei et al., 2016; Samek and Müller, 2019; Bastings and Filippova, 2020; Anand et al., 2022). RAG-Ex (Sudhi et al., 2024) explores model-independent explanations for retrieval-augmented generation by perturbing the input context and comparing the resulting model outputs. ContextCite (Cohen-Wang et al., 2024) formulates context attribution as identifying the parts of the context responsible for a generated statement, and uses context ablation to estimate attribution scores. SelfCite (Chuang et al., 2025) further leverages context ablation to construct self-supervised rewards for citation quality, measuring whether cited evidence is necessary and sufficient for generating a statement. These methods are closely related to our motivation, especially in their emphasis on identifying supporting evidence. However, attribution scores or scattered highlighted tokens are not always suitable as user-facing citations, since they may be fragmented and difficult to read. In contrast, our goal is to generate coherent sub-sentence evidence spans that follow natural reading patterns and can be directly presented to users.

There is also emerging work on more fine-grained or sub-sentence-level attribution. For example, LAQuer studies localized attribution queries in content-grounded generation, where users can query support for specific spans or fragments (Hirsch et al., 2025). Such work is closely related to ours in that it moves beyond sentence-level evidence. However, existing studies often focus on locating support for user-selected spans or generated claims, whereas our setting aims to automatically produce concise and sufficient cited content for question-answering outputs. In our formulation, a good citation should exclude irrelevant text while retaining the necessary information for verification, including antecedents or multiple evidence spans when required. Therefore, our work complements prior citation generation and attribution methods by studying a more fine-grained but still readable citation unit for retrieval-augmented generation.

Recent work also explores fine-grained attribu-

tion and provenance. Attribute First, then Generate selects concise source segments before generation (Slobodkin et al., 2024). ALiCE studies positional fine-grained citation generation (Xu et al., 2025), while Ding et al. incorporate dependency parsing for fine-grained attribution (Ding et al., 2025). TROVE traces generated text to source sentences for fine-grained provenance (Zhu et al., 2025). These works complement our focus on minimal, self-contained evidence spans.

3 Benchmark Construction

This section describes how we construct the benchmark for concise and sufficient sub-sentence citation generation. We first define the task and citation criteria, and then introduce the human annotation process. The construction of synthetic training data is described separately in Section 4. Figure 2 provides an overview of the overall pipeline.

3.1 Task Definition and Citation Criteria

Following prior work that evaluates long-form generations by decomposing them into sentence-level or atomic factual units (Min et al., 2023; Tang et al., 2024), we focus on question-answering instances whose answers can be verified by one or more spans in the retrieved context. Given a question q and a context C , the model generates an answer a together with supporting spans $E = \{e_1, e_2, \dots, e_m\}$ from C that verify the generated answer. Each e_i is an exact text span from the context and is associated with its location. Unlike sentence-level citation methods (Zhang et al., 2025; Xia et al., 2025), our task selects finer-grained evidence spans while keeping them readable and self-contained. This formulation emphasizes evidence selection rather than free-form explanation generation. The predicted citation spans must be copied from the retrieved context, which makes the output directly verifiable and avoids introducing unsupported statements.

In practical interfaces, sub-sentence citations need not replace sentence-level context. The sentence containing the predicted span can be highlighted as coarse contextual evidence, while the selected sub-sentence span can be highlighted as the precise supporting evidence. Such a two-level highlighting scheme preserves sentence-level readability and helps users quickly locate the exact information needed for verification.

We define a high-quality sub-sentence citation

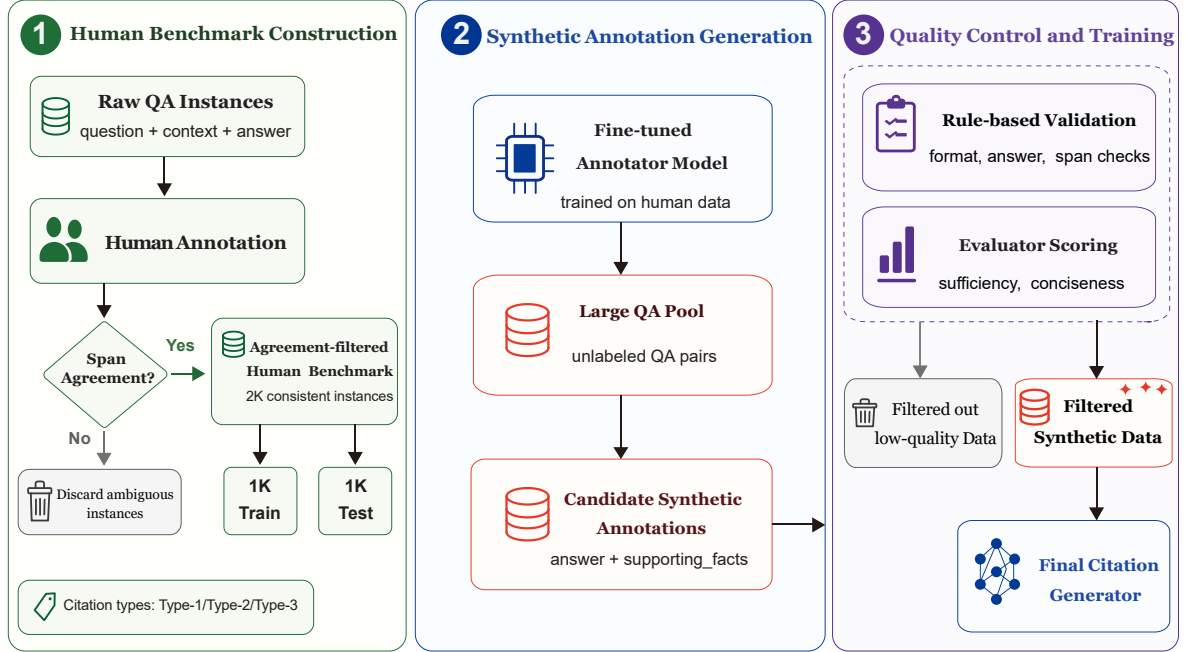


Figure 2: Overview of our benchmark construction and synthetic data filtering pipeline. We obtain reliable human annotations through double annotation and agreement filtering, then use the human training split to generate and filter synthetic annotations for training the final citation generator.

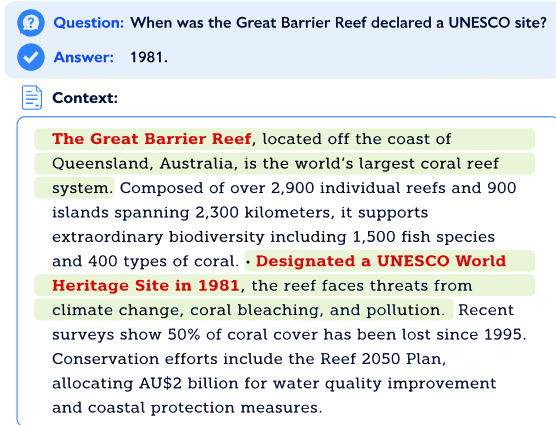


Figure 3: Sub-sentence-level citations adhering to our annotation principles.

using three criteria. First, it should be *sufficient*, containing enough information to verify the answer without requiring surrounding context. Second, it should be *concise*, excluding irrelevant clauses or background descriptions. Third, it should be *readable*, forming coherent phrases or clauses rather than isolated keywords. This distinguishes our task from post-hoc attribution methods that may highlight scattered tokens or assign importance scores (Lei et al., 2016; Bastings and Filippova, 2020; Sudhi et al., 2024; Cohen-Wang et al., 2024). For example, selecting only a date may identify the answer, but omit the event or subject it refers to.

Based on these criteria, we categorize instances into three types. **Type-1** instances are cases where a complete sentence is already concise and sufficient. **Type-2** instances contain irrelevant content in the supporting sentence, so only the answer-bearing part should be cited. **Type-3** instances require additional antecedent information or multiple supporting spans because the answer-bearing span alone is not self-contained. Figure 3 shows an example following these criteria, where the selected spans retain the necessary subject and evidence while excluding irrelevant surrounding content.

3.2 Human Annotation

We construct our benchmark from publicly available question-answering datasets, including XOR QA (Asai et al., 2021), SQuAD (Rajpurkar et al., 2016), and HotpotQA (Yang et al., 2018). XOR QA provides cross-lingual open-retrieval instances, SQuAD provides paragraph-level reading-comprehension examples, and HotpotQA contains multi-hop questions requiring supporting facts. These datasets complement one another in discourse structure, diversity, and reasoning depth.

For each candidate instance, annotators are given the question, context, and reference answer, and are asked to select minimal supporting spans that are sufficient for verification. The spans must be copied from the context and should avoid under-

citation, where necessary evidence is missing, and over-citation, where irrelevant content is included. Annotators were also encouraged to preserve natural phrase boundaries when possible, so that the selected spans remain readable after being shown independently. This helps avoid fragmented citations that contain the answer string but are difficult for users to interpret.

To improve reliability, each instance is independently annotated by two annotators following the same guidelines. Two annotations are considered in agreement when their selected supporting spans match exactly after whitespace normalization. Under this criterion, the annotators achieve 95% agreement on span selection. Instances with inconsistent citation spans are discarded rather than adjudicated, so the final manually annotated benchmark only includes examples on which both annotators agree. This conservative agreement-based filtering reduces ambiguity and improves the reliability of the benchmark.

After double annotation and agreement filtering, we obtain 2,000 instances. Each instance contains the question, context, reference answer, supporting facts, and citation type label. Type-1, Type-2, and Type-3 account for 20%, 30%, and 50% of the dataset, respectively. We split the benchmark into 1,000 training instances and 1,000 test instances.

4 Data Augmentation

Because sub-sentence citation annotation is costly, we construct synthetic annotations to expand the training data. Beyond increasing data scale, synthetic annotations also help cover more diverse citation patterns, including single-span evidence, shortened sentence-level evidence, and cases requiring antecedent information. Inspired by recent work on model-generated citation or attribution supervision (Gao et al., 2023; Zhang et al., 2025; Chuang et al., 2025), including preference-based optimization for attributed generation (Li et al., 2024) and synthetic-data construction for context attribution (Radevski et al., 2025), we first fine-tune a base LLM on the human-annotated training set.

Automatically generated annotations may contain incorrect answers, unsupported citations, incomplete evidence, overly long spans, or formatting errors. We therefore apply a two-stage filtering process. First, rule-based validation checks output format, answer consistency, citation validity, citation length, and citation count, and filters instances

according to the resulting rule-based score. Second, an evaluator model scores the remaining instances according to factual fidelity, answer correctness, sufficiency, conciseness, and coherence. These dimensions are related to recent work on citation sufficiency and context attribution (Chuang et al., 2025; Cohen-Wang et al., 2024).

We combine the rule-based validation result and evaluator score into a final quality score, and retain high-quality synthetic instances for training. The final citation generator is first trained on the filtered synthetic data and then refined on the human-annotated training set. This design separates broad pattern learning from final calibration. Filtered synthetic data exposes the model to diverse citation patterns, while the human-annotated data helps align the model with the target standards of conciseness, sufficiency, and readability.

5 Experiments

5.1 Experimental Setup and Evaluation Metrics

We evaluate our method on the 1K human-annotated test set introduced in Section 3. Given a question and its retrieved context, each model is required to generate both the answer and the corresponding supporting facts. The supporting facts are expected to be concise and sufficient sub-sentence citation spans extracted from the context.

We evaluate three open-source LLMs: Llama-3.1-8B, Qwen2.5-7B, and GLM-4-9B. For each model, we compare four training settings. *Base* denotes the original model without fine-tuning, where the target output format is induced only through prompting. *Manual Only* denotes fine-tuning only on human annotations. *Synthetic Only* denotes fine-tuning only on automatically generated annotations. *Full Method* denotes our complete training setting, where the model is first trained on filtered synthetic annotations and then refined on human annotations.

We additionally evaluate GPT-4o with few-shot prompting as a strong instruction-following baseline. The prompt also uses 10 annotated training examples (Brown et al., 2020) and requires the model to output a JSON object containing the answer and supporting facts; the full prompt template is provided in Appendix D.

We evaluate citation quality with Precision (P), Recall (R), and F1. Each supporting fact is represented as a tuple of a text span and a line identifier. To account for localization and span quality, a pre-

Model	Strategy	Performance Metrics					
		R	P	F1	ROUGE-L	BLEU	CL
Llama-3.1-8B	Base	26.85	18.09	20.61	58.49	43.87	142.21
	Synthetic Only (10K)	77.81	79.33	77.86	86.40	78.35	69.15
	Manual Only (0.5K)	75.28	76.13	74.84	85.51	77.68	70.07
	Full Method (10K+0.5K)	78.10	78.30	77.53	86.17	77.90	65.12
	Manual Only (1K)	77.23	78.50	77.15	86.26	77.97	67.56
	Full Method (10K+1K)	81.20	80.98	80.46	87.75	80.72	66.72
Qwen2.5-7B	Base	27.57	28.71	27.89	63.83	50.09	137.07
	Synthetic Only (10K)	77.80	79.17	77.67	86.28	78.23	68.30
	Manual Only (0.5K)	71.98	72.53	71.39	83.80	75.54	71.34
	Full Method (10K+0.5K)	78.29	78.19	77.56	85.37	76.97	67.21
	Manual Only (1K)	75.46	75.55	74.74	84.95	76.75	70.05
	Full Method (10K+1K)	78.58	77.46	77.81	86.01	77.64	67.52
GLM-4-9B	Base	34.47	35.77	34.74	59.63	47.81	128.32
	Synthetic Only (10K)	73.91	74.76	73.70	83.58	75.38	75.20
	Manual Only (0.5K)	69.98	70.05	69.30	82.65	74.17	75.67
	Full Method (10K+0.5K)	71.41	71.11	70.57	79.50	70.90	72.86
	Manual Only (1K)	71.42	72.57	71.99	82.45	73.44	73.74
	Full Method (10K+1K)	74.23	73.58	73.90	83.45	75.57	71.12

Table 1: Citation generation performance of Llama-3.1-8B, Qwen2.5-7B, and GLM-4-9B under different fine-tuning methods and settings.

dicted citation can match a reference citation only if they share the line identifier. For each eligible pair, span similarity is computed as character 3-gram overlap F1 after lowercasing, punctuation removal, and whitespace normalization.

Precision and Recall are computed by averaging best-match similarity scores over predicted and reference citations, respectively, and citation F1 is their harmonic mean. This best-match formulation supports multiple citation spans without requiring a fixed span ordering.

We additionally report ROUGE-L and BLEU between the concatenated predicted and reference citation texts as auxiliary overlap metrics. ROUGE-L is computed using token-level longest common subsequence F1, and BLEU is computed with add-one smoothing up to 4-grams. To quantify conciseness, we report citation length (CL), defined as the average number of normalized non-space characters in generated citation spans. A good citation model should achieve high citation F1 while keeping CL reasonably low. Because automatic overlap metrics cannot fully capture readability and sufficiency, we further provide qualitative analysis in Appendix B. Table 1 reports the main results across three base models and different training settings.

5.2 Main Results and Implementation Details

All models are fine-tuned with LoRA under comparable settings. We use the same input-output format for all models: the input consists of the question and retrieved context, and the output contains the answer and supporting facts. Unless otherwise specified, synthetic training uses 10K evaluator-filtered synthetic instances, and manual refinement uses either 0.5K or 1K human-annotated instances. For the full method, we first fine-tune the model on filtered synthetic data and then continue fine-tuning on the human-annotated training set.

We use the AdamW optimizer with an initial learning rate of 5×10^{-5} and a cosine learning rate scheduler. Models are trained for 4 epochs with gradient clipping using a maximum norm of 1.0. To improve training efficiency, we use bf16 mixed-precision training. Experiments are conducted with a per-GPU batch size of 2 and 8 gradient accumulation steps. For LoRA, we set the rank $r = 8$, scaling factor $\alpha = 32$, and dropout to 0. We also adopt the LoRA+ optimizer setting with a B-matrix learning rate ratio of 16.

The results show several clear trends. First, the base models perform poorly on sub-sentence citation generation. Although they can often generate

fluent answers, they frequently fail to follow the required citation format, cite overly long context, or miss necessary supporting spans. For example, Llama-3.1-8B achieves only 20.61 F1 in the base setting, showing that prompting alone is insufficient for this fine-grained citation task.

Second, both manual annotations and synthetic annotations are useful for fine-tuning. Compared with the base setting, manual-only training improves citation F1 and reduces citation length. For Llama-3.1-8B, 1K manual annotations improve F1 from 20.61 to 77.15. Synthetic-only training also achieves strong performance, suggesting that large-scale synthetic annotations can provide useful supervision when their quality is properly controlled. This result also suggests that the task contains learnable structural patterns, such as locating answer-bearing clauses and preserving nearby antecedents, which can be transferred from synthetic annotations to the final model.

Third, the full method generally achieves the best F1 performance, although synthetic-only training obtains slightly higher scores on some auxiliary overlap metrics. On Llama-3.1-8B, adding filtered synthetic data to 1K human annotations improves F1 from 77.15 to 80.46. On Qwen2.5-7B, the full method improves F1 from 74.74 to 77.81 compared with manual-only training. On GLM-4-9B, the full method improves F1 from 71.99 to 73.90. These results suggest that filtered synthetic annotations and human annotations play complementary roles: synthetic data provides broader coverage of citation patterns, while human annotations calibrate the model toward the target citation standard.

Finally, citation length decreases after fine-tuning. The base models produce long citations, with CL values above 120 for all three models. Fine-tuned models reduce CL to around 65–75 while achieving much higher F1. This indicates that the models learn not only to locate relevant evidence, but also to avoid excessive citation content.

5.3 Effectiveness of Evaluator-based Filtering

Synthetic annotations can reduce manual annotation cost, but they may also introduce noise. Automatically generated annotations may contain incorrect answers, unsupported citations, incomplete evidence, overly long spans, or formatting errors. To examine whether filtering is necessary, we compare synthetic annotations with and without evaluator-based filtering. The results are shown in Table 2.

Training on unfiltered synthetic annotations per-

Setting	F1
Manual Only 1K	77.15
Synthetic 10K w/o filtering	73.24
Synthetic 10K w/ evaluator filtering	77.86
Synthetic 10K w/ evaluator filtering + Manual 1K	80.46

Table 2: Effect of evaluator-based filtering on synthetic annotations.

forms worse than manual-only training, with F1 dropping from 77.15 to 73.24. This shows that increasing the amount of training data is not sufficient for sub-sentence citation generation. Since the task requires fine-grained judgments about conciseness and sufficiency, noisy synthetic annotations can mislead the model during supervised fine-tuning.

After applying evaluator-based filtering, training improves from 73.24 to 77.86 F1 and surpasses the manual-only setting. This indicates that the evaluator can remove low-quality synthetic instances while preserving useful supervision. Combining evaluator-filtered synthetic annotations with 1K manual annotations improves F1 to 80.46, achieving the best result. These results confirm that quality control is essential for making synthetic citation data effective.

5.4 Additional Analyses

Impact of Training Data Size. We analyze how the amount of synthetic and human annotation data affects model performance. Figure 4 shows the effect of increasing the number of filtered synthetic instances from 2K to 10K. The F1 score steadily improves as more synthetic annotations are added, increasing from 76.23 at 2K to 77.86 at 10K. However, the improvement becomes marginal after 8K, suggesting that the benefit of synthetic data gradually saturates. This indicates that data scale is useful, but quality and diversity remain important.

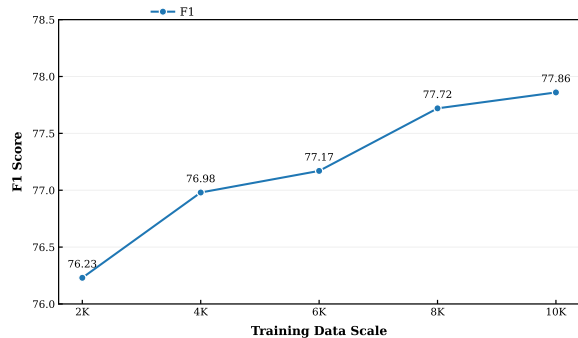


Figure 4: Impact of synthetic data scale on model performance.

Figure 5 compares the effect of adding different amounts of human-annotated data with and without synthetic pre-training. The results show that synthetic data is helpful in low-resource settings. With 200 human annotations, training with 10K synthetic data achieves 76.68 F1, while using only manual data achieves 71.35 F1. As the amount of annotated data increases, the gap becomes smaller, but the model trained with synthetic data remains stronger overall. This supports our design choice of using synthetic data for broad pattern learning and human annotations for final calibration. The trend suggests that increasing the number of synthetic instances is not the only factor behind the improvement. After the model has seen enough examples, data brings smaller gains, indicating that the remaining errors may depend more on citation difficulty and annotation quality than on data scale.

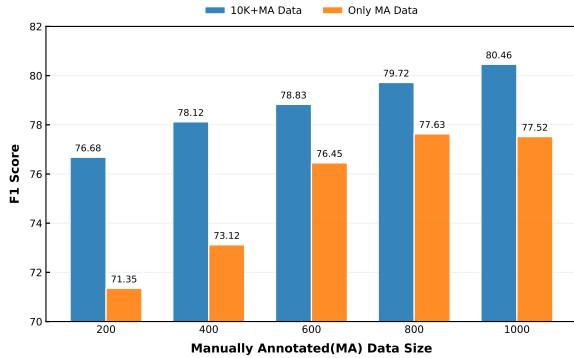


Figure 5: Impact of manually annotated data scale on the performance of models trained with and without synthetic data.

Comparison with Existing Methods. We compare our method with representative citation and attribution approaches, including LongCite, ALCE, RAG-Ex, ContextCite, and LAQuer, as well as strong instruction-following LLMs under few-shot prompting. We further construct two refinement baselines to examine whether sub-sentence citations can be obtained by simply shortening sentence-level citations. Specifically, Qwen and GPT-4o are prompted to refine the sentence-level citations generated by LongCite.

As shown in Table 3, our method achieves the highest citation F1 among all compared methods. LAQuer produces slightly shorter citations, but obtains substantially lower F1, indicating that compact evidence alone does not guarantee sufficient support. Strong prompting models also partially follow the sub-sentence citation format, but remain

Method	F1	CL
LongCite-Llama-3.1-8B	45.29	248.96
ALCE	52.46	–
RAG-Ex 2.0	58.91	–
ContextCite	47.70	143.28
LAQuer	70.04	58.97
LongCite + Qwen refine	51.20	94.62
LongCite + GPT-4o refine	53.61	73.86
GPT-4o prompting	65.26	–
GPT-5.4 prompting	67.42	70.60
Qwen-3.6 prompting	65.16	65.14
Ours	80.46	66.72

Table 3: Comparison with citation, attribution, refinement, and strong prompting baselines under our sub-sentence citation evaluation. CL denotes citation length; lower is better.

Type	#	F1	CL
Type-1	176	88.40	79.94
Type-2	305	79.94	62.64
Type-3	519	66.91	67.23

Table 4: Performance across different citation types.

below our fine-tuned model.

Importantly, refining LongCite substantially reduces citation length but yields only modest F1 improvements: its F1 increases from 45.29 to 51.20 with Qwen refinement and 53.61 with GPT-4o refinement, still far below our 80.46. This suggests that the task cannot be reduced to simply shortening sentence-level citations. Producing a concise and sufficient citation also requires preserving subjects, antecedents, or multiple evidence spans when necessary.

Type-wise Analysis. We further analyze performance across the three citation types defined in Section 3.1. As shown in Table 4, Type-1 is the easiest, while Type-3 is substantially more challenging because these examples often require antecedents, explicit subjects, or multiple supporting spans. Notably, Type-3 constitutes the largest portion of our benchmark, so the lower performance on this category also explains a substantial fraction of the remaining overall errors. Improving antecedent recovery and multi-span evidence selection is therefore a promising direction for future work.

The clear performance drop from Type-1 to Type-3 suggests that the main difficulty is not merely shortening evidence, but retaining the additional context required to make fine-grained citations self-contained.

Method	Suf.	Con.	Read.	Time↓
ContextCite	60.0	33.0	42.0	23.5
LAQuer	75.0	70.5	71.0	16.5
Ours	84.5	78.0	81.5	13.9

Table 5: Human evaluation on 100 test examples. Suf., Con., and Read. denote the percentage of positive judgments for sufficiency, conciseness, and readability, respectively. Time is the total verification time in minutes.

Effect of Two-stage Training. We examine how synthetic and human annotations should be combined. A simple strategy is to shuffle synthetic and human annotations together and fine-tune the model once. However, because synthetic data is much larger and noisier than annotated data, directly mixing the two sources may dilute the effect of high-quality manual annotations. We therefore compare mixed training with our two-stage strategy, where the model is first trained on synthetic data and then refined on human annotations.

Human Evaluation. Because automatic overlap metrics cannot fully capture whether a citation is self-contained and easy to verify, we conduct a human evaluation on 100 randomly sampled test instances. Two annotators independently evaluate anonymized outputs from ContextCite, LAQuer, and our method using binary judgments of sufficiency, conciseness, and readability. We also record the total time required to verify the 100 answers.

As shown in Table 5, our method receives the highest ratings on all three criteria and reduces total verification time from 23.5 minutes with ContextCite and 16.5 minutes with LAQuer to 13.9 minutes. These results complement the automatic evaluation and provide direct evidence that concise and sufficient citations can reduce the effort required for answer verification. In particular, the improvement in verification time is consistent with our motivation that citation quality should not be measured solely by whether the correct evidence is included, but also by how efficiently users can locate and interpret it. The simultaneous gains in sufficiency and conciseness further indicate that shorter citations need not come at the expense of evidential completeness.

Table 6 shows that two-stage training performs best when 1K manual annotations are available, improving F1 from 78.35 in the shuffled setting to 80.46. In the 0.5K manual setting, it achieves 77.53 F1, close to the synthetic-only setting and higher than manual-only training. These results suggest

Setting	1K Manual	0.5K Manual
Manual Only	77.15	74.84
Synthetic Only	77.86	77.86
Synthetic + Manual shuffled	78.35	77.16
Two-stage	80.46	77.53

Table 6: Comparison between direct mixed training and two-stage training.

that synthetic and human annotations should not be treated as interchangeable training instances. Synthetic data helps the model learn general citation behavior and output format, while human annotations better align it with the target citation standard. This distinction is important because small span-boundary differences can affect both citation conciseness and sufficiency.

Generalization and Robustness. We additionally evaluate robustness to substantially longer and noisier contexts, as well as out-of-domain transfer to ASQA and ELI5. With contexts expanded to approximately 500 and 1,000 words, our method achieves 78.26 and 76.49 F1, respectively, outperforming ContextCite and LAQuer while maintaining shorter citations. On 100 randomly sampled examples from both ASQA and ELI5, our method also obtains the highest F1 among the compared methods. These results provide initial evidence that the learned citation behavior is not restricted to the original benchmark. Detailed settings and results are provided in Appendix E.

6 Conclusion

In this work, we study concise and sufficient sub-sentence-level citations for RAG-based systems to address the problem that existing citation strategies may often contain substantial irrelevant information, or omit some content necessary for verification. We develop the annotation guidelines and construct a dataset for our improved citation requirements. A synthetic data generation pipeline that leverages LLMs to produce training instances with a rigorous automated filtering stage is proposed to augment manually annotated data. Experiments conducted on our constructed dataset verify the effectiveness of the proposed approach.

Limitations

Our benchmark mainly focuses on short-form QA rather than long-form responses with many inter-dependent claims. Although our noisy-context and out-of-domain experiments show encourag-

ing transfer, they do not fully represent long-form multi-claim citation generation. Extending the benchmark to such settings remains future work. Our approach also introduces additional computational cost due to synthetic annotation, evaluator-based filtering, and LLM inference. Finally, although citation spans are copied from the retrieved context, incorrect span selection may still yield incomplete or insufficient evidence.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (No.62306140), the Fundamental Research Funds for the Central Universities (NO.NXD2026009), the National Natural Science Foundation of China (No.62476127), the Natural Science Foundation of Jiangsu Province (No.BK20242039), the Fundamental Research Funds for the Central Universities (NO.NXD2026006), the Research Fund (No.PO250624101698), the Scientific Research Starting Foundation of Nanjing University of Aeronautics and Astronautics (No.YQR21022), and the High Performance Computing Platform of Nanjing University of Aeronautics and Astronautics.

References

- Avishek Anand, Lijun Lyu, Maximilian Idahl, Yumeng Wang, Jonas Wallat, and Zijian Zhang. 2022. Explainable information retrieval: A survey. *arXiv preprint arXiv:2211.02405*.
- Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. [XOR QA: Cross-lingual open-retrieval question answering](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online. Association for Computational Linguistics.
- Razvan Azamfirei, Sapna R Kudchadkar, and James Fackler. 2023. Large language models and the perils of their hallucinations. *Critical Care*, 27(1):120.
- Jasmijn Bastings and Katja Filippova. 2020. [The elephant in the interpretability room: Why use attention as explanation when we have saliency methods?](#) In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2020, Online, November 2020*, pages 149–155. Association for Computational Linguistics.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Yung-Sung Chuang, Benjamin Cohen-Wang, Zejiang Shen, Zhaofeng Wu, Hu Xu, Xi Victoria Lin, James R. Glass, Shang-Wen Li, and Wen-Tau Yih. 2025. [SelfCite: Self-supervised alignment for context attribution in large language models](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 10839–10858. PMLR.
- Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Mařdry. 2024. [ContextCite: Attributing model generation to context](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 95764–95807. Curran Associates, Inc.
- Qiang Ding, Lvzhou Luo, Yixuan Cao, and Ping Luo. 2025. [Attention with dependency parsing augmentation for fine-grained attribution](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 372–387, Vienna, Austria. Association for Computational Linguistics.
- 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.
- Eran Hirsch, Aviv Slobodkin, David Wan, Elias Stengel-Eskin, Mohit Bansal, and Ido Dagan. 2025. [LAQuer: Localized attribution queries in content-grounded generation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15355–15370, Vienna, Austria. Association for Computational Linguistics.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.
- Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kůttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

- 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.
- 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.
- 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. *arXiv preprint arXiv:2203.11147*.
- Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.
- Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.
- Gorjan Radevski, Kiril Gashteovski, Shahbaz Syed, Christopher Malon, Sebastien Nicolas, Chia-Chien Hung, Timo Szttyler, Verena Heußer, Wiem Ben Rim, Masafumi Enomoto, Kunihiro Takeoka, Masafumi Oyamada, Goran Glavaš, and Carolin Lawrence. 2025. [On synthesizing data for context attribution in question answering](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16929–16950, Vienna, Austria. Association for Computational Linguistics.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Wojciech Samek and Klaus-Robert Müller. 2019. Towards explainable artificial intelligence. In *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 5–22. Springer.
- Jiajun Shen, Tong Zhou, Yubo Chen, Kang Liu, and Jun Zhao. 2025. CiteLab: Developing and diagnosing llm citation generation workflows via the human-llm interaction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 490–501.
- Aviv Slobodkin, Eran Hirsch, Arie Cattat, 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.
- Viju Sudhi, Sinchana Ramakanth Bhat, Max Rudat, and Roman Teucher. 2024. Rag-ex: A generic framework for explaining retrieval augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2776–2780.
- Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient fact-checking of llms on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847.
- Sirui Xia, Xintao Wang, Jiaqing Liang, Yifei Zhang, Weikang Zhou, Jiaji Deng, Fei Yu, and Yanghua Xiao. 2025. [Ground every sentence: Improving retrieval-augmented LLMs with interleaved reference-claim generation](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 969–988, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yilong Xu, Jinhua Gao, Xiaoming Yu, Baolong Bi, Huawei Shen, and Xueqi Cheng. 2025. [ALiCE: Evaluating positional fine-grained citation generation](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 545–561, Albuquerque, New Mexico. Association for Computational Linguistics.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [HotpotQA: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.
- Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2024. [Cognitive mirage: A review of hallucinations in large language models](#). In *Proceedings of the First International OpenKG Workshop: Large Knowledge-Enhanced Models co-located with The International Joint Conference on Artificial Intelligence (IJCAI 2024)*, Jeju Island, South Korea, August 3, 2024, volume 3818 of *CEUR Workshop Proceedings*, pages 14–36. CEUR-WS.org.
- Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong,

Ling Feng, and Juanzi Li. 2025. [LongCite: Enabling LLMs to generate fine-grained citations in long-context QA](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5098–5122, Vienna, Austria. Association for Computational Linguistics.

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

A Annotation Guidelines

This section summarizes the guidelines provided to annotators for constructing sub-sentence citations. Given a question, its context, and a reference answer, annotators were asked to select exact text spans copied from the context that provide sufficient evidence for verifying the answer. The selected spans should satisfy three criteria: **sufficiency**, **conciseness**, and **readability**.

Sufficiency. The selected evidence should contain enough information to verify the answer without relying on unstated surrounding context. In particular, annotators were instructed to preserve necessary subjects, antecedents, or other contextual information when an answer-bearing phrase alone would be ambiguous. For example, a span such as “completed in 80 AD” is insufficient if the entity being completed is not identifiable from the citation itself.

Conciseness. Annotators should exclude clauses, modifiers, examples, or background information that are not needed to verify the answer. When only part of a sentence supports the answer, annotators should select the shortest natural span that remains sufficient. This avoids over-citation, where a correct citation contains substantial irrelevant information.

Readability. Selected spans should form coherent phrases or clauses rather than isolated keywords whenever possible. Annotators were encouraged to preserve natural phrase boundaries so that the citation remains understandable when presented independently.

Annotation Procedure. For each instance, annotators first identify the sentence or sentences that contain evidence for the answer. They then determine whether the complete sentence is already

concise and sufficient. If so, the full sentence is retained. Otherwise, irrelevant content is removed while preserving all information required for verification. If the remaining answer-bearing span is not self-contained because of a pronoun, implicit subject, or other context-dependent expression, an additional antecedent span is included. Multiple spans may also be selected when the answer requires evidence from more than one location in the context.

Annotators were specifically instructed to avoid two common errors:

- **Under-citation:** the selected span omits information needed to determine what entity, event, or claim the evidence refers to.
- **Over-citation:** the selected span contains information unrelated to verifying the answer.

The cited text must be copied exactly from the context without rewriting or paraphrasing. When multiple valid supporting spans are available, annotators should prefer the shortest natural span that satisfies the above criteria.

Examples. We distinguish three representative annotation cases.

Type-1: Complete sentence is already sufficient and concise.

Question: What is the capital of France?

Context: Paris is the capital of France.

Citation: Paris is the capital of France.

Here, removing any part of the sentence would not provide a clear advantage, so the complete sentence is retained.

Type-2: Remove irrelevant content from the supporting sentence.

Question: When was the Colosseum completed?

Context: The Colosseum was completed in 80 AD under Emperor Titus and later hosted gladiatorial contests and public spectacles.

Citation:
The Colosseum was completed in 80 AD

The later description of gladiatorial contests is irrelevant to the question and is therefore excluded.

Type-3: Additional antecedent or multiple spans are required.

Question: When was the Colosseum completed?

Context: The Colosseum is an ancient amphitheater in Rome. It was completed in 80 AD under Emperor Titus.

Citation: The Colosseum +
was completed in 80 AD

Selecting only “was completed in 80 AD” would be insufficient because the subject is implicit. The antecedent is therefore included as an additional span.

These examples illustrate the general principle followed throughout annotation: retain the minimum amount of context needed to make the evidence both verifiable and independently understandable.

B Case Study

Table 7 showcases the citation outputs of different methods across two representative open-domain QA scenarios. Q1, concerning the state capital of Assam, involves a context containing multiple sentences that can support the answer independently. Q2, regarding YouTube Kids’ supported operating systems, features a high density of extraneous background information. These two scenarios test the methods’ ability to balance informativeness and conciseness in supporting fact extraction.

For Q1, the single-sentence perturbation strategy of RAG-Ex 2.0 fails to produce any supporting facts. This is because when multiple sentences can independently support the answer, removing any one still leaves sufficient information, leading to negligible perturbation differences that prevent identifying the core supporting content. Meanwhile, LongCite-Llama-3.1-8B and ALCE exhibit significant citation redundancy. Their outputs include multiple supporting sentences even when a single citation would suffice, thereby increasing the cognitive load on users who must verify the answer across repetitive claims.

For Q2, the context contains extensive irrelevant details, such as parental control features and upload restrictions, that are unrelated to the OS support question. RAG-Ex 2.0, LongCite-Llama-3.1-8B, and ALCE all fail to filter these distractions: their citations retain non-core content alongside the key information, making the supporting facts hard to locate quickly.

In contrast, our method tailors citations to each scenario’s needs: for Q1, it retains only one core

claim confirming Dispur as Assam’s capital, excluding redundant repetitions and irrelevant background. For Q2, it strips away non-essential app-function details and focuses solely on the OS support information. This ensures that our citations are sufficient for answer verification while remaining concise.

This comparison demonstrates that our sub-sentence-level citation generation method identifies the exact fragments necessary for verification, reducing citation redundancy without sacrificing informativeness.

C Synthetic Data Filtering Details

This section describes the filtering procedure used to select high-quality synthetic citation annotations. Given a synthetic instance, the filter first applies rule-based validation and then optionally uses an evaluator model to score citation quality. The final score combines both components, and instances above a predefined threshold are retained for training.

C.1 Rule-based Validation

For each synthetic instance, we parse the model output into a JSON object with two required fields: answer and supporting_facts. Instances with invalid JSON or missing required fields are rejected or penalized. Each supporting fact is expected to contain a cited text span and a line identifier.

The rule-based validator checks answer consistency, citation validity, citation length, and citation count. Answer consistency is measured using a normalized fuzzy matching score between the generated answer and the reference answer. The generated answer is accepted if it exactly matches, contains, or is contained by the reference answer after normalization, or if its fuzzy matching score is at least 0.85. Otherwise, the instance is rejected.

Citation validity requires each cited text span to be non-empty and appear in the input context after normalization. Line identifiers are also checked when available, but line mismatches are only penalized rather than used as hard rejection, since some contexts are stored as single-paragraph text. We further penalize citations that are too long relative to the context. Let L_c be the total number of words in all generated citation spans and L_x be the number of words in the context. The citation ratio is

$$r = \frac{L_c}{L_x}.$$

Methods	Q1: “What is the capital of the state of Assam?” Answer: “Dispur.”	Q2: “On what OS was the YouTube Kids app available?” Answer: “Android and iOS”
	Generated Citations	Generated Citations
RAG-Ex 2.0	Dispur is the capital of the state of Assam in India. Dispur, a locality of Guwahati, became the capital of Assam in 1973. This was after Shillong, the erstwhile capital, became the capital of the state of Meghalaya that was carved out of Assam. Dispur is the seat of Government of Assam. ...	In February 2015, YouTube announced the launch of a new app specifically for use by children visiting the site, called YouTube Kids. It allows parental controls and restrictions on who can upload content, and is available for both Android and iOS devices. Later on August 26, 2015, YouTube Gaming was launched. ...
LongCite-Llama-3.1-8B	Dispur is the capital of the state of Assam in India. Dispur, a locality of Guwahati, became the capital of Assam in 1973. This was after Shillong, the erstwhile capital, became the capital of the state of Meghalaya that was carved out of Assam. Dispur is the seat of Government of Assam. ...	In February 2015, YouTube announced the launch of a new app specifically for use by children visiting the site, called YouTube Kids. It allows parental controls and restrictions on who can upload content, and is available for both Android and iOS devices. Later on August 26, 2015, YouTube Gaming was launched. ...
ALCE	Dispur is the capital of the state of Assam in India. Dispur, a locality of Guwahati, became the capital of Assam in 1973. This was after Shillong, the erstwhile capital, became the capital of the state of Meghalaya that was carved out of Assam. Dispur is the seat of Government of Assam. ...	In February 2015, YouTube announced the launch of a new app specifically for use by children visiting the site, called YouTube Kids. It allows parental controls and restrictions on who can upload content, and is available for both Android and iOS devices. Later on August 26, 2015, YouTube Gaming was launched
Ours	Dispur is the capital of the state of Assam in India. Dispur, a locality of Guwahati, became the capital of Assam in 1973. This was after Shillong, the erstwhile capital, became the capital of the state of Meghalaya that was carved out of Assam. Dispur is the seat of Government of Assam. ...	In February 2015, YouTube announced the launch of a new app specifically for use by children visiting the site, called YouTube Kids. It allows parental controls and restrictions on who can upload content, and is available for both Android and iOS devices. Later on August 26, 2015, YouTube Gaming was launched. ...

Table 7: Comparison of citation results generated by different methods on two queries.

If $r > 0.60$, the instance receives a stronger length penalty; if $0.40 < r \leq 0.60$, it receives a smaller penalty. Instances with more than five citation spans are also penalized.

The rule-based score s_h is initialized to 1.0 and reduced according to the validation errors:

$$s_h = \max(0, \min(1, 1 - \sum_i p_i)),$$

where p_i denotes the penalty associated with each detected issue. An instance passes the rule-based stage only if $s_h \geq 0.50$, all cited spans appear in the context, and at least one citation span is provided.

C.2 Evaluator-based Scoring

For instances that pass rule-based validation, we use an evaluator model to assess citation quality. The evaluator receives the context, question, reference answer, generated answer, and generated supporting facts. It outputs integer scores from 1 to 5 for six dimensions: factual fidelity, answer correctness, sufficiency, conciseness, coherence, and

overall quality. These dimensions are designed to capture whether the synthetic instance is faithful to the context, whether the answer matches the reference, whether the cited spans are sufficient for verification, whether they avoid irrelevant content, and whether they are readable.

Let the normalized evaluator scores be

$$f_f, f_a, f_s, f_c, f_r, f_o \in [0, 1],$$

corresponding to factual fidelity, answer correctness, sufficiency, conciseness, coherence, and overall score, respectively. The evaluator score is computed as

$$s_e = 0.25f_f + 0.25f_a + 0.20f_s + 0.15f_c + 0.05f_r + 0.10f_o.$$

This weighting emphasizes factual fidelity, answer correctness, and sufficiency, while still considering conciseness and coherence.

Component	Prompt content
Role instruction	You are a question-answering assistant that strictly follows the required output format. You must output only a valid JSON object and no additional explanation.
Task definition	Given a context and a question, find the answer from the context and return supporting facts. The output must contain exactly two fields: <code>answer</code> and <code>supporting_facts</code> .
Citation format	Each supporting fact should be a list containing the cited text span and its line identifier. Line identifiers start from 0. The cited text must be copied exactly from the context.
Output constraint	Do not output explanations, Markdown, or any extra text.
Output format	<code>{ "answer": "direct answer", "supporting_facts": [{"evidence span", line_id}] }</code>
Few-shot examples	The system prompt then provides 10 annotated training examples in the same input-output format.
Test-time input	For each test instance, the user message contains the instruction, context, and question.

Table 8: Prompt template used for the GPT-4o few-shot prompting baseline.

C.3 Final Filtering Score

The final filtering score combines the rule-based score and evaluator score:

$$S = 100 \times (0.40s_h + 0.60s_e).$$

Thus, S ranges from 0 to 100. In our experiments, synthetic instances are retained if their final score is no lower than the filtering threshold. We set the default threshold to 80. Alternatively, the same scoring function can be used to select the top- k highest-scoring synthetic instances.

This two-stage design ensures that structurally invalid or unsupported citations are removed by deterministic checks, while the evaluator model further filters instances according to higher-level criteria such as sufficiency, conciseness, and readability.

D GPT-4o Prompting Baseline

We evaluate GPT-4o as a few-shot prompting baseline. The system prompt contains 10 annotated examples from the training set and instructs the model to output only a JSON object with two fields: `answer` and `supporting_facts`. Each supporting fact contains a citation span copied from the context and its line identifier. We use temperature 0 for deterministic generation. The prompt template is shown in Table 8.

E Additional Generalization Experiments

To further evaluate whether our method remains effective beyond the original benchmark setting, we conduct additional experiments under longer and noisier contexts as well as on external long-form QA datasets. For the noisy-context experiments, we retain the original question, answer, and

Dataset	Method	F1	CL
Noisy-500	ContextCite	44.07	145.01
	LAQuer	75.87	78.56
	Ours	78.26	69.49
Noisy-1000	ContextCite	38.12	142.12
	LAQuer	75.13	80.30
	Ours	76.49	70.46
ASQA	ContextCite	19.14	111.77
	LAQuer	40.18	143.11
	Ours	41.82	36.11
ELI5	ContextCite	14.10	99.29
	LAQuer	40.46	155.12
	Ours	42.44	47.93

Table 9: Robustness to noisy contexts and transfer to external long-form QA datasets. ASQA and ELI5 results are evaluated on 100 randomly sampled examples from each dataset.

gold supporting evidence, while inserting irrelevant or weakly related sentences into the context. We construct two settings with approximately 500 and 1,000 words, denoted as Noisy-500 and Noisy-1000, respectively. These settings test whether a model can still identify concise supporting spans when the relevant evidence is surrounded by substantially more distracting content.

We also evaluate out-of-domain transfer on ASQA and ELI5. Since these datasets were not constructed specifically for our sub-sentence citation task, we randomly sample 100 examples from each dataset and evaluate the compared methods under the same citation extraction and scoring protocol. This experiment is intended to provide an initial assessment of whether the learned citation behavior transfers to different QA domains and longer-form contexts.

As shown in Table 9, increasing context noise causes a performance drop for all methods, but our model remains the strongest under both noisy-

Gold Reference	F1	CL
Annotator 1	80.41	72.51
Annotator 2	81.00	72.25
Third-expert adjudication	80.34	72.45

Table 10: Evaluation using alternative human span references.

context settings. On Noisy-500, our method achieves 78.26 F1 with a citation length of 69.49, compared with 75.87 F1 and 78.56 CL for LA-Quer. When the context is further expanded to approximately 1,000 words, our method still achieves 76.49 F1 and maintains shorter citations than LA-Quer. In contrast, ContextCite degrades more substantially as additional distracting context is introduced.

The transfer results on ASQA and ELI5 show a similar trend. Although absolute F1 scores are lower than those on our in-domain benchmark, our method consistently achieves the highest F1 while producing substantially shorter citations. On ASQA, it reaches 41.82 F1 with 36.11 CL, while on ELI5 it obtains 42.44 F1 with 47.93 CL. These results suggest that the learned sub-sentence citation behavior is not restricted to the original benchmark and can partially transfer to different QA settings.

Nevertheless, these experiments should be viewed as an initial generalization analysis rather than a full evaluation of long-form multi-claim citation generation. In particular, ASQA and ELI5 contain more complex and open-ended answers than our primary benchmark. Extending the method and annotation protocol to large-scale long-form citation benchmarks remains an important direction for future work.

F Annotation Disagreement Analysis

To examine whether different valid span boundaries substantially affect evaluation, we additionally evaluate the final model on the 100 disagreement instances discarded during benchmark construction. We use each annotator’s selection as an independent reference and also construct an adjudicated reference by a third expert.

The similar results across the three references suggest that the main performance conclusion is robust to reasonable alternative span selections. Nevertheless, agreement filtering may reduce coverage of intrinsically ambiguous cases, and future benchmarks could retain multiple valid references or use adjudication.