TROVE: A Challenge for Fine-Grained Text Provenance via Source Sentence Tracing and Relationship Classification

Junnan Zhu[1*†], **Min Xiao**[1,2†], **Yining Wang**[4], **Feifei Zhai**[1,3], **Yu Zhou**[1,3], **Chengqing Zong**[1,2]
[1] State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, CAS, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] Fanyu AI Laboratory, Zhongke Fanyu Technology Co., Ltd, Beijing, China
[4] Unisound AI Technology Co.Ltd
{junnan.zhu, yzhou, cqzong}@nlpr.ia.ac.cn

Abstract

LLMs have achieved remarkable fluency and coherence in text generation, yet their widespread adoption has raised concerns about content reliability and accountability. In high-stakes domains, it is crucial to understand where and how the content is created. To address this, we introduce the Text pROVEnance (TROVE) challenge, designed to trace each sentence of a target text back to specific source sentences within potentially lengthy or multi-document inputs. Beyond identifying sources, TROVE annotates the fine-grained relationships (*quotation*, *compression*, *inference*, and *others*), providing a deep understanding of how each target sentence is formed. To benchmark TROVE, we construct our dataset by leveraging three public datasets covering 11 diverse scenarios (e.g., QA and summarization) in English and Chinese, spanning source texts of varying lengths (0-5k, 5-10k, 10k+), emphasizing the multi-document and long-document settings essential for provenance. To ensure high-quality data, we employ a three-stage annotation process: sentence retrieval, GPT-4o provenance, and human provenance. We evaluate 11 LLMs under direct prompting and retrieval-augmented paradigms, revealing that retrieval is essential for robust performance, larger models perform better in complex relationship classification, and closed-source models often lead, yet open-source models show significant promise, particularly with retrieval augmentation. We make our dataset available here: https://github.com/ZNLP/ZNLP-Dataset.

1 Introduction

Large language models (LLMs) have demonstrated great potential in natural language generation, producing highly coherent and fluent human-like text. However, their rapidly increasing prevalence raises significant concerns regarding content accountability and reliability. While considerable efforts have been made in citation (Li et al., 2024; Huang et al., 2024; Cao and Wang, 2024; Aly et al., 2024) and grounded generation (Li et al., 2022; Brahman et al., 2022; Slobodkin et al., 2024), most existing studies focus on single-document-level source identification, leading to a significant gap in meeting the requirements of real-world scenarios. For instance, in domains like legal document drafting or medical reporting, it is crucial to identify where a sentence originates and understand how it has been generated from the sources.

To bridge this gap, we introduce the challenge of text provenance (TROVE), which involves tracing a target text to the given source document(s) and establishing fine-grained relationships between the target and its source. TROVE is critical and challenging for large-scale information, as tracing sources becomes more complicated with longer or more numerous documents. To benchmark TROVE, we construct our dataset based on three public datasets: LongBench (Bai et al., 2024), LooGLE (Li et al., 2022), and CRUD-RAG (Lyu et al., 2024), considering the perspectives of multi-document, long-document, or their combination.

Specifically, we first select examples with multi-sentence outputs to ensure sufficient sources and categorize data by original scenarios (question answering or summarization), languages (Chinese or English), input text length (0-5k, 5k-10k, or 10k+), and number of input documents (single-document or multi-document), aiming for a balanced distribution across each dimension. Next, we employ multiple information retrieval methods to recall the candidate source sentences for each sentence in the target text. Then, we use crafted prompts to guide GPT-4o in preliminary annotation, aiming to identify the sources of target sentences from retrieved candidate sentences and classify the target-source relationships into quotation (verbatim or partial

[*] Corresponding Author.
[†] Equal Contribution.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11755–11771
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

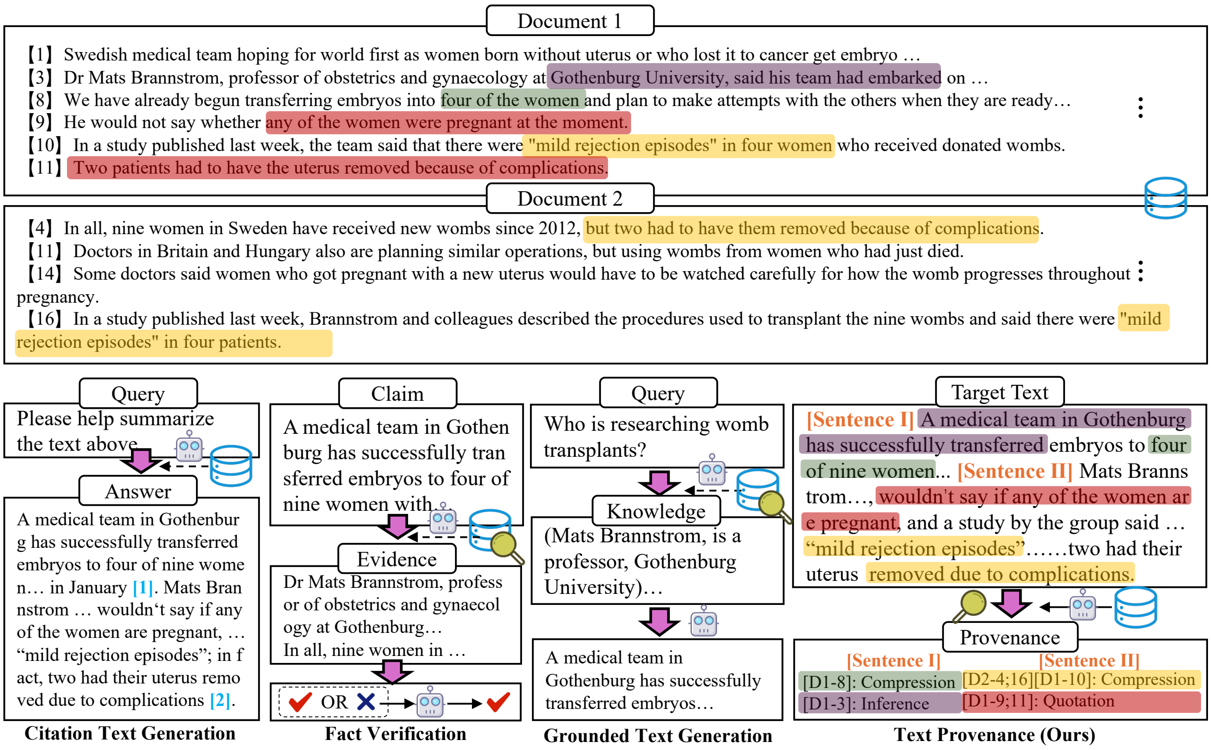

Document 1
【1】 Swedish medical team hoping for world first as women born without uterus or who lost it to cancer get embryo …
【3】 Dr Mats Brannstrom, professor of obstetrics and gynaecology at Gothenburg University, said his team had embarked on …
【8】 We have already begun transferring embryos into four of the women and plan to make attempts with the others when they are ready…
【9】 He would not say whether any of the women were pregnant at the moment.
【10】 In a study published last week, the team said that there were "mild rejection episodes" in four women who received donated wombs.
【11】 Two patients had to have the uterus removed because of complications.
Document 2
【4】 In all, nine women in Sweden have received new wombs since 2012, but two had to have them removed because of complications.
【11】 Doctors in Britain and Hungary also are planning similar operations, but using wombs from women who had just died.
【14】 Some doctors said women who got pregnant with a new uterus would have to be watched carefully for how the womb progresses throughout pregnancy.
【16】 In a study published last week, Brannstrom and colleagues described the procedures used to transplant the nine wombs and said there were "mild rejection episodes" in four patients.
Query
Please help summarize the text above
Answer
A medical team in Gothenburg has successfully transferred embryos to four of nine women… in January [1]. Mats Brannstrom … wouldn't say if any of the women are pregnant, … "mild rejection episodes"; in fact, two had their uterus removed due to complications [2].
Citation Text Generation
Claim
A medical team in Gothenburg has successfully transferred embryos to four of nine women with…
Evidence
Dr Mats Brannstrom, professor of obstetrics and gynaecology at Gothenburg…
In all, nine women in …
✓ OR ✗
✓
Fact Verification
Query
Who is researching womb transplants?
Knowledge
(Mats Brannstrom, is a professor, Gothenburg University)…
A medical team in Gothenburg has successfully transferred embryos…
Grounded Text Generation
Target Text
[Sentence I] A medical team in Gothenburg has successfully transferred embryos to four of nine women... [Sentence II] Mats Brannstrom…, wouldn't say if any of the women are pregnant, and a study by the group said … "mild rejection episodes"……two had their uterus removed due to complications.
Provenance
[Sentence I]
[D1-8]: Compression
[D1-3]: Inference
[Sentence II]
[D2-4;16][D1-10]: Compression
[D1-9;11]: Quotation
Text Provenance (Ours)

Figure 1: Overview of the difference between text provenance and related tasks. Solid arrows indicate required inputs and outputs, while dashed arrows represent optional ones. Compared with existing studies focusing on single-document or coarse-grained scenarios, our TROVE involves finer-grained provenance.

copy), compression (summarization or paraphrase), inference (expansion, generalization, or specification), and others (e.g., negation). Each target sentence can be traced to multiple source sentences, and different relationships may apply simultaneously. For instance, a target sentence sourced from [a, b, c] might have [a, b] labeled as compression and [c] as quotation. Finally, we conduct a human review, requiring annotators to verify each annotation by considering both the source document(s) and the results produced by GPT-4o.

We perform a comprehensive evaluation of 11 models under two paradigms: *direct prompting* and *retrieval-augmented*, yielding valuable insights into the capabilities of current models in TROVE.

Our main contributions are as follows:

- We introduce TROVE, a new challenge that traces each target sentence to its originating sources and classifies fine-grained target-source relationships beyond coarse-grained or single-document source identification.

- We present a carefully curated dataset covering multiple scenarios, languages, and source lengths. Our three-stage annotation produces high-quality, fine-grained provenance data.

- We systematically evaluate 11 LLMs (both closed-source and open-source) under multiple scenarios, revealing the necessity of retrieval augmentation, the advantages of larger models for relationship classification, and relationship classification remains challenging.

2 Related Work

Citation Text Generation. Citation text generation focuses on producing text enriched with citations to enhance verifiability, making it widely applicable in academic writing (Jurgens et al., 2018; Xing et al., 2020; Lauscher et al., 2022; Mandal et al., 2024) and LLM-based chatbots (Li et al., 2024; Huang et al., 2024; Cao and Wang, 2024; Aly et al., 2024). Existing approaches can be categorized into parametric and non-parametric methods. Parametric methods (Mandal et al., 2024; Gu and Hahnloser, 2024) rely on knowledge and patterns implicitly encoded within the model parameters to generate citation text. However, they face challenges in incorporating new citations or knowledge updates and are prone to hallucinations. Non-parametric methods (Gao et al., 2023; Huang et al., 2024; Li et al., 2024) directly access external knowledge sources, such as citation databases, documents, or retrieval systems, to produce more reliable citation text. These approaches often leverage retrieval-augmented generation (RAG) techniques to integrate retrieved information with text generation. However, citations are typically generated in a post-hoc manner, which increases latency. Most existing methods focus on producing document-level, single-reference citations and emphasize the quality of the generated text.

Fact Verification. Existing studies on fact verification (Chen et al., 2024; Zheng et al., 2024; Churina et al., 2024) typically follow a two-stage approach: evidence retrieval and claim verification. Evidence retrieval aims to identify relevant passages or documents using information retrieval (Chen et al., 2022; Zheng et al., 2024) or neural ranking models (Malviya and Katsigiannis, 2024). Claim verification aims to determine the authenticity of a claim by comparing it with the retrieved evidence, which has received more attention (Zhong et al., 2020; Zou et al., 2023).

Grounded Text Generation. Grounded text generation aims to produce text consistent with external sources of information, such as knowledge bases (Li et al., 2022; Lu et al., 2022), documents (Slobodkin et al., 2024; Hsu et al., 2024), or real-world facts (Godbole et al., 2024; Brahman et al., 2022). This task ensures factual accuracy and coherence in generated content, as seen in applications like dialogue generation (Li et al., 2022; Lu et al., 2022) and factual summarization (Slobodkin et al., 2024; Song et al., 2022).

Text Provenance: Unique Challenges and Contributions. As shown in Figure 1, while citation text generation, fact verification, and grounded text generation all involve interactions between generated text and external sources, they each emphasize different aspects. Citation text generation focuses on incorporating references to support claims, fact verification aims to validate the truthfulness of statements, and grounded text generation ensures consistency with external information. In contrast, text provenance uniquely concentrates on identifying the specific source sentences for each target sentence and classifying the precise nature of their relationships, such as direct quotation, compression, inference, or negation. It requires retrieving relevant source sentences and performing a detailed semantic analysis to categorize the type of relationship, thereby providing a deeper understanding of how generated text originates from its sources. Consequently, text provenance extends beyond the capabilities of existing tasks by offering a more granular and relationship-focused approach to tracing the origins of generated content.

3 Task Formulation

Text Provenance aims to identify the source sentences for each target sentence in a generated text and classify the relationship between them using a given document collection. Specifically, given a target text $T = \{t_1, t_2, ..., t_n\}$, where each t_i is a target sentence, and a document collection $D = \{d_1, d_2, ..., d_m\}$, where each documents d_i contains a set of sentences $\{s_{i,1}, s_{i,2}, ..., s_{i,k_i}\}$, the task is to determine, for each target sentence t_i, a set of source sentences $\{s_{i,j_1}, s_{i,j_2}, ..., s_{i,j_k}\}$ from D and classify their relationships.

Each target sentence t_i may derive from multiple source sentences. The relationship between t_i and each source sentences $s_{i,j}$ is categorized into one of the following types: **Quotation**, where t_i is a verbatim or partial copy of $s_{i,j}$; **Compression**, where t_i is a paraphrase or a summary derived from multiple source sentences, such as s_{i,j_1} and s_{i,j_2}; **Inference**, where t_i is logically inferred from one or more source sentences, such as $s_{i,j}$; **Others**, where t_i does not fit the above categories, including cases like contradiction or negation.

For example, a target sentence t_i may be derived from multiple source sentences, such as $s_{1,2}$, $s_{1,3}$, and $s_{2,1}$, where the relationship between t_i and $s_{1,2}$, $s_{1,3}$ could be classified as compression, and the relationship between t_i and $s_{2,1}$ could be classified as inference. This task thus requires a system to identify the appropriate source sentences and determine the precise relationship between each target sentence and its sources, which presents a complex challenge in understanding the fine-grained relationships between generated text and its origins.

4 Datasets

4.1 Data Collection

Text provenance becomes particularly crucial when dealing with large volumes of information, as tracing sources becomes more difficult with increasing document length and number. Therefore, we construct our text provenance dataset from the perspective of multi-document, long-document, or a combination of both, based on three public datasets: LongBench (Bai et al., 2024), LooGLE (Li et al., 2022), and CRUD-RAG (Lyu et al., 2024).

LongBench focuses on long-context understanding and comprises 21 datasets in both English and Chinese across 6 task categories, covering key long-text application areas such as single-doc QA, multi-doc QA, and summarization. LooGLE is a com-

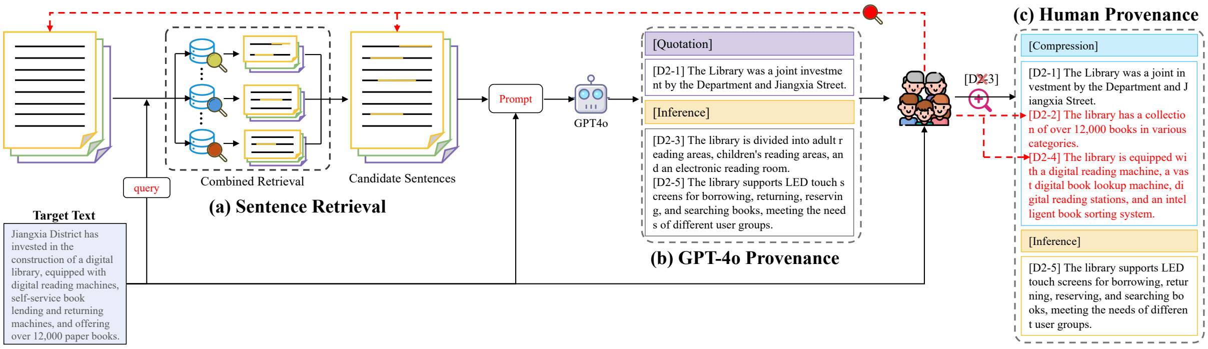

Target Text
Jiangxia District has invested in the construction of a digital library, equipped with digital reading machines, self-service book lending and returning machines, and offering over 12,000 paper books.
query
Combined Retrieval
Candidate Sentences
(a) Sentence Retrieval
Prompt
GPT4o
[Quotation]
[D2-1] The Library was a joint investment by the Department and Jiangxia Street.
[Inference]
[D2-3] The library is divided into adult reading areas, children's reading areas, and an electronic reading room.
[D2-5] The library supports LED touch screens for borrowing, returning, reserving, and searching books, meeting the needs of different user groups.
(b) GPT-4o Provenance
[D2-3]
(c) Human Provenance
[Compression]
[D2-1] The Library was a joint investment by the Department and Jiangxia Street.
[D2-2] The library has a collection of over 12,000 books in various categories.
[D2-4] The library is equipped with a digital reading machine, a vast digital book lookup machine, digital reading stations, and an intelligent book sorting system.
[Inference]
[D2-5] The library supports LED touch screens for borrowing, returning, reserving, and searching books, meeting the needs of different user groups.

Figure 2: Overview of our data annotation. (a) Sentence Retrieval: selecting candidate provenance sentences using multiple retrievers; (b) GPT-4o Provenance: automatically annotating provenance relationships based on retrieved sentences; (c) Human Provenance: reviewing and refining GPT-4o's annotations while independently checking source documents to identify missing provenance sentences. Di-j denotes the j-th sentence in the i-th document.

prehensive benchmark for evaluating long-context understanding in large language models. It features extremely long documents (post-2022) with over 24,000 tokens each and 6,000 questions across diverse domains, designed to assess short- and long-dependency tasks. CRUD-RAG is a comprehensive Chinese benchmark for evaluating Retrieval-Augmented Generation (RAG) systems. It categorizes RAG applications into four CRUD operations, i.e., Create, Read, Update, and Delete. It provides diverse evaluation tasks such as text continuation, question answering, hallucination modification, and multi-document summarization.

From these datasets, we select examples where the output (or reference) contains multiple sentences, using these as the target text for our task. This approach ensures that the target text provides sufficient material for detailed source tracing, as each sentence may originate from different fragments of the input documents. We then treat each sample's original inputs as a unified document collection. Specifically, we sample from *GovReport*, *QMSum*, *SAMSum*, *VCSum*, and *MultiNews* in LongBench and the long-dependency summarization (LongSum) task in LooGLE. Additionally, we include samples from *EventSum*, *QA1doc*, *QA2doc*, and *QA3doc* in CRUD-RAG. We categorize data by different tasks, languages (English and Chinese), input text length (0-5k, 5k-10k, and 10k+), and the number of input documents (single and multiple), trying to achieve a balanced distribution across these dimensions. Although we strive for a balanced distribution across these categories, some subsets inevitably remain underrepresented.

4.2 Data Annotation

We employ GPT-4o to alleviate the manual annotation workload, as shown in Figure 2. For each target sentence t_i, the annotation procedure consists of three steps: (a) sentence retrieval, (b) GPT-4o provenance, and (c) human provenance.

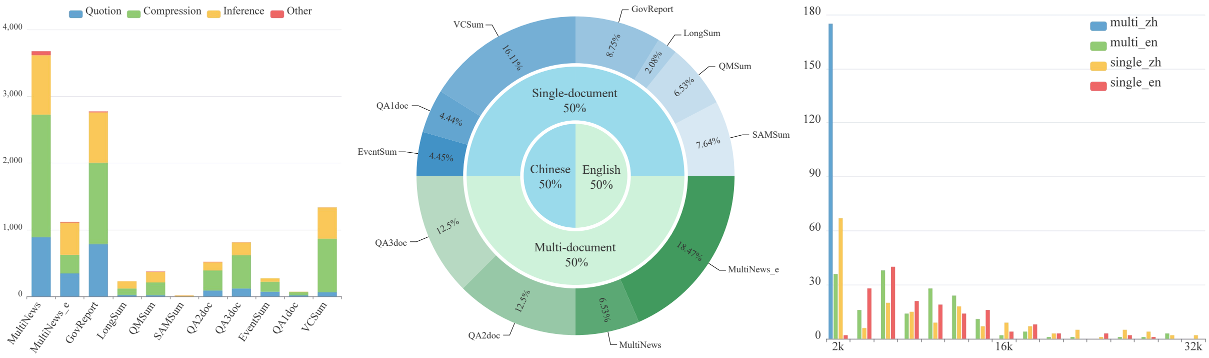
Sentence Retrieval. Due to the lengthy input text, GPT-4o may ignore key sentences when directly tracing provenance through long passages. To mitigate this, we first retrieve candidate provenance sentences using the target sentences as queries and then perform provenance based on these sentences. We have presented ablation studies to validate this approach. Moreover, different retrievers capture diverse semantic features. To maximize the recall rate of candidate provenance sentences, we aggregate the results from M distinct retrievers. Each retriever selects the top-k most relevant sentences, denoted as $R_i(D, t_i)$, forming the following set of candidate provenance sentences:

$$cands = \bigcup_{i}^{M} R_i(D, t_i) \tag{1}$$

To reduce recall errors, only the union of sentences recalled by at least two retrievers is considered. We employ three retrievers: BM25 (Robertson et al., 2009), Dense (Luan et al., 2021), and LCS (Saadah et al., 2013). Each retriever selects the top-k most relevant sentences, where $k = 10$.

GPT-4o Provenance. Based on the candidate provenance sentences, GPT-4o conducts fine-grained annotation and classifies the provenance relationship types, as depicted in Figure 2(b). The detailed prompt is provided in the appendix.

Human Provenance. Annotators review GPT-4o's results to verify the provenance sentences and their corresponding relationship types. Noting that GPT-4o can ignore critical details, the annotators examine the document collections to address any omissions. As illustrated in Figure 2(c), the sentences “D2-2” and “D2-4”, which contain “12,000



(a) Relationship across the task.

(b) Task and language distributions.

(c) Input lengths across the datasets.

Figure 3: Dataset distribution.

#Doc	Lang	Source			Target		Provenance			
		docs	sentences	tokens	sentences	tokens	sentences /example	tokens /example	sentences /sentence	tokens /sentence
single	zh	1.00	196.44	7,981.33	1.65	189.79	8.52	509.21	7.04	421.56
	en	1.00	636.61	9,074.38	9.83	253.30	19.19	620.33	1.74	52.46
multi	zh	2.51	20.95	903.48	2.44	146.02	6.78	400.33	2.98	182.12
	en	3.77	327.93	7,222.53	12.13	320.47	23.10	694.17	1.97	59.78

Table 1: Statistics of the dataset.

#Docs	Lang	Trace	Type	GPT-4o
single	zh	.6696	.5788	.4391
	en	.6410	.5336	.5325
multi	zh	.7400	.6187	.5328
	en	.6004	.4862	.6997

Table 2: Consistency of the annotation.

books” and “digital reading machine” respectively, exhibit a strong connection to the target sentence, yet GPT-4o fails to identify them. Therefore, the annotators will incorporate these two sentences into the final analysis. We invite 8 graduate students to spend about 510 hours annotating the provenance of 4,388 English and 811 Chinese sentences, costing an average of $0.20 per sentence.

4.3 Statistics of the Dataset

Table 1 provides detailed statistics of our dataset, including (1) Source: the average number of documents, sentences, and tokens per sample. (2) Target: the average number of target sentences and tokens per sample. (3) Provenance Results: the average number of provenance sentences and tokens per sample and the average number of provenance sentences and tokens per target sentence.

Figure 3 illustrates the dataset's distribution across key characteristics: (a) relationship distributions across the task, (b) task and language distributions, and (c) source length distributions. These visualizations highlight the dataset's high diversity.

4.4 Consistency Analysis

To ensure dataset quality, 10% of the examples are assigned to different annotators for consistency assessment. We evaluate annotator agreement from three perspectives: (1) tracing provenance sentences, (2) classification of relationship types, and (3) determination of necessary corrections to GPT-4o's provenance sentences. To quantify agreement, we use Fleiss' Kappa (Falotico and Quatto, 2015) to measure the reliability across multiple annotators. The results, presented in Table 2, demonstrate that the annotation process is reliable.

5 Experiment

5.1 Experimental Setup

We evaluate LLMs ranging from 6B to 671B parameters, including open-source and closed-source models. Open-source models include Qwen1.5-Instruct series (Bai et al., 2023): Qwen1.5-Instruct-7B-chat, Qwen1.5-Instruct-14B-chat; Qwen2.5-Instruct series (Qwen et al., 2025): Qwen2.5-Instruct-7B, Qwen2.5-Instruct-14B; Llama-3-Instruct-8B (Grattafiori et al., 2024); ChatGLM2-6B (Zeng et al., 2023); Vicuna-7B-V1.5 (Chiang et al., 2023); DeepSeek-V3 (671B) (DeepSeek-AI et al., 2024). Closed-source models include GPT-4o[1], Gemini-1.5-pro, Kimi[2].

The source length (0–32k) sometimes exceeds the context length supported by most LLMs. There-

[1] https://chat.openai.com/

[2] https://kimi.moonshot.cn/

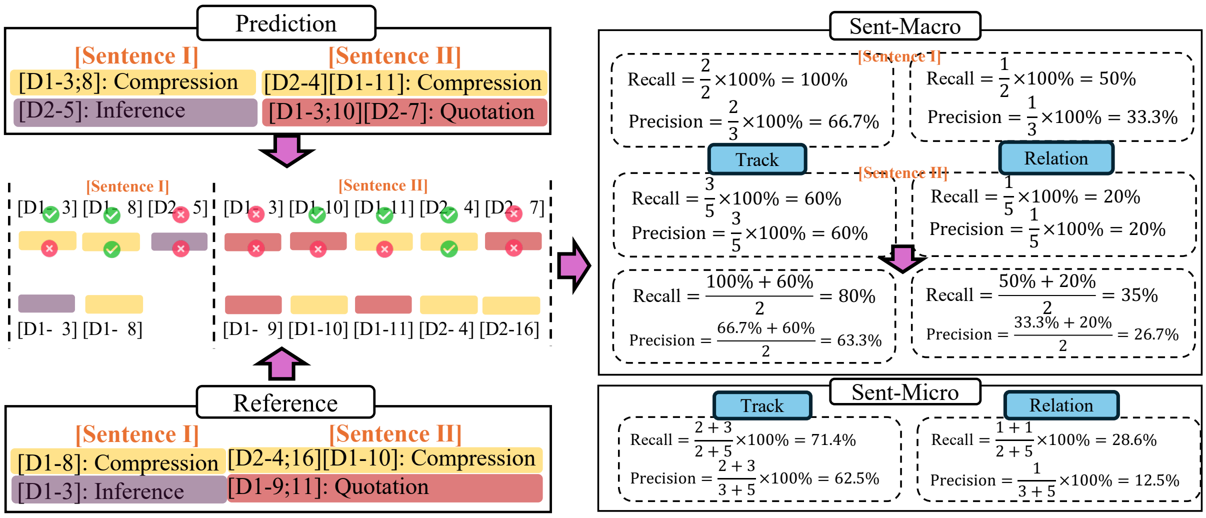

Prediction
[Sentence I]
[Sentence II]
[D1-3;8]: Compression
[D2-4][D1-11]: Compression
[D2-5]: Inference
[D1-3;10][D2-7]: Quotation
[Sentence I]
[Sentence II]
[D1- 3] [D1- 8] [D2- 5]
[D1- 3] [D1-10] [D1-11] [D2- 4] [D2- 7]
[D1- 3] [D1- 8]
[D1- 9] [D1-10] [D1-11] [D2- 4] [D2-16]
Reference
[Sentence I]
[Sentence II]
[D1-8]: Compression
[D2-4;16][D1-10]: Compression
[D1-3]: Inference
[D1-9;11]: Quotation
Sent-Macro
[Sentence I]
Recall = 2/2 ×100% = 100%
Precision = 2/3 ×100% = 66.7%
Recall = 1/2 ×100% = 50%
Precision = 1/3 ×100% = 33.3%
Track
Relation
[Sentence II]
Recall = 3/5 ×100% = 60%
Precision = 3/5 ×100% = 60%
Recall = 1/5 ×100% = 20%
Precision = 1/5 ×100% = 20%
Recall = (100% + 60%)/2 = 80%
Precision = (66.7% + 60%)/2 = 63.3%
Recall = (50% + 20%)/2 = 35%
Precision = (33.3% + 20%)/2 = 26.7%
Sent-Micro
Track
Relation
Recall = (2 + 3)/(2 + 5) ×100% = 71.4%
Precision = (2 + 3)/(3 + 5) ×100% = 62.5%
Recall = (1 + 1)/(2 + 5) ×100% = 28.6%
Precision = 1/(3 + 5) ×100% = 12.5%

Figure 4: Overview of evaluation metrics for TROVE, including source tracing and relationship classification.

fore, we adopt a sliding-window approach for samples where the source text exceeds the model's maximum length limit, denoted as M. Specifically, the input text is split into chunks of $0-M$, $M-2\times M$, $2\times M-3\times M$, etc. Each chunk is processed independently, and the final result is obtained by merging the predictions of all chunks.

During GPT-4o provenance, initial retrieval significantly enhances GPT's recall rate. Thus, each model is evaluated under two approaches: (1) direct prompting tracing, where the model processes the input directly, and (2) retrieval-augmented tracing, where retrieval is performed first, followed by tracing based on the retrieved results.

5.2 Provenance Automatic Evaluation

We propose an evaluation method to assess model performance on this task, as shown in Figure 4.

First, to evaluate model accuracy in tracing target sentences and texts, we introduce macro-average and micro-average metrics at the sentence level. Macro-average metrics compute precision and recall for each sentence and average them across all target sentences in a sample. In contrast, micro-average metrics aggregate true predicted categories across all target sentences and calculate precision and recall based on global statistics.

In addition to evaluating source sentence tracing, we assess the model's ability to determine relationships between traced and target sentences. Our evaluation system includes 13 metrics for both source tracing and relationship classification: macro-average and micro-average precision and recall. Specifically, we compute **Macro-Track-P**, **Macro-Track-R**, **Micro-Track-P**, and **Micro-Track-R** for source tracing, as well as **Macro-Relation-P**, **Macro-Relation-R**, **Micro-Relation-P**, and **Micro-Relation-R** for relationship classification. To intuitively compare models, we calculate the F1 scores for Macro-Track-P, Macro-Relation-P, Micro-Track-P, and Micro-Relation-P, averaging them to derive the overall **F1-score**.

5.3 Evaluation Results

Impact of Retrieval-Augmented Tracing vs. Direct Prompting Tracing. Across almost all models, retrieval-augmented tracing outperforms direct prompting in F1 scores, often by a large margin. For example, Qwen2.5-14B's F1 jumps from 26.02 to 40.68 with retrieval, while ChatGLM-6B, which nearly fails in direct prompting tracing with an F1 of 0.02, improves to 3.47. Even closed-source models show the same trend, as Gemini-1.5-Pro significantly improves from an F1 of 9.61 to 51.18 with retrieval. It suggests that retrieval helps overcome context-length limits and brings in the relevant source text, making it much easier for models to match target sentences with their sources.

Impact of Model Size. As shown in Table 3, larger models generally achieve higher scores in source tracing and relationship classification. For example, Qwen2.5-14B (retrieval) outperforms its smaller counterparts in most metrics, such as Track-P and Relation-P. However, Qwen2.5-7B (retrieval) achieves the highest Track-R scores, indicating that smaller models can also perform well in specific aspects of source tracing even if they do not lead in the overall F1-score. While the trend favors larger models, specific architectures or training strategies allow smaller models to remain competitive in the provenance task. Notably, for relationship classification, the advantage of larger models is more

Model	Method	Macro						Micro						F1
		T-P	T-R	T-F1	R-P	R-R	R-F1	T-P	T-R	T-F1	R-P	R-R	R-F1	
Retrieval	LCS	19.71	63.42	29.41	-	-	-	19.71	61.28	29.25	-	-	-	14.67
	BM25	23.73	78.66	35.70	-	-	-	23.73	76.92	35.56	-	-	-	17.81
	Dense	17.85	69.54	28.28	-	-	-	18.85	67.41	28.11	-	-	-	14.10
	Union	33.89	76.83	46.17	-	-	-	33.16	74.99	45.13	-	-	-	22.82
Open-Source														
Vicuna-7b	DP	6.80	23.22	10.50	2.37	8.53	3.69	6.78	23.10	10.44	2.38	8.43	3.69	7.08
	RA	27.14	41.76	32.50	9.79	17.22	12.38	29.10	40.78	33.64	10.39	15.76	12.44	22.74
LLama3-8b	DP	5.16	16.63	6.97	2.03	7.21	2.96	5.45	15.50	6.43	1.84	6.20	2.49	4.71
	RA	43.74	38.19	40.61	22.49	19.51	20.82	49.81	35.04	41.07	25.42	18.40	21.33	30.96
Chatglm-6b	DP	0.02	0.04	0.02	0.01	0.00	0.01	0.04	0.02	0.03	0.01	0.01	0.01	0.02
	RA	3.68	4.06	3.84	1.50	1.76	1.60	11.97	3.98	5.93	4.85	1.73	2.53	3.47
Qwen1.5-7b	DP	6.47	41.00	11.18	1.36	11.71	2.36	6.50	40.82	10.84	1.27	10.04	2.20	6.65
	RA	35.99	53.25	42.26	11.02	19.08	13.80	34.83	52.19	41.00	10.26	16.48	12.48	27.39
Qwen2.5-7b	DP	8.88	49.56	14.77	2.89	15.49	4.81	7.67	49.50	12.94	1.93	13.56	3.29	8.95
	RA	41.96	**71.92**	52.23	14.73	28.56	19.23	39.50	**69.72**	49.65	12.32	24.55	16.21	34.33
Qwen1.5-14b	DP	12.50	38.97	16.84	3.71	13.35	5.25	13.36	37.51	16.93	3.65	11.32	4.74	10.94
	RA	45.07	53.68	48.33	14.06	20.96	16.72	47.99	51.65	49.20	14.50	17.90	15.93	32.54
Qwen2.5-14b	DP	29.77	52.25	36.12	15.11	27.37	18.69	29.50	49.55	33.01	14.37	24.74	16.24	26.02
	RA	**54.60**	50.24	**51.99**	**29.22**	27.43	28.12	**64.68**	47.02	**54.23**	**33.49**	24.80	**28.37**	**40.68**
DeepSeek-V3 (671B)	DP	44.79	56.56	49.80	21.88	28.40	24.63	39.54	54.31	44.95	17.46	25.61	20.41	34.95
	RA	49.17	55.85	51.94	26.10	**31.20**	**28.24**	50.75	53.76	51.93	26.17	**28.54**	27.19	39.83
Closed-Source														
GPT-4o	DP	59.31	55.46	57.18	36.55	33.98	35.15	57.32	52.43	54.55	34.39	31.61	32.81	44.92
	RA	73.14	55.45	62.72	42.68	32.87	36.94	**74.81**	51.59	60.84	42.93	30.25	35.38	48.97
Gemini-1.5-pro	DP	13.30	13.00	13.10	7.54	7.44	7.46	11.38	11.71	11.49	6.29	6.54	6.40	9.61
	RA	**74.13**	58.53	**64.94**	**45.02**	34.45	**38.75**	73.80	54.62	**62.43**	**46.00**	**33.52**	**38.58**	**51.18**
Kimi	DP	39.75	47.29	43.12	20.38	24.82	22.34	36.01	44.65	39.68	17.52	22.03	19.42	31.14
	RA	60.27	**64.69**	62.25	32.44	**36.05**	34.07	57.25	**61.50**	59.15	30.32	33.41	31.73	46.80

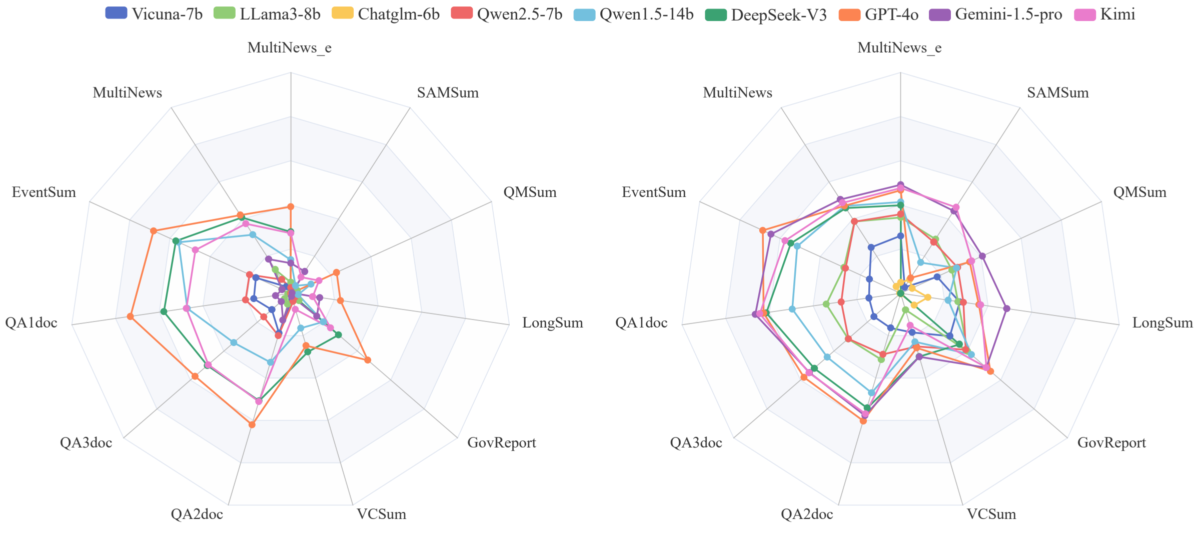
Table 3: Experiment results of LLMs. DP and RA denote direct prompting tracing and retrieval-augmented tracing. In both open-source and closed-source models, pink denotes the best DP results, while green marks the best RA results. The **bold values** highlight the best results within open and closed-source models, respectively. Since the union retrieval method outperforms each single retrieval method, we use the union retrieval method in RA.

consistent, suggesting that capturing complex relationships (such as paraphrasing, summarization, and logical inference) demands the enhanced representational capacity of increased parameterization.

Precision–Recall Trade-offs Across Models. We can find some interesting trade-offs when examining the precision and recall metrics for each model. Some models, like Qwen2.5-7B with retrieval, prioritize recall, identifying more traced sources with a recall of 71.92, but at the cost of lower precision at 41.96. Others, such as Qwen2.5-14B with retrieval, achieve a better balance, reaching a higher precision of 54.60 while maintaining a recall of 50.24. In real-world applications, a high-recall system may be preferable when capturing all possible source sentences, which is crucial, even if some false positives appear. On the other hand, a precision-focused system is better suited when avoiding false positives is a priority.

Open-Source vs. Closed-Source. Among open-source models, parameter sizes vary widely, from a few billion (e.g., 6B–14B) to the much larger Deepseek V3 with 671B parameters. Despite these differences, larger models generally perform better in direct prompting and retrieval-augmented settings, especially in relationship classification. Deepseek-V3 (DP) shows strong performance with an F1 score of 34.95, outperforming many smaller models. However, when retrieval is applied, models like Qwen2.5-14B begin to reduce the gap with leading closed-source systems. For closed-source models, Gemini-1.5-Pro (RA) and GPT-4o (RA) achieve the highest F1 scores at 51.18 and 48.97, performing well in both source tracing and relationship classification. However, Gemini-1.5-Pro struggles with direct prompting, with an F1 score of only 9.61, highlighting the importance of retrieval. While closed-source models still lead overall, their advantage is reduced significantly when open-source LLMs use strong retrieval methods.

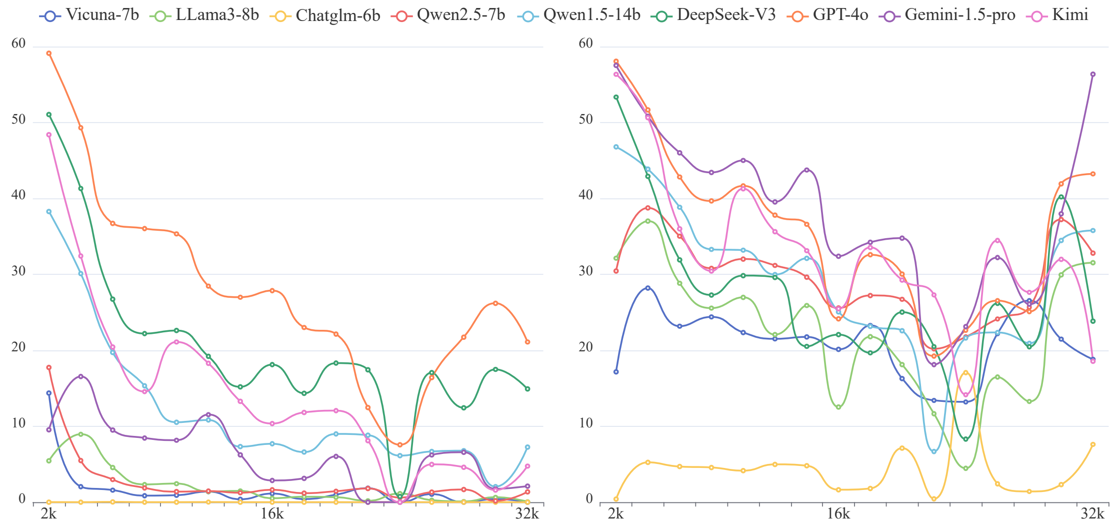
Relationship Classification. Besides source sentence tracing, models must identify the relationship between traced and target sentences (e.g., quotation, compression, and inference). Relationship



(a) Direct Prompting

(b) Retrieval-Augmented

Figure 5: The performance of different models varies on different scenarios.



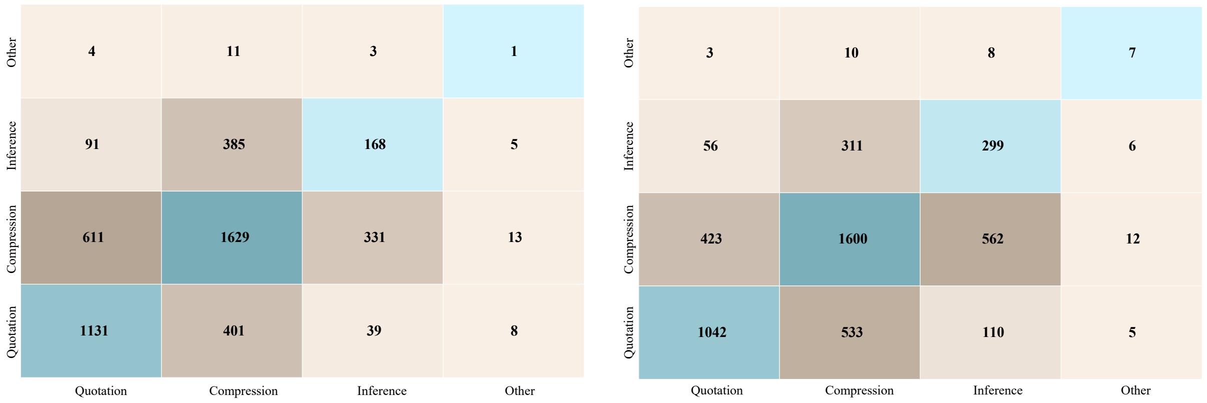
(a) Direct Prompting

(b) Retrieval-Augmented

Figure 6: The performance of different models varies on different sourth lengths.

classification is more challenging than sentence tracing, requiring the model to understand deeper semantic and structural differences. Larger models (e.g., Qwen2.5-14B, Deepseek-V3, and GPT-4o) tend to perform more consistently, showing higher precision and recall in relationship classification than smaller open-source models. However, no model achieves highly reliable performance, suggesting that accurately capturing deep semantic relationships remains a challenging problem.

To conclude our analysis, we highlight the following key insights: 1) **Retrieval is essential.** Every model benefits significantly from retrieval, often turning poor performance in direct prompting into much stronger results when relevant context is provided. 2) **Larger models handle complex tasks better.** Larger models tend to perform better in relationship classification, indicating that richer representations are crucial for handling complex tasks. 3) **Precision and recall involve trade-offs.** Some models focus on capturing more potential sources, leading to higher recall but lower precision, while others do the opposite. The choice between high recall and high precision depends on the specific application. 4) **Closed-source models dominate, but open-source is catching up.** Models like Gemini-1.5-Pro and GPT-4o achieve the highest F1 scores, maintaining a clear advantage. However, retrieval-augmented open-source models, such as Qwen2.5-14B, are making significant progress and, in some cases, reaching comparable performance. 5) **Relationship classification remains a challenge.** No model achieves consistently strong performance in detecting complex relationships, showing that there is still room for improvement in fine-grained provenance tasks.



(a) Confusion matrix for GPT-4o (DP).

(b) Confusion matrix for GPT-4o (RA).

Figure 7: Confusion matrix for GPT-4o. X-axis: predictions, y-axis: ground truth.

5.4 Analysis

Figure 5 and Figure 6 show model performance across scenarios and source lengths respectively.

Models Performance Across Scenarios. Under the DP method, GPT-4o performs better than all other models across all scenarios. With the RA method, Kimi, GPT-4o, and Gemini-1.5-Pro each exhibit distinct advantages in different scenarios. For instance, GPT-4o leads in EventSum, QA3doc, and QA2doc; Gemini-1.5-Pro outperforms others in MultiNews, MultiNews_e, QMSum, QA1doc, and LongSum; and Kimi shows outstanding performance in SAMSum. These results suggest that each model adapts differently depending on the scenario. They also show that RA leads to more significant improvements than DP, especially in multi-document, dialogue, and meeting scenarios.

Models Performance Across Source Lengths. Regarding source length, the RA is generally less affected by longer texts. In the DP, once the source length reaches 32k, only GPT-4o and DeepSeek-V3 maintain a passable but somewhat lower level of performance, while the others see a significant drop. Interestingly, Qwen2.5-14b usually falls behind Kimi and Gemini-1.5-Pro, but it surpasses both when the source length exceeds 20k.

Error Analysis. To understand model behavior, we analyze the confusion matrices for GPT-4o under both direct prompting and retrieval-augmented conditions, with results shown in Figure 7a and Figure 7b respectively. The matrices reveal a clear hierarchy of relationship difficulty: *Quotation* is easiest, followed by *Compression*, while *Inference* proves most challenging. Three error patterns emerge: (1) Compression bias — models overpredict this category, with 533 *Quotation* and 311 *Inference* instances misclassified as *Compression* in RA; (2) "Inference→Compression" confusion — 311 out of 672 true *Inference* cases are misclassified as *Compression*, indicating difficulty distinguishing between summarization and logical derivation; (3) "Other" underrepresentation — only 7 out of 28 instances correctly identified, highlighting challenges with rare relationship types.

Impact of Retrieval. While RA improves overall performance (*Inference* correct predictions increase from 168 to 299), it also intensifies misclassification attempts. The Quotation→Compression errors increase from 401 in DP to 533 in RA, suggesting that additional context sometimes causes models to overinterpret simple quotations as more complex relationships. These patterns reveal fundamental challenges in relationship classification that extend beyond performance metrics.

6 Conclusion

We present TROVE, a fine-grained text provenance challenge to enhance transparency and accountability in text generation. TROVE traces each target sentence to its source, classifying their relationship as *quotation*, *compression*, *inference*, or *others*. TROVE offers a rigorous foundation for understanding where and how text is derived. Our dataset construction leverages three public datasets, LongBench, LooGLE, and CRUD-RAG, covering 11 scenarios, 2 languages, and 3 source length ranges.

Experiments with major LLMs show that retrieval augmentation significantly improves performance, especially for multi- and long-document settings. Larger models handle complex target-source relationships better, and while closed-source models lead in performance, open-source models reduce the gap with retrieval methods. However, relationship classification remains a key challenge.

Limitations

We conclude the limitations of our study as follows: (1) Lack of Hallucination Cases. Our dataset construction relies on existing public datasets rather than texts generated directly by language models. As a result, hallucinations are absent in TROVE. In future work, we will enrich the dataset by incorporating model-generated content. (2) Scalability and Context Window Constraints. Although we include long-document and multi-document settings, current LLMs are constrained by finite context windows. In extremely lengthy documents, crucial source sentences might be ignored during retrieval.

Acknowledgements

The research work described in this paper has been partially supported by the Natural Science Foundation of China under Grant No. 62336008. We sincerely thank the anonymous reviewers for their insightful comments and constructive suggestions.

References

Rami Aly, Zhiqiang Tang, Samson Tan, and George Karypis. 2024. Learning to generate answers with citations via factual consistency models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 11876–11896.

Jinze Bai, Shuai Bai, Yunfei Chu, et al. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Yushi Bai, Xin Lv, Jiajie Zhang, et al. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3119–3137.

Faeze Brahman, Baolin Peng, Michel Galley, et al. 2022. Grounded keys-to-text generation: Towards factual open-ended generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7397–7413.

Shuyang Cao and Lu Wang. 2024. Verifiable generation with subsentence-level fine-grained citations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15584–15596.

Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022. Gere: Generative evidence retrieval for fact verification. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, page 2184–2189.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024. Complex claim verification with evidence retrieved in the wild. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 3569–3587.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, et al. 2023. Vicuna: An open source chatbot impressing gpt-4 with 90%* chatgpt quality.

Svetlana Churina, Anab Maulana Barik, and Saisamarth Rajesh Phaye. 2024. Improving evidence retrieval on claim verification pipeline through question enrichment. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 64–70.

DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, et al. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *Preprint*, arXiv:2405.04434.

Rosa Falotico and Piero Quatto. 2015. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49:463–470.

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 (EMNLP)*, pages 6465–6488.

Ameya Godbole, Nicholas Monath, Seungyeon Kim, et al. 2024. Analysis of plan-based retrieval for grounded text generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 13101–13119.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nianlong Gu and Richard Hahnloser. 2024. Controllable citation sentence generation with language models. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 22–37.

I-Hung Hsu, Zifeng Wang, Long Le, et al. 2024. CaLM: Contrasting large and small language models to verify grounded generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12782–12803.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuxuan Gu, Weihong Zhong, Xiachong Feng, Weijiang Yu, Weihua Peng, Duyu Tang, Dandan Tu, and Bing Qin. 2024. Learning fine-grained grounded citations for attributed large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14095–14113.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics (TACL)*, 6:391–406.

Anne Lauscher, Brandon Ko, Bailey Kuehl, et al. 2022. MultiCite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1875–1889.

Weitao Li, Junkai Li, Weizhi Ma, and Yang Liu. 2024. Citation-enhanced generation for LLM-based chatbots. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1451–1466.

Yu Li, Baolin Peng, Yelong Shen, Yi Mao, et al. 2022. Knowledge-grounded dialogue generation with a unified knowledge representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 206–218.

Hongyuan Lu, Wai Lam, Hong Cheng, and Helen Meng. 2022. On controlling fallback responses for grounded dialogue generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2591–2601.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics (TACL)*, 9:329–345.

Yuanjie Lyu, Zhiyu Li, Simin Niu, Feiyu Xiong, et al. 2024. Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models. *ACM Transactions on Information Systems*, pages 1–32.

Shrikant Malviya and Stamos Katsigiannis. 2024. SK_DU team: Cross-encoder based evidence retrieval and question generation with improved prompt for the AVeriTeC shared task. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 99–107.

Biswadip Mandal, Xiangci Li, and Jessica Ouyang. 2024. Contextualizing generated citation texts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3849–3854.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, et al. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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

Munjiah Nur Saadah, Rigga Widar Atmagi, Dyah S Rahayu, and Agus Zainal Arifin. 2013. Information retrieval of text document with weighting tf-idf and lcs. *Jurnal Ilmu Komputer dan Informasi*, 6(1):34–37.

Aviv Slobodkin, Eran Hirsch, Arie Cattan, et al. 2024. Attribute first, then generate: Locally-attributable grounded text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3309–3344. Association for Computational Linguistics.

Kaiqiang Song, Chen Li, Xiaoyang Wang, Dong Yu, and Fei Liu. 2022. Towards abstractive grounded summarization of podcast transcripts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4407–4418.

Xinyu Xing, Xiaosheng Fan, and Xiaojun Wan. 2020. Automatic generation of citation texts in scholarly papers: A pilot study. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6181–6190.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, et al. 2023. Glm-130b: An open bilingual pre-trained model. *Preprint*, arXiv:2210.02414.

Liwen Zheng, Chaozhuo Li, Xi Zhang, Yu-Ming Shang, Feiran Huang, and Haoran Jia. 2024. Evidence retrieval is almost all you need for fact verification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9274–9281.

Wanjun Zhong, Jingjing Xu, Duyu Tang, et al. 2020. Reasoning over semantic-level graph for fact checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6170–6180.

Anni Zou, Zhuosheng Zhang, and Hai Zhao. 2023. Decker: Double check with heterogeneous knowledge for commonsense fact verification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11891–11904.

A Dataset

Table 4 presents a detailed statistical overview of our dataset, categorized across multiple dimensions: document type (single vs. multi-document), language (English vs. Chinese), scenario, domain, origin dataset, and average document length.

Our dataset consists of English and Chinese sources, covering multiple scenarios such as news summarization, academic summarization, and question answering. It includes domains such as news, government reports, scientific papers, meetings, and dialogues, ensuring broad coverage across different textual data types. The origin datasets include well-established resources, i.e., Long-Bench, LooGLE, and CRUD.

To account for variations in document length, we report #AvgLen, which measures the average length of source documents in words for English texts and characters for Chinese texts. Multi-document datasets (e.g., MultiNews) tend to have longer text sequences, while single-document datasets vary significantly based on their domain (e.g., academic papers in LongSum have much longer texts than news articles in EventSum).

B Detail Experiment Results

We present experimental results for open-source and closed-source LLMs under single- and multi-document settings in English and Chinese. Table 6 shows the results for open-source models. Table 7 provides results for closed-source models.

We report metrics for direct prompting (DP) and retrieval-augmented (RA) tracing. Each table includes macro- and micro-averaged precision, recall, and F1 metrics for source tracing (T), relationship classification (R), and an overall F1 score.

In single-document English tasks, among open-source models, Qwen2.5-7B with retrieval-augmented tracing achieves the highest F1 (35.02), outperforming other open-source alternatives (e.g., Qwen2.5-14B with 34.50). However, the closed-source Gemini-1.5-Pro obtains an even higher F1 of 48.39 with retrieval, making it the top performer overall in this single-document English scenario. Notably, GPT-4o is quite capable under direct prompting (33.97), exceeding the retrieval-augmented baselines of most open-source LLMs. However, almost all models (open or closed) show significant gains when retrieval is introduced.

In single-document Chinese tasks, among open-source models, DeepSeek-V3 (RA) leads with an F1 score of 44.52, outperforming Qwen and Llama. Among closed-source models, GPT-4o (DP) scores 42.99, while Gemini-1.5-Pro (RA) gets higher at 46.95. Although these closed-source models exceed most open-source options except DeepSeek-V3, GPT-4o also performs well without retrieval, scoring 43.61 with direct prompting, even better than its RA variant. In contrast, Gemini relies heavily on retrieval, as shown by its sharp jump from a very low direct prompting score of 2.85 to 46.95 when retrieval is applied. This highlights the varying levels of dependence on reducing candidates among different models.

In multi-document English tasks, Qwen2.5-14b (RA) leads among open-source models with an F1 score of 44.54, slightly ahead of DeepSeek-V3 (43.39). However, the closed-source Gemini-1.5-Pro gets the top score with 51.25, outperforming GPT-4o (48.34) and Kimi (49.48). GPT-4o also shows strong performance without retrieval, scoring 41.26, while Gemini struggles with a much lower 15.80. This suggests that GPT-4o is naturally better at direct prompting, whereas Gemini and Kimi depend more on retrieved context to handle complex multi-document provenance.

In multi-document Chinese tasks, DeepSeek-V3 (RA) leads open-source models with an F1 score of 54.52, far ahead of Qwen2.5-14B (47.51). However, GPT-4o achieves the best overall with 61.09, just ahead of Gemini-1.5-Pro (58.11) and Kimi (57.33). This highlights GPT-4o's strong ability to handle multi-source Chinese text.

In all, in single-document tasks, Qwen2.5-7B and DeepSeek-V3 emerge as strong open-source choices for English and Chinese, respectively, yet Gemini-1.5-Pro can outperform them once retrieval is incorporated. GPT-4o stands out for its relatively high direct-prompting scores across both languages, showing strong built-in tracing capabilities. Under multi-document conditions, the complexity increases, and the top results often come from closed-source solutions (e.g., Gemini-1.5-pro, GPT-4o, Kimi), although Qwen2.5-14b and DeepSeek-V3 hold their own in the open-source domain. Models integrating retrieval, whether open- or closed-source, generally exhibit greater gains and more accurate sentence-level provenance.

B.1 Confusion Matrix Analysis Details

The observed discrepancies in total counts between confusion matrices for different methods are attributed to the following methodological factors:

#Doc	Lang	Tasks	Number	Origin Dataset	Domain	#AvgLen
multi	en	MultiNews_e	133	Long-Bench	News	8,672.77
		MultiNews	47	Long-Bench	News	3,118.66
single	en	GovReport	63	Long-Bench	Report	6,836.30
		LongSum	15	LooGLE	ArXiv	21,797.40
		QMSum	47	Long-Bench	Meeting	9,222.79
		SAMSum	55	Long-Bench	Dialogue	7,587.96
multi	zh	QA2doc	90	CRUD	News	713.97
		QA3doc	90	CRUD	News	1,070.50
single	zh	EventSum	32	CRUD	News	758.97
		QA1doc	32	CRUD	News	676.41
		VCSum	116	Long-Bench	Meeting	11,993.00

Table 4: Detailed statistics of our dataset. #AvgLen denotes the average length of the source document(s), measured in Chinese characters for Chinese texts and words for English texts. *Tasks* indicates the data's original task (scenario).

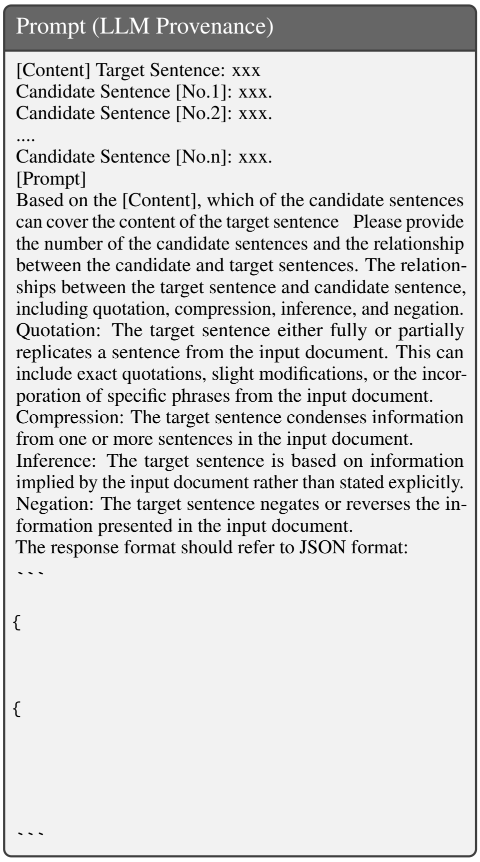
Final-Lable	Pre-Label
Quotion	Copy
	Paraphrase
	Reordering
Compression	Fusion
	Summary
	Distillation
Inference	Inference
	Expansion
	Generalization
Other	Negation

Table 5: Mapping between pre-labels and final-labels.

(1) The retrieval-augmented method may fail to retrieve certain sentences from the source documents, leading to variations in the number of ground-truth sentences available for classification. (2) The retrieval-augmented and direct prompting methods trace different sets of source sentences due to their distinct retrieval mechanisms. Sentences that remain untraced by either method are excluded from the subsequent relationship classification task, resulting in different sample sizes across experimental conditions. It is worth noting that we employ the pass@5 evaluation metric for all experimental assessments to ensure consistent and robust performance measurement.

C Prompts

To prevent large language models from mislabeling, the pre-labeling process of GPT-4o adopts a more fine-grained classification, specifically as: Copy, Paraphrase, Summary, Inference, Expansion, Fusion, Distillation, Reordering, Negation, Generalization. And the mapping between pre-labels and final labels is shown in Table 5.



Prompt (LLM Provenance)

~~~
[Content] Target Sentence: xxx
Candidate Sentence [No.1]: xxx.
Candidate Sentence [No.2]: xxx.
....
Candidate Sentence [No.n]: xxx.
[Prompt]
Based on the [Content], which of the candidate sentences can cover the content of the target sentence   Please provide the number of the candidate sentences and the relationship between the candidate and target sentences. The relationships between the target sentence and candidate sentence, including quotation, compression, inference, and negation.
Quotation: The target sentence either fully or partially replicates a sentence from the input document. This can include exact quotations, slight modifications, or the incorporation of specific phrases from the input document.
Compression: The target sentence condenses information from one or more sentences in the input document.
Inference: The target sentence is based on information implied by the input document rather than stated explicitly.
Negation: The target sentence negates or reverses the information presented in the input document.
The response format should refer to JSON format:
```

{


{



```
~~~

Figure 8: Prompt for LLM provenance in experiments.

LLM	#Doc	Lang.	Method	Macro						Micro						F1
				T_P	T_R	T_F1	R_P	R_R	R_F1	T_P	T_R	T_F1	R_P	R_R	R_F1	
Vicuna	single	en	DP	0.45	2.41	0.76	0.21	1.52	0.37	0.42	2.47	0.72	0.19	1.43	0.34	0.55
			RA	29.82	46.96	36.48	10.50	19.69	13.69	32.01	45.45	37.56	11.05	16.64	13.28	25.25
		zh	DP	8.97	30.53	13.86	3.19	10.94	4.94	8.52	30.51	13.33	3.04	11.08	4.77	9.22
			RA	29.75	32.47	31.05	10.12	12.75	11.28	30.34	32.81	31.53	10.26	12.63	11.33	21.30
	multi	en	DP	2.45	10.85	4.00	0.82	3.64	1.33	2.21	10.70	3.66	0.68	3.55	1.15	2.54
			RA	30.89	48.43	37.72	11.53	21.08	14.91	34.62	46.45	39.68	12.84	18.39	15.12	26.86
		zh	DP	15.33	49.10	23.37	5.25	18.01	8.13	15.96	48.70	24.04	5.60	17.66	8.50	16.01
			RA	18.09	39.17	24.75	7.03	15.34	9.64	19.43	38.40	25.81	7.42	15.39	10.02	17.55
LLama3	single	en	DP	3.53	30.86	6.33	1.37	12.05	2.45	2.55	29.67	4.69	0.95	10.75	1.75	3.81
			RA	55.74	43.15	48.64	24.80	18.66	21.30	58.79	39.01	46.90	25.79	17.86	21.10	34.49
		zh	DP	1.51	1.70	1.60	0.52	0.65	0.58	2.10	1.46	1.72	0.59	0.45	0.51	1.10
			RA	27.66	20.11	23.29	14.95	11.26	12.84	32.05	19.55	24.28	17.24	11.30	13.65	18.52
	multi	en	DP	8.24	26.55	12.57	4.57	13.90	6.88	7.12	23.76	10.96	3.75	11.58	5.67	9.02
			RA	49.07	47.30	48.17	26.08	25.94	26.01	55.06	42.50	47.97	29.72	23.10	26.00	37.04
		zh	DP	7.37	7.42	7.39	1.67	2.25	1.92	10.04	7.13	8.34	2.05	2.03	2.04	4.92
			RA	42.50	42.18	42.34	24.14	22.19	23.13	53.36	39.13	45.15	28.94	21.35	24.57	33.80
Chatglm	single	en	DP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			RA	8.95	9.73	9.32	3.80	4.58	4.16	25.06	9.92	14.21	11.35	4.70	6.65	8.58
		zh	DP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			RA	0.42	0.20	0.27	0.28	0.06	0.10	0.83	0.17	0.29	0.28	0.06	0.10	0.19
	multi	en	DP	0.06	0.17	0.09	0.04	0.01	0.02	0.16	0.08	0.11	0.02	0.02	0.02	0.06
			RA	5.32	6.23	5.74	1.91	2.39	2.13	21.90	5.79	9.15	7.75	2.15	3.37	5.10
		zh	DP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			RA	0.02	0.07	0.04	0.00	0.00	0.00	0.10	0.05	0.07	0.00	0.00	0.00	0.03
Qwen1.5 -7b	single	en	DP	1.30	30.84	2.50	0.39	12.05	0.76	1.29	30.19	2.47	0.40	9.52	0.76	1.62
			RA	33.91	57.06	42.54	11.07	22.61	14.87	31.66	55.25	40.26	9.96	18.58	12.97	27.66
		zh	DP	7.22	37.80	12.13	1.91	10.51	3.23	7.07	37.76	11.91	1.87	10.20	3.16	7.61
			RA	43.16	42.11	42.63	10.25	11.65	10.90	42.95	42.11	42.53	10.01	11.02	10.49	26.64
	multi	en	DP	2.04	29.08	3.81	0.48	8.94	0.91	1.88	28.85	3.53	0.48	7.17	0.91	2.29
			RA	36.45	54.01	43.52	14.62	24.34	18.27	34.91	50.12	41.15	13.16	19.59	15.75	29.67
		zh	DP	16.40	66.28	26.30	2.67	15.33	4.54	15.75	66.48	25.46	2.34	13.28	3.98	15.07
			RA	30.44	59.81	40.35	8.13	17.71	11.15	29.79	61.26	40.08	7.89	16.73	10.72	25.58
Qwen2.5 -7b	single	en	DP	1.41	25.92	2.68	0.67	9.42	1.25	1.03	25.63	1.97	0.29	6.95	0.55	1.61
			RA	44.35	67.54	53.54	15.43	28.65	20.06	41.56	64.62	50.58	12.21	22.88	15.92	35.02
		zh	RA	10.38	46.09	16.95	3.54	14.21	5.67	9.30	46.08	15.48	2.68	13.32	4.46	10.64
			RA	43.67	59.50	50.37	12.06	20.57	15.20	42.69	59.01	49.54	11.32	19.68	14.38	32.37
	multi	en	DP	4.36	40.63	7.87	2.27	14.96	3.94	2.59	40.36	4.87	0.72	12.24	1.35	4.51
			RA	45.50	70.35	55.26	20.70	33.92	25.71	41.16	66.15	50.74	16.45	27.40	20.55	38.07
		zh	DP	19.36	85.59	31.58	5.10	23.38	8.37	17.75	85.93	29.43	4.03	21.75	6.80	19.04
			RA	34.32	90.28	49.73	10.73	31.09	15.96	32.60	89.09	47.74	9.29	28.26	13.98	31.85
Qwen1.5 -14b	single	en	DP	0.77	31.02	1.50	0.17	7.48	0.33	0.72	27.75	1.41	0.13	5.76	0.26	0.88
			RA	46.18	44.82	45.49	9.95	14.05	11.65	50.84	43.02	46.61	12.25	12.13	12.19	28.99
		zh	DP	15.79	32.88	21.34	4.84	10.01	6.52	15.73	32.76	21.26	4.46	9.54	6.08	13.80
			RA	45.66	41.57	43.52	11.60	14.17	12.75	47.04	41.15	43.90	12.02	13.12	12.55	28.18
	multi	en	DP	6.37	48.85	11.27	2.85	22.98	5.07	4.38	46.69	8.01	1.58	18.51	2.92	6.82
			RA	50.35	64.90	56.71	23.14	32.69	27.10	49.75	60.59	54.64	20.70	26.41	23.21	40.41
		zh	DP	27.05	43.14	33.25	6.99	12.95	9.08	32.60	42.85	37.03	8.44	11.46	9.72	22.27
			RA	38.09	63.45	47.61	11.57	22.93	15.38	44.33	61.84	51.64	13.02	19.92	15.75	32.59
Qwen2.5 -14b	single	en	DP	15.92	49.57	24.10	9.88	25.78	14.29	7.90	45.88	13.48	4.40	22.10	7.34	14.80
			RA	44.43	42.18	43.28	24.75	23.71	24.22	60.35	37.72	46.43	31.53	19.50	24.10	34.50
		zh	DP	36.61	50.26	42.36	17.77	25.25	20.86	36.70	49.58	42.18	17.12	24.10	20.02	31.35
			RA	57.24	39.46	46.71	29.73	21.04	24.64	62.04	39.30	48.12	31.79	20.85	25.18	36.16
	multi	en	DP	21.86	62.59	32.40	12.43	36.63	18.56	16.92	57.50	26.14	8.92	31.71	13.92	22.76
			RA	53.62	57.14	55.33	31.55	34.13	32.79	65.04	51.61	57.55	36.25	29.43	32.49	44.54
		zh	DP	44.69	46.59	45.62	20.37	21.81	21.06	56.50	45.24	50.25	27.04	21.07	23.69	35.16
			RA	63.09	62.18	62.63	30.86	30.84	30.85	71.27	59.46	64.83	34.40	29.42	31.72	47.51
DeepSeek-V3	single	en	DP	17.34	23.46	19.94	11.26	15.33	12.98	11.53	21.44	15.00	6.83	12.31	8.78	14.17
			RA	18.22	25.15	21.13	10.56	16.07	12.75	20.07	23.07	21.47	11.27	13.24	12.18	16.88
		zh	DP	51.57	62.88	56.67	23.24	29.34	25.94	50.08	62.59	55.64	21.56	28.57	24.58	40.71
			RA	63.86	56.21	59.79	31.08	28.85	29.92	63.37	55.65	59.26	30.15	28.16	29.12	44.52
	multi	en	DP	39.48	60.90	47.91	23.39	35.69	28.26	26.40	56.82	36.05	14.01	30.79	19.26	32.87
			RA	48.45	59.18	53.28	29.25	36.83	32.61	55.20	55.54	55.37	31.88	32.74	32.30	43.39
		zh	DP	70.78	79.02	74.67	29.63	33.24	31.33	70.15	76.37	73.13	27.45	30.77	29.02	52.04
			RA	66.13	82.86	73.56	33.52	43.04	37.69	64.37	80.78	71.65	31.39	40.01	35.18	54.52

Table 6: Experiment results of open-source LLMs under single- and multi-document settings in English and Chinese. DP and RA denote direct prompting tracing and retrieval-augmented tracing.

LLM	#Doc	Lang.	Method	Macro						Micro						F1
				T_P	T_R	T_F1	R_P	R_R	R_F1	T_P	T_R	T_F1	R_P	R_R	R_F1	
GPT-4o	single	en	DP	46.54	41.11	43.66	28.69	25.08	26.76	46.72	37.16	41.40	27.16	21.68	24.11	33.98
			RA	65.96	50.93	57.48	35.99	27.69	31.29	70.60	46.50	56.07	36.01	24.30	29.02	43.47
		zh	DP	59.17	48.70	53.43	38.02	31.44	34.42	58.57	48.25	52.91	37.27	30.71	33.68	43.61
			RA	75.47	43.72	55.36	41.82	24.84	31.17	75.76	42.81	54.71	41.70	24.32	30.73	42.99
	multi	en	DP	50.62	55.20	52.81	33.67	35.36	34.49	45.03	50.64	47.67	28.38	31.93	30.05	41.26
			RA	64.98	58.04	61.31	39.79	36.00	37.80	67.07	52.12	58.66	40.26	31.85	35.57	48.34
		zh	DP	80.90	76.83	78.81	45.83	44.06	44.93	78.95	73.67	76.22	44.75	42.11	43.39	60.84
			RA	86.15	69.12	76.70	53.14	42.96	47.51	85.81	64.93	73.92	53.76	40.51	46.20	61.09
Gemini-1.5 -pro	single	en	DP	16.01	13.85	14.86	8.83	7.40	8.05	12.85	11.57	12.17	6.53	5.96	6.23	10.33
			RA	69.39	55.51	61.68	42.20	31.83	36.28	70.53	50.68	58.98	44.17	31.30	36.64	48.39
		zh	DP	4.34	3.81	4.06	1.95	1.73	1.83	3.98	3.61	3.79	1.80	1.67	1.73	2.85
			RA	77.54	45.72	57.52	49.98	29.20	36.86	76.98	44.75	56.60	50.00	29.16	36.84	46.95
	multi	en	DP	19.53	22.00	20.69	12.53	14.12	13.28	16.41	19.64	17.88	10.51	12.35	11.35	15.80
			RA	67.21	64.36	65.75	40.16	38.04	39.07	66.92	58.74	62.56	40.23	35.31	37.61	51.25
		zh	DP	13.31	12.33	12.80	6.86	6.49	6.67	12.27	12.01	12.13	6.33	6.20	6.26	9.47
			RA	82.38	68.53	74.82	47.75	38.75	42.78	80.78	64.30	71.60	49.60	38.31	43.23	58.11
Kimi	single	en	DP	24.62	30.16	27.11	13.05	15.65	14.23	19.51	27.57	22.85	9.92	13.21	11.33	18.88
			RA	63.26	62.36	62.81	30.85	31.43	31.14	59.41	58.49	58.94	28.53	29.07	28.79	45.42
		zh	DP	29.54	32.02	30.73	15.70	17.22	16.43	28.31	31.59	29.86	14.59	16.35	15.42	23.11
			RA	46.88	43.39	45.07	25.98	25.13	25.55	46.07	42.72	44.33	25.40	24.42	24.90	34.96
	multi	en	DP	37.73	51.20	43.45	21.44	29.85	24.96	30.42	46.25	36.70	15.77	24.77	19.27	31.09
			RA	60.13	70.01	64.70	35.08	42.29	38.35	56.07	64.78	60.11	32.00	38.05	34.76	49.48
		zh	DP	67.12	75.76	71.18	31.32	36.54	33.73	65.79	73.18	69.29	29.79	33.80	31.67	51.47
			RA	70.82	83.01	76.43	37.84	45.33	41.25	67.46	80.01	73.20	35.37	42.11	38.45	57.33

Table 7: Experiment results of closed-source LLMs under single- and multi-document settings in English and Chinese. DP and RA denote direct prompting tracing and retrieval-augmented tracing.

Prompt (GPT-4o Provenance)

[Content] Target Sentence 1:
Candidate Sentence 1: xxx.
Candidate Sentence 2: xxx.
...
Target Sentence 2:
Candidate Sentence 1: xxx.
Candidate Sentence 2: xxx.
...
[Prompt]
Given a target sentence and a set of candidate provenance sentences, your task is to identify which candidate sentence(s) most likely served as the source for the information in the target sentence. The selected candidate sentence(s) should directly support the target sentence, ensuring that every word in the target sentence can be traced back to the selected candidate sentences, either in the exact form or as a synonym. And give the key relationships between the the target sentence and selected candidate sentence, including: copy, paraphrase, summary, inference, expansion, fusion, distillation, reordering, negation, generalization.
Key Relationships:
- Copy: The target sentence either fully or partially replicates a sentence from the input document. This can include exact quotations, slight modifications, or the incorporation of specific phrases from the input document.
- Paraphrase: The target sentence conveys the same meaning as a sentence in the input document but uses different wording.
- Summary: The target sentence condenses information from one or more sentences in the input document.
- Inference: The target sentence is based on information implied by the input document rather than stated explicitly.
- Expansion: The target sentence elaborates on or adds new details to the information found in a sentence in the input document.
- Fusion: The target sentence combines information from multiple sentences or sections of the input document.
- Distillation: The target sentence distills key points from a longer, more detailed section of the input document.
- Reordering: The target sentence presents the same information as in the input document but in a different order.
- Negation: The target sentence negates or reverses the information presented in the input document.
- Generalization/Specification: The target sentence either generalizes the information (making it less specific) or specifies it (making it more detailed) based on the input document.

Instructions:
1. For each target sentence, review the candidate provenance sentences provided and give the key relationships.
2. Select the candidate sentence(s) that best support the information in the target sentence.
3. Ensure that all elements of the target sentence are present in the selected candidate sentence(s) in the same form or as synonyms.
4. Copy the selected candidate sentence(s) exactly as they appear in the [Content] section.
5. Do not generate any new content or modify the sentences.
6. The key relationship includes: copy, paraphrase, summary, inference, expansion, fusion, distillation, reordering, negation, generalization.

The response format should refer to JSON format:

```
    {


    {



```

Figure 9: Prompt for GPT4o provenance in annotation.

文本溯源标注手册

1. 引言

文本溯源任务旨在通过追溯目标文本中的句子，识别其源自输入文档的哪些句子。这个过程有助于确定目标文本与源材料之间的关系，包括内容是直接引用、改写、总结、推理，还是以其他方式转换的。

2. 任务定义

目标：对于目标文本中的每个句子，您的任务是识别一个或多个可能作为该目标句子来源的输入文档中的句子。无论是直接引用、压缩、推理还是否定，所识别的句子应反映目标句子中所包含的信息。

关系类型

1. 直接引用（单句）：包含复制、改写、重排序

复制/部分复制（Copy）：目标句子完全或部分复制了输入文档中的句子，包括精确引用、轻微修改或采用特定短语。

改写（Paraphrasing）：目标句子传达了与输入文档中某个句子相同的含义，但使用了不同的措辞。

重排序（Reordering）：目标句子呈现了与输入文档中相同的信息，但顺序不同。

2. 压缩（单句或多句）：包含融合、总结

融合（Fusion）：目标句子将来自输入文档多个句子或部分的信息组合在一起。

总结（Summary）：目标句子对输入文档中的一个或多个句子的信息进行了精简概述。

3. 推理（单句或多句）：包含扩展、泛化和细化

扩展（Expansion）：目标句子对输入文档中某个句子的信息进行了详细阐述或添加了新细节。

泛化（Generalization）：目标句子基于输入文档中具体的信息进行了概括，使之变得不那么具体。

细化（Specification）：目标句子基于输入文档中信息进行了更详细的说明，使之更加具体。

4. 否定（单句或多句）：包含否定、矛盾和反驳

否定（Negation）：目标句子否定或推翻了输入文档中呈现的信息。

矛盾/歪曲：目标句子与输入文档中的信息存在矛盾或歪曲了原意。

反驳：目标句子对输入文档中的信息进行了反驳。

3. 标注说明

步骤1：阅读输入文档和目标文本

仔细阅读整个输入文档，理解其上下文和内容。

浏览目标文本，确定需要溯源的句子。

步骤2：识别候选句子

对于目标文本中的每个句子，在输入文档中找到包含该句子所反映信息的一个或多个句子。

在选择候选句子时，考虑所有可能的关系类型（如直接引用、融合、推理、否定）。

步骤3：选择标准

全面覆盖：确保目标句子的每个元素都可以在所选的候选句子中找到。如果输入文档中的单个句子无法完全涵盖目标句子，请考虑选择多个候选句子。

关系识别：将目标句子与适当的关系类型匹配（如改写、融合）。使用关系定义来指导您决定哪些候选句子最能反映目标句子。

同义匹配：如果目标句子使用了同义词或不同的表达方式，请确保其含义仍能在输入文档的候选句子中准确反映。

上下文相关性：考虑输入文档和目标文本中周围的上下文。从段落或章节的更广泛背景来看，候选句子可能更为相关。

步骤4：标注

从输入文档中复制您认为支持目标句子的确切候选句子。

不要生成新内容或修改现有句子。

步骤5：核对

仔细检查您的标注，确保所选的候选句子在内容、上下文和所识别的关系类型方面完全支持目标句子。

Figure 10: The guideline for human annotation.