GenProve: Learning to Generate Text with Fine-Grained Provenance

Anonymous ACL submission

Abstract

Large language models (LLM) often hallucinate, and while adding citations is a common solution, it is frequently insufficient for accountability as users struggle to verify how a cited source supports a generated claim. Existing methods are typically coarse-grained and fail to distinguish between direct quotes and complex reasoning. In this paper, we introduce **Generation-time Fine-grained Provenance**, a task where models must generate fluent answers while simultaneously producing structured, sentence-level provenance triples. To enable this, we present **ReFInE** (**Re**lation-aware **F**ine-grained **In**terpretability & **E**vidence), a dataset featuring expert-verified annotations that distinguish between *Quotation*, *Compression*, and *Inference*. Building on ReFInE, we propose **GenProve**, a framework that combines Supervised Fine-Tuning (SFT) with Group Relative Policy Optimization (GRPO). By optimizing a composite reward for answer fidelity and provenance correctness, GenProve significantly outperforms 14 strong LLMs in joint evaluation. Crucially, our analysis uncovers a reasoning gap where models excel at surface-level quotation but struggle significantly with inference-based provenance, suggesting that verifiable reasoning remains a frontier challenge distinct from surface-level citation.

1 Introduction

While LLMs demonstrate impressive fluency, their tendency to hallucinate remains a major barrier to widespread adoption (Fan et al., 2025). Users need to verify not just whether an answer sounds correct, but exactly where it comes from in the external evidence. To address this issue, current systems typically use Retrieval-Augmented Generation (RAG) or simply add citations to the text (Gao et al., 2023; Li et al., 2024). However, these standard approaches often function as *black boxes* because they provide a list of documents yet fail to specify *which* exact sentence supports a claim or *how* that evidence is used. This ambiguity leaves users guessing whether the model is directly quoting a fact, summarizing scattered details, or making a logical inference, which makes verification difficult and inefficient. For rigorous verification, knowing *how* a model uses evidence (e.g., inferring vs. quoting) is as important as knowing *which* document it used.

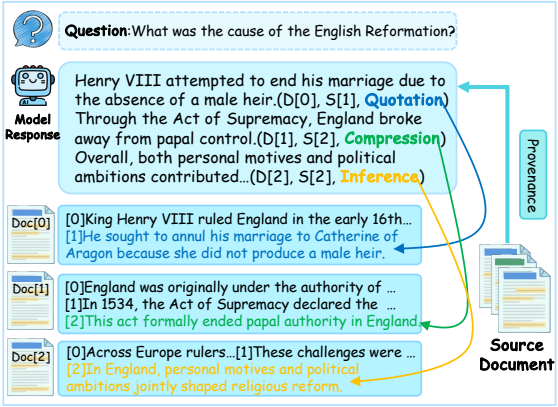

Figure 1: Overview of **Generation-time Fine-grained Provenance**. Given a query and source documents, the model simultaneously produces the answer and structured triples (DocID, SentID, Relation) to explain how the evidence supports each generated sentence.

We advocate for a more transparent approach, which we refer to as **Generation-time Fine-grained Provenance**. Unlike simple citation generation, this task requires the model to function as a transparent reasoner. For every generated sentence, the model must simultaneously identify the specific supporting source sentence and explicitly classify the evidence usage relation as *Quotation*, *Compression*, or *Inference* (Figure 1).

A major obstacle to this goal is the lack of training data. Most existing benchmarks are designed for analysis after generation or lack structured supervision on evidence types (Gao et al., 2023; Zhu et al., 2025). To bridge this gap, we construct **ReFInE** (**Re**lation-aware **F**ine-grained **In**terpretability & **E**vidence). Unlike previous

heuristic-based datasets, ReFInE is built via a rigorous human-in-the-loop pipeline where LLM-assisted proposals undergo multi-stage expert verification. This ensures the dataset accurately captures complex evidence usage patterns, serving as a reliable testbed for transparent generation.

Building on ReFInE, we propose **GenProve**, a training framework tailored for this objective. We observe that standard SFT is insufficient; models often struggle to balance the fluidity of the answer with the strict structural constraints of provenance triples. GenProve overcomes this by integrating GRPO (Guo et al., 2025) with a novel multi-dimensional reward modeling strategy. Specifically, we design a composite reward that goes beyond simple text quality. While strictly adhering to the structural constraints learned during SFT, our objective explicitly optimizes for *content fidelity* and *provenance correctness*, penalizing hallucinations where the cited evidence does not semantically support the generated claim. This holistic alignment forces the model to treat citation not as a stylistic decoration, but as an intrinsic reasoning constraint.

We evaluate GenProve against 14 strong LLMs. Results show that GenProve establishes a new state-of-the-art, significantly outperforming competitors in both answer quality and provenance accuracy. Crucially, our diagnostic analysis exposes a reasoning gap where models easily master exact *Quotation*, yet they struggle significantly with *Inference*. This suggests that the reliability of logical deductions remains a key challenge for future research.

Our contributions are summarized as follows:

- We define Generation-time Fine-grained Provenance, shifting from coarse document-level citations to sentence-level attribution with explicit relation typing.
- We release **ReFInE**, the first expert-annotated QA dataset that provides dense, typed provenance supervision for multi-document generation, enabling rigorous training and evaluation of model interpretability.
- We propose **GenProve**, integrating SFT with GRPO alignment to master structured provenance generation. Experiments demonstrate that GenProve establishes a new state-of-the-art, while our analysis reveals the difficulty of verifying inference-based claims compared to direct quotation.

2 Related Work

Citation-Aware Text Generation. Incorporating citations into generated text is a critical step towards verifiable AI. Early approaches focus on stylistic imitation of academic writing (Xing et al., 2020) or utilize post-hoc retrieval to verify generated claims after the fact (Li et al., 2024; Hsu et al., 2024). With the advent of LLMs, the focus has shifted to RAG. Benchmarks like ALCE (Gao et al., 2023) have standardized the evaluation of citation quality, emphasizing document recall and precision. However, these methods typically operate at a coarse granularity, retrieving entire documents or passages without pinpointing the specific evidence used. Recent training-based methods (Aly et al., 2024; Slobodkin et al., 2024) attempt to improve robustness by fine-tuning models to cite sources. However, these approaches are limited by treating citations as untyped pointers (e.g., simply linking to [1]). Our study advances this paradigm by enforcing typed relations, requiring the model to demonstrate an explicit understanding of the semantic relationship between the claim and the evidence, such as whether it is quoting or inferring.

Fine-Grained Provenance & Verification. To improve interpretability, granularity in attribution has evolved from document-level to sentence-level. Previous works like GERE (Chen et al., 2022) and SCIFI (Cao and Wang, 2024) explore generating sentence identifiers, while Kambhamettu et al. (2024) introduce phrase-level links. Most relevant to our work is TROVE (Zhu et al., 2025), which introduces a comprehensive taxonomy for provenance relations. We **adopt their core taxonomy** (*Quotation*, *Compression*, *Inference*) to ensure rigorous classification. However, we explicitly exclude their "Other" category. We omit this label for two reasons: first, it represents a negligible long-tail of the distribution; second, it serves as an ambiguous catch-all bucket. Such undefined signals lack clear semantic boundaries, making them unsuitable optimization targets for precise model alignment. Furthermore, while TROVE focuses on post-hoc analysis of static text, GenProve targets **generation-time** provenance. We integrate these fine-grained types directly into the training process, shifting the paradigm from checking after generation to generating with inherent verification.

Reasoning with Evidence. Research studies investigate how LLMs reason with retrieved context. Studies like FRONT (Huang et al., 2024)

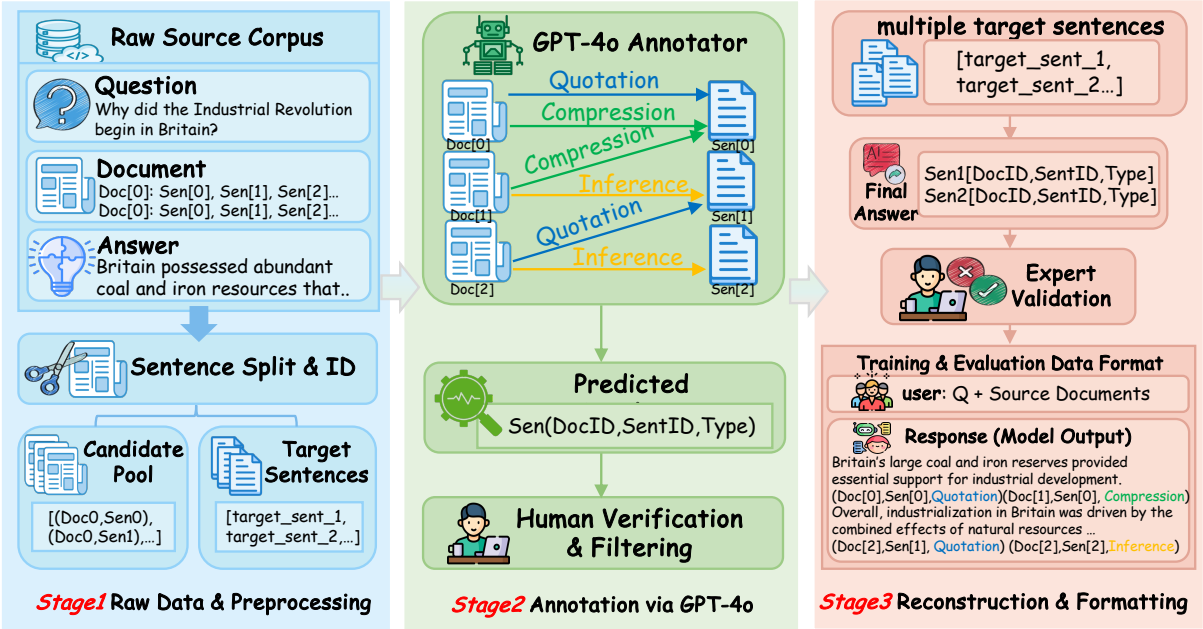

Figure 2: The construction pipeline of **ReFInE**. The process ensures high-quality provenance supervision through three stages: (1) preprocessing, (2) LLM-assisted annotation with filtering, and (3) reconstruction with rigorous **human-in-the-loop expert validation** to verify evidence sufficiency and relation correctness.

and SciRGC (Li and Chen, 2025) have begun to model the latent reasoning process behind citations, inspired by chain-of-thought prompting. GenProve pushes this direction further by explicitly supervising the *Inference* relation. Unlike previous works that often conflate simple retrieval with complex reasoning, our framework distinguishes between surface-level copying and deep synthesis. By optimizing for specific relation types, we evaluate and improve the model's ability to abstract and deduce information rather than merely retrieving and copying verbatim segments.

3 Dataset

3.1 Overview

We study GenProve, a generation-time fine-grained provenance task where a system produces an answer together with sentence-level evidence links and typed provenance relations. A central obstacle to learning GenProve is the lack of training data that aligns each answer sentence to specific source sentences while also distinguishing how the evidence is used, such as quotation, compression, or inference. To address this gap, we construct ReFInE, a supervised dataset that provides multi-document inputs and reference answers annotated with structured provenance tags.

Each ReFInE instance pairs a user question with multiple source documents and a reference answer, where every sentence carries a provenance annotation linking it to source sentences via Quotation, Compression, or Inference. We build the dataset through a three-stage pipeline comprising sentence-level preprocessing, GPT-4o–based provenance annotation with human screening, and expert-validated reconstruction into a unified message format. Figure 2 summarizes the construction process.

3.2 Dataset construction

Raw data and sentence-level preprocessing We build ReFInE on top of a public long-form QA corpus with retrieved multi-document evidence (Yehudai et al., 2024), where each example contains a user question, a set of source documents, and a long-form reference answer (Figure 2, Stage 1). For each instance, we treat the user query as the question Q, segment the long answer into sentences to obtain a sequence of target sentences $\{t_1, t_2, \dots\}$, and segment all source documents into sentences. Each source sentence in D is assigned a unique pair $(\text{Doc_ID}, \text{Sent_ID})$, which later serves as the indexing scheme for the provenance triples in Eq. 11. This stage produces a candidate pool of sentence-level evidence drawn from the source documents, together with a set of target sentences that require provenance labels.

GPT-4o-based provenance annotation and preliminary filtering. Given a question Q, a document set D, and a target sentence t_j, we prompt GPT-4o to predict provenance triples. These triples follow the (DocID, SentID, Relation) format, cov-

ering Quotation, Compression, and Inference (Figure 2, Stage 2). This process annotates each sentence independently to form a candidate pool. Next, three annotators screen the outputs for quality. They verify instruction compliance, fluency, and ethical safety. Crucially, they enforce strict [PROVE] formatting, checking for tag completeness, evidence merging, and index consistency. Violating samples are removed or corrected. Appendix B.1 details this protocol.

Reconstruction, expert validation, and final formatting. We reconstruct instance-level examples by aggregating sentence-level annotations (Figure 2, Stage 3). Target sentences are sorted by their original order and concatenated, with each sentence receiving a [PROVE] tag that enumerates its evidence and relations (Eq. 12). Subsequently, three experts conduct a second-round validation. Under a dual-check protocol, they rigorously verify evidence sufficiency and relation correctness (Quotation, Compression, Inference); instances failing either check are revised or discarded. Finally, valid samples are formatted for training: the user message comprises the question Q and documents D, while the assistant message contains the long answer A with embedded [PROVE] tags. Appendix B.2 details this protocol.

3.3 Dataset Analysis

Split composition and relation distribution. ReFInE consists of three subsets that support SFT, RL-based training (GRPO), and held-out evaluation (EVAL), containing 12,540, 5,256, and 4,838 instances, respectively. Figure 3 summarizes the split composition and the relation-type mixture within each subset. The split proportions are 55.4% (SFT), 23.22% (GRPO), and 21.38% (EVAL). The outer ring further shows that Quotation dominates across all splits, whereas GRPO allocates a larger share to Inference and Compression than EVAL, making it more suitable for optimizing reasoning and abstraction under sentence-level evidence constraints.

Provenance density and corpus-level statistics. In ReFInE, each answer contains 3.96 provenance tags on average, with each tag aggregating 1.98 provenance triples. Full corpus-level statistics are reported in Appendix B.5.

3.4 Comparison with Existing Benchmarks

Table 4 compares ReFInE with representative citation-aware and provenance benchmarks along axes central to fine-grained accountability, including relation expressivity, provenance granularity, task form, and whether provenance is produced alongside the answer using structured tags.

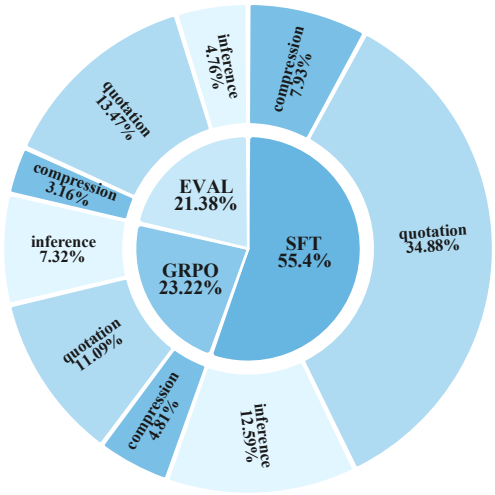

Figure 3: Relation-type distribution in ReFInE.

Fine-grained, typed sentence-level provenance. Many prior benchmarks emphasize attribution but collapse provenance into a single untyped support signal or operate at a coarser granularity, which limits sentence-level inspection and weakens the interpretability of how evidence supports each claim. In contrast, ReFInE annotates each answer sentence with sentence-level evidence links and explicit relation types, enabling relation-aware auditing beyond identifying the source alone.

Generation-time supervision. Several settings perform provenance analysis post hoc or append/verify citations in multi-step pipelines. ReFInE instead requires the answer and its provenance to be produced simultaneously, providing direct supervision and evaluation for generation-time provenance-aware decoding. Full benchmark comparison details are provided in Appendix B.6.

4 Method

We develop GenProve, a two-step training framework that enables a model to generate answers with fine-grained provenance. Given a question Q and a document collection $D = \{d_1, \ldots, d_m\}$, the model produces an answer $A = (t_1, \ldots, t_n)$. Each answer sentence t_j is accompanied by a set of provenance triples that identify supporting source sentences in D and their relation types. Figure 4 shows the overall procedure.

4.1 Supervised Fine-Tuning

We treat SFT as a foundational warm-up stage primarily designed to enforce structure adher-

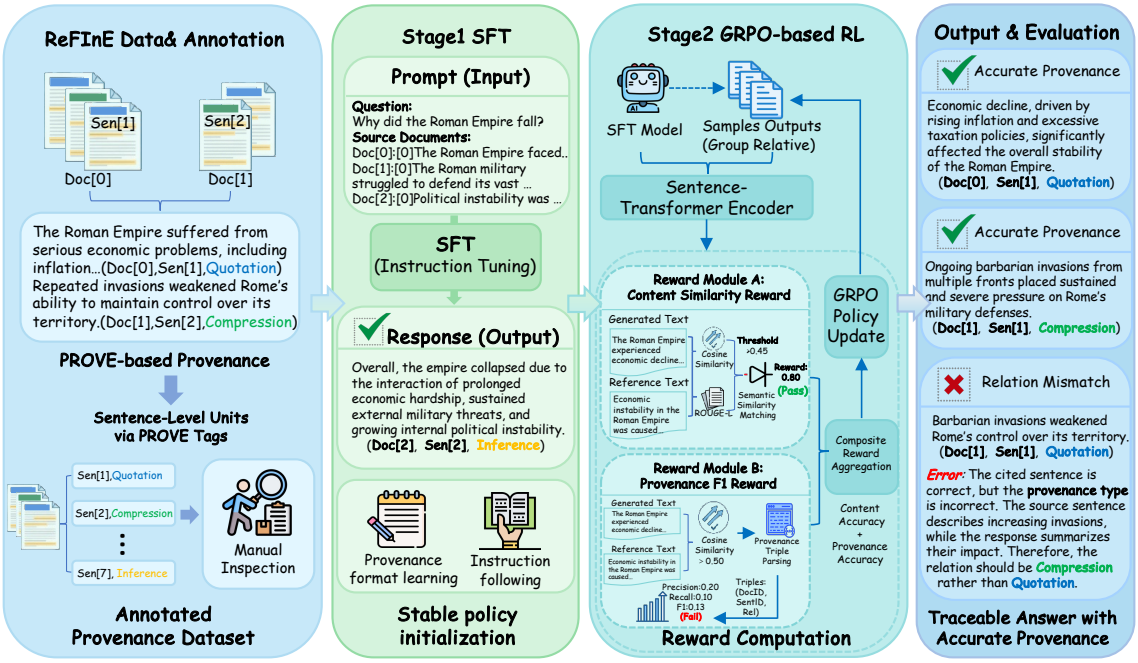

Figure 4: The **GenProve** framework. The model first undergoes SFT for instruction following and format learning. It is then aligned using GRPO with a **composite reward mechanism** that jointly optimizes for answer fidelity (content similarity reward) and fine-grained provenance accuracy (F1 Reward).

ence. It trains the base model to follow instructions and to emit well-formed provenance annotations together with fluent answers. Let $\mathcal{D}_{\text{SFT}} = \{(Q_i, D_i, A_i^{\text{ref}})\}_{i=1}^{N}$ denote the training set constructed from ReFInE, where A_i^{ref} is the reference answer annotated with sentence-level provenance. We fine-tune the model by maximizing the conditional likelihood of the reference output:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{i=1}^{N} \log p_\theta\left(A_i^{\text{ref}} \mid Q_i, D_i\right). \tag{1}$$

This step provides a stable policy initialization that reliably produces syntactically valid provenance tags and on-topic content. Crucially, this structural foundation enables the subsequent RL stage to focus on refining the model's provenance accuracy rather than struggling with basic formatting errors.

4.2 GRPO-based Reinforcement Learning

Reinforcement learning improves provenance accuracy and reduces unsupported statements by optimizing a reward that evaluates both content and provenance. Starting from the SFT policy π_θ, we sample a group of candidate answers for each input (Q, D) and update the policy using GRPO. The objective maximizes the expected reward:

$$\mathcal{J}(\theta) = \mathbb{E}_{(Q,D)\sim\mathcal{D}_{\text{GRPO}}}\left[\mathbb{E}_{A\sim\pi_\theta(\cdot|Q,D)}\left[R(A, A^{\text{ref}})\right]\right]. \tag{2}$$

The reward $R(A, A^{\text{ref}})$ aggregates two components: a sentence-matching content reward and a reference-guided provenance F1 reward (Figure 4).

Reward Design. GenProve computes rewards at sentence granularity by parsing provenance tags and splitting both the generated answer and the reference into sentence units. Let $A = (t_1, \ldots, t_n)$ and $A^{\text{ref}} = (t_1^{\text{ref}}, \ldots, t_M^{\text{ref}})$. We represent both answers as sentence–provenance pairs:

$$\begin{cases} A = \{(t_j, P_j)\}_{j=1}^{n}, \\ A^{\text{ref}} = \{(t_k^{\text{ref}}, P_k^{\text{ref}})\}_{k=1}^{M}. \end{cases} \tag{3}$$

Each P_j (or P_k^{ref}) is a set of triples of the form $(\text{doc_id}, \text{sent_id}, r)$ with $r \in \{\text{Quotation}, \text{Compression}, \text{Inference}\}$.

Reward A: Sentence-matching content similarity. This reward encourages semantic alignment with the reference while preserving sentence-level structure. For each generated sentence t_j, we find the best-matching reference sentence by cosine similarity between sentence embeddings produced by a Sentence-Transformer encoder:

$$k(j) = \arg\max_{k\in\{1,\ldots,M\}} \cos\left(\phi(t_j), \phi(t_k^{\text{ref}})\right). \tag{4}$$

Here $\phi(\cdot)$ denotes the encoder. If the best cosine score is below a threshold τ_c, the reward for this sentence is zero; otherwise we compute ROUGE-L

between the matched pair:

$$r_{\text{sim}}(t_j) = \mathbb{I}\Big[\cos\big(\phi(t_j), \phi(t^{\text{ref}}_{k(j)})\big) \geq \tau_c\Big] \cdot \text{ROUGE-L}\big(t_j, t^{\text{ref}}_{k(j)}\big). \quad (5)$$

The content reward is the mean across sentences:

$$R_{\text{sim}}(A, A^{\text{ref}}) = \frac{1}{n}\sum_{j=1}^{n} r_{\text{sim}}(t_j). \quad (6)$$

Reward B: Reference-guided provenance F1. This reward encourages generating correct provenance triples and relation types. To reduce missed provenance, we align sentences inversely. For each reference sentence t^{ref}_k, we retrieve the closest generated sentence by cosine similarity:

$$j(k) = \arg\max_{j\in\{1,\ldots,n\}} \cos\big(\phi(t^{\text{ref}}_k), \phi(t_j)\big). \quad (7)$$

We gate mismatched pairs using a similarity threshold τ_p. Given an aligned pair, we compare their provenance sets. Let $I_k = P_{j(k)} \cap P^{\text{ref}}_k$ denote the set of correctly reproduced provenance triples. We compute sentence-level precision and recall as $\text{Prec}_k = |I_k|/|P_{j(k)}|$ and $\text{Rec}_k = |I_k|/|P^{\text{ref}}_k|$, and define the provenance score by

$$F1_k = \frac{2\,\text{Prec}_k\,\text{Rec}_k}{\text{Prec}_k + \text{Rec}_k + \epsilon}. \quad (8)$$

The provenance reward averages sentence-level scores over all reference sentences, while gating out mismatched pairs:

$$R_{\text{prov}}(A, A^{\text{ref}}) = \frac{1}{M}\sum_{k=1}^{M} \mathbb{I}\Big[\cos\big(\phi(t^{\text{ref}}_k), \phi(t_{j(k)})\big) \geq \tau_p\Big] \cdot F1_k. \quad (9)$$

Composite reward. We combine the two components into a single scalar reward:

$$R(A, A^{\text{ref}}) = \alpha\, R_{\text{sim}}(A, A^{\text{ref}}) + \beta\, R_{\text{prov}}(A, A^{\text{ref}}), \quad (10)$$

where α and β balance content fidelity and provenance correctness. This design penalizes common failure modes shown in Figure 4, including incorrect relation typing and unsupported or out-of-document provenance.

5 Experiments

5.1 Experimental Setup

Models. We evaluate 14 LLMs, covering both open- and closed-source systems. The open-source models include Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Gemma-3-12B-it (Team et al., 2025a), Yi-1.5-9B-Chat (Liu et al.), Qwen3-8B (Yang et al., 2025), InternLM2.5-7B-Chat (Cai et al., 2024), Hunyuan-7B-Instruct (Zheng et al., 2025), Vicuna-7B-v1.5 (Zheng et al., 2023), Baichuan2-7B-Chat (Yang et al., 2023), Qwen3-14B (Yang et al., 2025), GLM-4-9B (GLM et al., 2024), and GLM-4.5 (GLM et al., 2024). The closed-source models include Gemini 2.5 Pro (Comanici et al., 2025), GPT-5 (Achiam et al., 2024), and Kimi (Team et al., 2025b). All models use a unified input format of questions and source documents. The full inference prompt is given in Appendix C.1.

Training configuration. GenProve is trained in two steps. For supervised fine-tuning, we start from Qwen3-8B (Yang et al., 2025) and perform full-parameter optimization with AdamW, using a learning rate of 2×10^{-5}, a maximum sequence length of 2048, and gradient accumulation to achieve an effective batch size of 16. For GRPO alignment, we initialize from the SFT model and continue optimization under the same learning rate and sequence length settings, with temperature set to 1, $\beta = 0.02$, and 4 iterations per update. For reward computation, the sentence-matching and provenance-alignment thresholds are set to $\tau_c = 0.45$ and $\tau_p = 0.50$, respectively.

Evaluation Metrics. We evaluate models along three axes: answer quality, provenance accuracy, and format validity. Answer quality is measured using ROUGE-L (Lin, 2004), BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), BERTScore (Zhang et al., 2020), and MoverScore (Zhao et al., 2019), computed on answers excluding provenance tags. Provenance accuracy is assessed by sentence-level precision, recall, and F1 via exact matching over document id, sentence id, and relation type, while format validity reports the percentage of outputs that strictly follow the required provenance schema. Additionally, we conduct subjective evaluations with LLM and human judges: the former provides relation-specific scores, while the latter assesses answer quality and provenance correctness (prompts and guidelines in Appendices C.2–C.3).

5.2 Main Results

Table 1 presents the main results on ReFInE. GenProve achieves the best overall performance and ranks first on all evaluation axes, including answer quality, provenance accuracy, and the LLM-as-judge score. The gains are consistent across automatic metrics and subjective judging, demonstrating that generation-time fine-grained provenance

Model	ROUGE-L↑	BLEU↑	METEOR↑	MoverScore↑	Prec.↑	Rec.↑	F1↑	Format (%)↑	LLM-as-judge (1–5)↑
Baichuan2-7B (Yang et al., 2023)	34.68	19.41	48.08	32.03	3.22	3.41	3.04	26.60	0.78
Vicuna-7b-v1.5 (Zheng et al., 2023)	38.83	24.37	50.08	34.73	9.01	5.85	6.70	92.40	1.10
InternLM2.5-7B (Cai et al., 2024)	47.79	30.60	53.47	43.60	12.64	13.35	11.94	80.24	1.67
Hunyuan-7B (Zheng et al., 2025)	39.91	24.70	43.74	32.99	22.78	22.86	21.43	90.88	1.72
Yi-1.5-9B (Liu et al.)	48.26	30.11	47.77	43.35	20.71	22.91	20.11	96.81	1.80
Llama-3.1-8B (Grattafiori et al., 2024)	47.71	26.36	42.04	41.14	21.78	20.30	20.05	99.85	2.00
GLM-4-9B (GLM et al., 2024)	50.33	32.62	50.00	44.93	34.57	34.12	32.79	**100.0**	2.16
Qwen3-8B (Yang et al., 2025)	51.80	35.56	55.38	45.90	37.78	30.56	32.53	**100.0**	2.25
Gemma-3-12B (Team et al., 2025a)	48.97	32.13	50.80	43.79	41.06	31.11	34.03	**100.0**	2.47
Qwen3-14B (Yang et al., 2025)	52.85	35.70	55.34	47.06	45.80	40.33	41.16	99.70	2.59
GLM-4.5-355B (GLM et al., 2024)	49.81	35.05	57.69	44.92	48.84	44.03	44.55	98.63	2.63
Kimi (Team et al., 2025b)	49.33	31.24	51.56	43.84	31.55	32.09	29.70	99.09	2.20
GPT-5 (Achiam et al., 2024)	41.79	20.01	38.83	36.38	21.37	16.73	17.88	99.85	2.23
Gemini 2.5 Pro (Comanici et al., 2025)	48.75	31.77	53.09	44.79	46.68	42.86	42.92	**100.0**	2.57
GenProve (Ours)	**57.25**	**42.22**	**59.39**	**51.04**	**54.96**	**51.26**	**51.21**	99.85	**3.14**

Table 1: Main results on ReFInE. Our proposed GenProve consistently outperforms strong open-source and closed-source LLMs across answer quality, provenance accuracy, and LLM-based evaluation.

training improves both the usefulness of answers and the reliability of sentence-level provenance.

Across model groups, open-source systems exhibit substantial variance. Earlier chat-style or lightly instruction-tuned models, such as Baichuan2-7B and Vicuna-7B-v1.5, often fail to follow the provenance schema, leading to low format validity and weak provenance accuracy. In contrast, more recent open-source models, including Qwen3 and GLM-4, generate valid outputs more consistently and achieve markedly higher provenance F1 and LLM-judge scores. Among non-GenProve systems, GLM-4.5 is the strongest baseline, ranking second in both provenance quality and LLM-judge score. Closed-source models are competitive: Gemini 2.5 Pro is the strongest closed-source baseline, but still trails GenProve on the overall judge score.

From the metric perspective, answer-quality metrics show that GenProve generates more faithful and fluent responses after provenance tags are removed. It exceeds the strongest baseline on ROUGE-L, BLEU, METEOR, and MoverScore, indicating improvements in both surface overlap and semantic similarity. Provenance metrics show the largest margin: GenProve achieves a substantially higher provenance F1 than the strongest baseline, suggesting more accurate sentence-level evidence localization and relation typing. Correct format highlights that formatting is necessary but not sufficient: several strong baselines already achieve near-perfect parseability, whereas weaker baselines fail frequently; GenProve maintains similarly high compliance. Finally, the LLM-as-judge score summarizes end-to-end quality under joint requirements of correctness, fluency, and traceability, where GenProve attains the highest overall score.

Model	BLEU	BERTScore	F1	Format	Judge
GenProve	42.22	61.98	51.21	99.85	3.14
w/o Prov Reward	11.94	46.96	24.67	96.66	2.20
w/o Sim Reward	24.60	44.50	60.32	95.74	2.71
w/o GRPO	41.82	60.70	50.48	99.70	2.62

Table 2: Ablation study on ReFInE. The results validate the necessity of GRPO alignment and the complementary roles of content similarity and provenance rewards.

5.3 Ablation Study

Table 2 shows that GRPO alignment significantly boosts end-to-end quality, raising judge scores substantially over the SFT baseline. Reward ablations further confirm the two components are complementary. Removing the provenance reward drops provenance F1 and judge scores, implying similarity alone cannot enforce precise provenance. Conversely, removing the similarity reward improves F1 yet harms answer quality, showing provenance optimization alone is insufficient. Combining both maximizes the judge score, effectively balancing fluent answers with correct, typed provenance.

5.4 Diagnostic Analysis

Performance by Relation Type. Figure 5 reports F1 by provenance relation type, reflecting the reliability of sentence-level provenance beyond verbatim reuse. The heatmap reveals a clear difficulty ordering: Quotation is easiest, while Compression and Inference are substantially harder, indicating challenges in evidence abstraction and integration. GenProve achieves the strongest performance across all three relations and the highest average F1, with its largest gains on Compression and Inference, reflecting improved evidence localization and relation typing in harder cases.

GRPO Training Dynamics. Figure 6 visualizes reward trajectories during GRPO alignment. Both component rewards increase and stabilize, indicating the policy improves content alignment and

Figure 5: Performance breakdown by relation type (F1 score). The heatmap reveals a reasoning gap: while most models handle verbatim *Quotation* well, they struggle significantly with *Inference*. **GenProve** consistently outperforms baselines, showing the largest gains in complex provenance tasks (Compression and Inference).

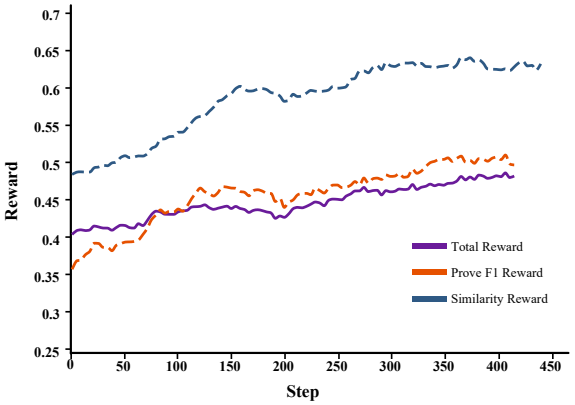

Figure 6: Learning dynamics during GRPO. Consistent upward trends in content similarity and provenance F1 rewards indicate GenProve improves provenance reliability without compromising answer faithfulness, achieving coordinated optimization of dual objectives.

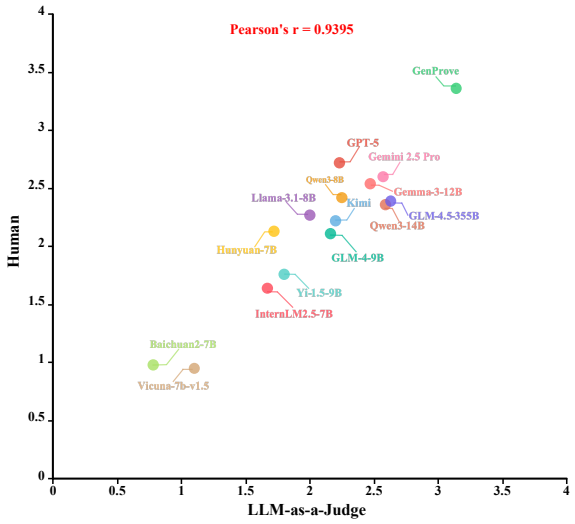

Figure 7: Correlation between LLM-as-a-Judge scores and human ratings. The high Pearson correlation ($r = 0.9395$) validates our automatic metric. Notably, **GenProve** occupies the top-right corner, demonstrating significantly superior performance over all baselines under both automated and human evaluations.

provenance correctness jointly rather than oscillating between objectives. The total reward follows this upward trend, mirroring the complementary roles found in ablations: similarity optimization strengthens faithfulness, whereas PROVE-F1 strengthens typed provenance, and their combination supports superior overall quality.

5.5 Consistency with Human Evaluation

To assess the reliability of LLM-as-a-Judge as our primary evaluation signal, we measure its consistency with human evaluation at the model level. Figure 7 shows a strong positive correlation between LLM-as-a-Judge scores and human ratings, with a Pearson correlation coefficient of $r = 0.9395$. This result indicates that the automatic judge closely aligns with human preferences under the same provenance-aware evaluation criteria, supporting its use for large-scale comparison in the main experiments. Detailed human evaluation results and scoring analyses are provided in Appendix C.4 and C.6.

6 Conclusion

We introduce a paradigm shift from coarse citations to generation-time fine-grained provenance. By constructing the **ReFInE** dataset and developing the **GenProve** framework, we demonstrate that LLMs can be trained to transparently document their evidence usage via structured triples. Experiments confirm that GenProve balances generation quality with strict provenance constraints, establishing a new state-of-the-art across 14 strong LLMs. Despite these advances, the performance gap between simple *Quotation* and complex *Inference* suggests that verifiable reasoning remains a frontier challenge. We position ReFInE as a stepping stone towards self-auditing LLMs, models that not only generate knowledge but explicitly reason about the provenance of their own assertions.

Limitations

While GenProve establishes a new standard for fine-grained provenance, we identify three limitations to address in future work. (1) **Inference latency**. Generating structured provenance triples inevitably increases the output token count compared to standard generation. Although essential for trustworthiness, this introduces a slight latency trade-off in real-time applications. (2) **Linguistic scope**. Our current ReFInE dataset and evaluation primarily focus on English. Extending the *Quotation-Compression-Inference* taxonomy to multilingual or cross-lingual settings remains an open avenue for research. (3) **Retrieval dependency**. Our framework focuses on the generation stage. Like all RAG systems, end-to-end performance is bounded by the quality of the retriever; if retrieved documents contain no relevant information, the model cannot generate valid provenance.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, and others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, and others. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

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)*, pages 2184–2189.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, and others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Mingyuan Fan, Chengyu Wang, Cen Chen, Yang Liu, and Jun Huang. 2025. On the trustworthiness landscape of state-of-the-art generative models: A survey and outlook. *International Journal of Computer Vision*, 133(7):1–32.

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.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, and others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, and others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645:633–638.

I-Hung Hsu, Zifeng Wang, Long Le, Lesly Miculicich Werlen, Nanyun Peng, Chen-Yu Lee, and Tomas Pfister. 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 others. 2024. Learning fine-grained grounded citations for attributed large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14095–14113.

Hita Kambhamettu, Jamie Flores, and Andrew Head. 2024. Traceable text: Deepening reading of ai-generated summaries with phrase-level provenance links. *Preprint*, arXiv:2409.13099.

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.

Xiangyu Li and Jingqiang Chen. 2025. Scirgc: Multi-granularity citation recommendation and citation sentence preference alignment. *Preprint*, arXiv:2505.20103.

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

Jiaheng Liu, Riza Batista-Navarro, Qian Liu, Niklas Muennighoff, Ge Zhang, Yizhi LI, Xinyi Wang, and Willie Neiswanger. Open science for foundation models. In *ICLR 2025 Workshop Proposals*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

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

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, and others. 2025a. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, and others. 2025b. Kimi k2: Open agentic intelligence.

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.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, and others. 2023. Baichuan 2: Open large-scale language models. *Preprint*, arXiv:2309.10305.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, and others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Asaf Yehudai, Boaz Carmeli, Yosi Mass, Ofir Arviv, Nathaniel Mills, Eyal Shnarch, and Leshem Choshen. 2024. Achieving human parity in content-grounded datasets generation. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578.

Lianmin Zheng, Wei-Lin Chiang, and others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 46595–46623.

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. 2025. Hunyuan-mt technical report. *Preprint*, arXiv:2509.05209.

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 (ACL)*, pages 11755–11771.

A Task Definition

We study a text provenance task in which a system answers a user question and, for each answer sentence, specifies which source sentences support it and how they relate. Formally, given a question Q and a collection of source documents $D = \{d_1, d_2, \ldots, d_m\}$, where each document d_i is a sequence of sentences $d_i = \{s_{i,1}, s_{i,2}, \ldots, s_{i,k_i}\}$, the system produces an answer $A = (t_1, t_2, \ldots, t_n)$ and a provenance set $P = (P_1, P_2, \ldots, P_n)$ aligned with the answer. For each answer sentence t_j, the provenance P_j is a set of triples:

$$P_j = \{(\text{doc_id}, \text{sent_id}, r)\}, \tag{11}$$

where doc_id indexes a document in D, sent_id indexes a sentence within that document, so that $(\text{doc_id}, \text{sent_id})$ uniquely identifies a source sentence $s_{i,\ell}$ in D, and r denotes the provenance relation type.

Each answer sentence t_j may be supported by multiple source sentences, possibly drawn from different documents, and different links may carry different relation types. We consider three provenance relation types between an answer sentence t_j and a source sentence $s_{i,\ell}$: **Quotation**, where t_j copies or closely paraphrases the wording of $s_{i,\ell}$; **Compression**, where t_j summarizes or paraphrases information that is distributed across one or more source sentences, such as s_{i,ℓ_1} and s_{i,ℓ_2}; and **Inference**, where t_j states a conclusion that is logically supported by one or more source sentences, such as $s_{i,\ell}$. For a single answer sentence, different supporting source sentences may be associated with different relation types, and all corresponding triples are collected in P_j as in Eq. 11.

In our setting, the answer and its provenance are presented in a structured textual format by interleaving each answer sentence t_j with a provenance annotation that enumerates the triples in P_j. Concretely, an answer sentence and its provenance may be rendered as:

$$Sentence.[PROVE : (d1, s6, ' Quotation'), (d2, s3, ' Inference')] \tag{12}$$

where Sentence. corresponds to an answer sentence t_j, each pair (d1, s6) or (d2, s3) is a concrete instance of $(\text{doc_id}, \text{sent_id})$ in Eq. 11, and the strings 'Quotation' and 'Inference' instantiate the relation type r. The task thus requires a system to generate an answer to Q and, at the same time, provide fine-grained, sentence-level provenance that identifies supporting source sentences and labels the type of relationship for every answer sentence.

B Additional Details of ReFInE Construction

B.1 Preliminary filtering criteria

After GPT-4o produces sentence-level answers and provenance candidates, all instances undergo a preliminary filtering stage conducted by three annotators. The goal is to remove clearly unusable or structurally invalid samples before expert validation. We adopt four main criteria.

Instruction compliance. The generated answer must directly address the user question. Annotators discard responses that ignore the query, explicitly refuse to answer (e.g., "I do not know"), or simply restate the question without providing new information.

Fluency and completeness. Annotators remove outputs that exhibit severe grammatical errors, broken sentence structure, heavy repetition, or truncated content due to length limits that render the answer semantically incomplete.

Ethical compliance. Any sample containing hate speech, discriminatory language, personal identifiable information, or harmful recommendations is removed to ensure that ReFInE does not propagate unsafe content.

Format validity. We enforce strict constraints on the structure of provenance annotations. Annotators check:

- **Tag completeness**: every factual sentence in the answer must be accompanied by a parsable [PROVE] tag; samples with missing or unparsable tags are discarded.
- **Merged multi-source references**: if a single sentence is supported by multiple evidence sentences, all evidence must appear within a single [PROVE: (...)] block. Instances that split evidence across multiple tags for the same sentence are considered format violations.
- **Index consistency**: all DocID and SentID values must follow the zero-based indexing scheme and stay within the valid ranges of

the input documents and sentences; out-of-bounds or misaligned indices lead to removal.

- **Tuple well-formedness**: each provenance tuple must contain exactly three fields ("doc_id", "sent_id", "relation") with correct types; malformed tuples are grounds for discarding the sample.

B.2 Expert validation protocol

Instances that pass the preliminary filter then enter an expert validation stage. Three annotators with experience in NLP and factuality assessment independently review the remaining samples. For each answer sentence and its [PROVE] tag, annotators perform a two-part check that jointly considers evidence sufficiency and relation correctness.

Evidence sufficiency. Annotators examine whether the cited (DocID, SentID) tuples provide adequate support for the generated sentence. A provenance set is accepted if the content of the sentence can be fully derived from the cited source sentences without introducing unsupported external facts or contradicting the documents. If the answer contradicts the sources, lacks supporting evidence, or relies on hallucinated information, the instance is marked invalid and removed.

Relation correctness. Annotators standardize the use of the three relation types in ReFInE and verify that each predicted label matches the underlying evidence–sentence relationship.

Quotation. A link is labeled as Quotation when the answer sentence partially or fully copies the wording of the source sentence, allowing only minor grammatical adjustments such as tense or word order changes. If the sentence substantially rephrases or loosely paraphrases the source while being labeled as quotation, the label is corrected.

Compression. A link is labeled as Compression when the answer sentence is a faithful condensation of long or multi-sentence content in the source. Annotators check that the compressed sentence preserves the key information while shortening or simplifying the original wording. If a sentence merely copies the source or omits crucial information while being tagged as compression, the label is revised. For example, the sentence "The dam releases water because of heavy rain" can be accepted as a compression of a longer source description that explains rising water levels and forced discharge due to continuous rain.

Inference. A link is labeled as Inference when the answer states a conclusion that is logically supported by one or more source sentences. Annotators verify that the conclusion follows from the cited evidence, possibly requiring multi-hop reasoning, cross-paragraph or cross-document integration, or reasonable commonsense inference. If the answer introduces content that is not implied by the sources or breaks the reasoning chain, the instance is marked invalid. Typical accepted cases include multi-hop reasoning that combines two or more source sentences, cross-document aggregation of facts, and commonsense conclusions that extend but do not contradict the given evidence.

B.3 Prompt for GPT-4o Annotation

To ensure consistent sentence-level provenance annotation across ReFInE, we provide GPT-4o with a structured prompt that defines the three relation types, specifies the expected formats for ground_truth_global, ground_truth_local, and Candidate Text, and illustrates the mapping between global and local sentence indices. The prompt also includes an example input object that clarifies how candidate sentences and their relationships should be encoded. This prompt governs Stage 2 of the dataset construction pipeline, and the full prompt specification is shown in Figure 8, which standardizes all annotations prior to human verification.

B.4 Examples of ReFInE Instances

To illustrate the final "message" format used in ReFInE, we provide examples covering all three provenance relation types. Each instance follows a uniform structure: the `user` message contains the question together with its associated source documents, and the `assistant` message provides the answer enriched with sentence-level provenance tags in the `[PROVE:(doc_id, sent_id, relation)]` format. The examples demonstrate how Quotation, Compression, and Inference relations appear in fully constructed data.

Figure 9 shows a Quotation instance, where each answer sentence closely matches its supporting source sentence. Figure 10 illustrates a Compression example in which the answer condenses information across multiple source sentences. Figure 11 presents an Inference case where the answer requires multi-sentence or cross-document reasoning grounded in the provided evidence.

Dataset Construction Prompt

You need to complete three fields in the dataset: `ground_truth_global`, `ground_truth_local`, and `Candidate Text`. The specific tasks are:
1. Analyze the relationship between the target sentence and each candidate sentence Relationship types: • **Quotation**: The target sentence partially or fully replicates content from a candidate sentence, including exact quotes, slight edits, or incorporation of phrases. • **Compression**: The target sentence condenses information from one or more candidate sentences. • **Inference**: The target sentence is based on information implied rather than explicitly stated.
2. Populate the fields:
• **ground_truth_global**: Key format: "DocID-SentID" → Relationship Only include candidate sentences relevant to the target sentence.
• **ground_truth_local**: Key format: local candidate index (based on `global2local_id`) → Relationship Only include relevant candidate sentences.
• **Candidate Text**: A list of dictionaries: `{"Text Address": "NULL", "Doc_ID": "X", "Original Sentence": [{"Critical Sentence": "...", "Relationship": "...", "Sent_ID": "..."}]}` Include only sentences that support the target sentence.
—
Example Input Object:

```
{"id": -5742330000000000000,
 "target_id": 0,
 "target_sent": "\"Bunk'd\" was renewed for a third season by Disney Channel
 on August 31, 2017, and it premiered on June 18, 2018.",
 "Candidate Relationship Sets": ["Quotation", "Compression", "Inference"],
 "prompt_local": "[Content]=\"\"\"
Target Sentence: \"Bunk'd\" was renewed for a third season by Disney Channel on August 31, 2017, and it premiered on June 18, 2018.
Candidate Sentence [1]: \"The series was renewed for a third season by Disney Channel on August 31, 2017.
Candidate Sentence [2]: On June 1, 2018, it was announced that Peyton List, ...
Candidate Sentence [3]: The third season premiered on Disney Channel on June 18, 2018.
Candidate Sentence [4]: In March 2018, actress Skai Jackson stated ...
Candidate Sentence [5]: ...
\"\"\"
",
 "prompt_global": "[Content]=\"\"\"
Target Sentence: \"Bunk'd\" was renewed for a third season ...
Candidate Sentence [0-0]: \"The series was renewed for a third season ...
Candidate Sentence [0-1]: ...
Candidate Sentence [0-2]: The third season premiered on Disney Channel ...
Candidate Sentence [0-3]: ...
Candidate Sentence [1-0]: ...
Candidate Sentence [2-6]: The second season premiered on August 23, 2016.
\"\"\"
",
 "global2local_id": {"0-0": "1", "0-1": "2", "0-2": "3", ...},
 "ground_truth_global": {"0-0": "Quotation", "0-2": "Quotation"},
 "ground_truth_local": {"1": "Quotation", "3": "Quotation"},
 "Candidate Text": [
   {"Text Address": "NULL",
    "Doc_ID": "0",
    "Original Sentence": [
       {"Critical Sentence": "\"The series was renewed ... 2017.\"",
        "Relationship": "Quotation",
        "Sent_ID": "0"},
       {"Critical Sentence": "The third season premiered ... 2018.",
        "Relationship": "Quotation",
        "Sent_ID": "2"}
    ]}
 ]}
```

Figure 8: The prompt used in Stage 2 of dataset construction. It defines the schema for generating candidate provenance triples and mapping global-local indices for GPT-4o annotation.

B.5 Additional Dataset Statistics

Table 3 reports corpus-level statistics of ReFInE, including provenance density, provenance-triple aggregation, answer length (excluding provenance tags), and source-document granularity.

B.6 Benchmark Comparison Details

Table 4 compares ReFInE with representative benchmarks. We use ✓/✗ to denote Yes/No. "Avg. citations" denotes the average number of provenance triples in the reference answers. "Structured tags" indicates metadata-rich provenance tags (DocID, SentID, Relation) rather than plain indices.

C Additional Experimental Details

C.1 Inference Prompt for Provenance-Aware Generation

Figure 12 presents the unified inference prompt used across all models in our experiments. The prompt enforces evidence-based generation conditioned on the provided source documents and requires each factual sentence to be accompanied by a structured provenance tag. It explicitly defines the three provenance relation types—Quotation, Compression, and Inference—and specifies exclusion rules to avoid annotating non-factual content.

C.2 LLM Judge for Traceable QA

To achieve a more precise and interpretable evaluation, we decouple LLM-based judging into two independent components: **text generation qual-**

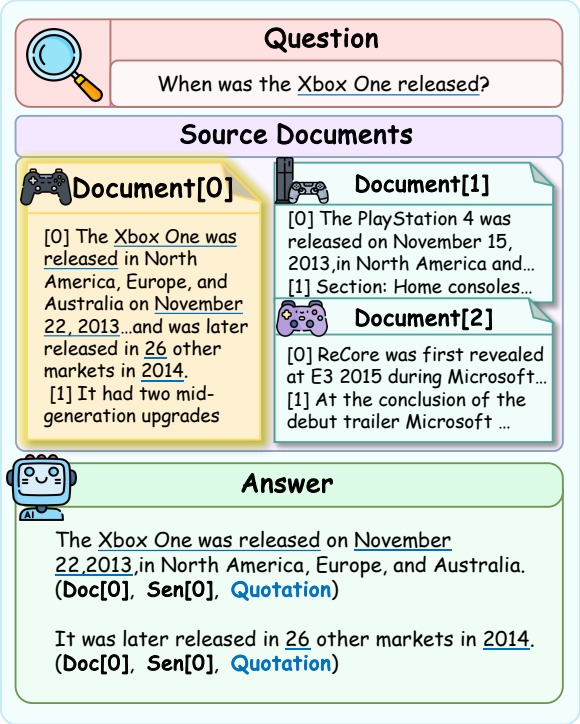

Figure 9: Example ReFInE instance illustrating the **Quotation** relation type.

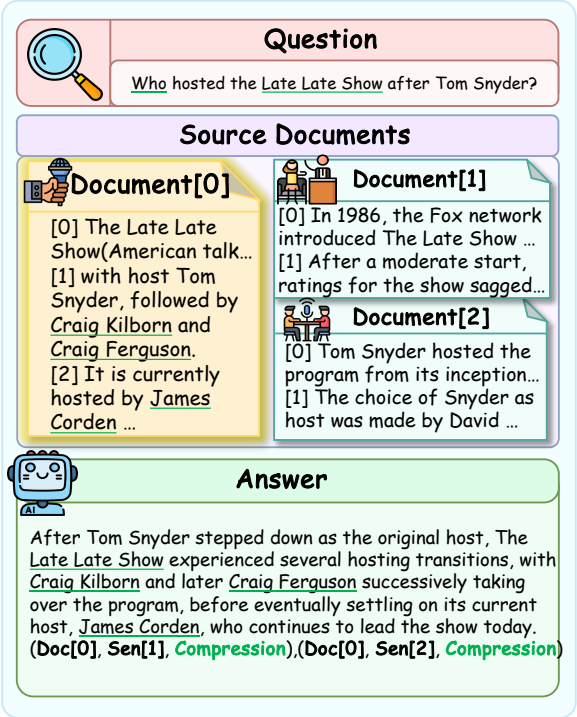

Figure 10: Example ReFInE instance illustrating the **Compression** relation type.

ity and **traceability quality**. Instead of relying on a single judge that conflates linguistic quality with citation correctness, we adopt two specialized LLM judges, each focusing on a distinct evaluation objective.

The first judge evaluates the natural language quality of the generated answer, while the second judge exclusively assesses the correctness and completeness of the provenance annotations. This separation enables finer-grained diagnosis of model errors, distinguishing between deficiencies in answer quality and failures in attribution or reasoning.

Text Generation Quality Judge Figure 13 shows the prompt used to evaluate the textual quality of model responses. The judge compares the generated answer against the question, source documents, and ground-truth labels, and assigns a score from 0 to 5 based on accuracy, fluency, and completeness. This judge does not consider provenance tags and focuses solely on the quality of the natural language answer.

Traceability Quality Judge Figure 14 illustrates the prompt used to evaluate provenance correctness. This judge focuses exclusively on the `[PROVE]` tags, verifying both the accuracy and completeness of attribution relationships with respect to the ground-truth labels and source documents.

Crucially, the judge explicitly penalizes missing relationship types required by the labels, enabling principled detection of recall errors in traceability generation.

C.3 Human Evaluation Guidelines

We conduct human evaluation to complement automatic and LLM-based metrics. Three expert annotators with prior experience in question answering and evidence annotation participate in the study. We randomly sample 200 evaluation instances from the test set and collect ratings for each model output. Each instance is independently evaluated by all annotators, and final scores are obtained by averaging across raters.

Answer quality (0–5). Raters score the natural-language answer while ignoring all provenance tags. Scores reflect accuracy with respect to the reference answer and the source documents, as well as fluency and completeness. A score of 5 indicates a fully correct and fluent answer with no substantive omissions or errors, while 0 indicates an unusable response (e.g., empty, refusal, or entirely incorrect).

Provenance quality (0–5). Raters assess whether provenance tags correctly and sufficiently support the answer. Relation types follow the same definitions as in our task (Quotation, Compression, Inference). We apply two key rules for relation-type scoring: (i) if a relation type appears in the

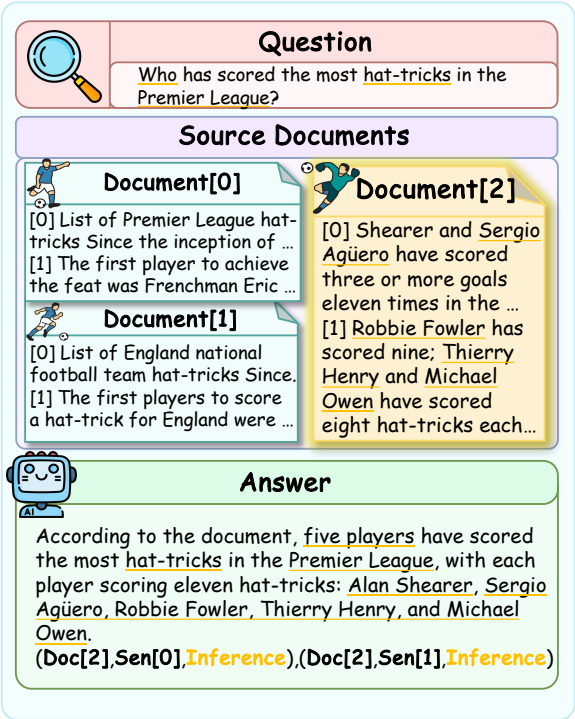

Figure 11: Example ReFInE instance illustrating the **Inference** relation type.

reference but is missing from the model output, the score for that type is 0; (ii) if a relation type appears in neither the reference nor the model output, the score for that type is null. When a relation type is used by the model, raters assign a 0–5 score based on correctness of the cited evidence and relation typing. Raters also provide an overall provenance score (0–5) that summarizes citation correctness and coverage. We do not penalize purely textual mentions of 1-based document numbering if the provenance tags correctly map to the underlying 0-based document identifiers.

C.4 LLM-as-a-Judge Breakdown

Table 5 reports the LLM-as-a-judge breakdown on ReFInE. The **Avg** column matches the LLM-as-judge score reported in the main results, while **Text Quality** and the provenance columns explain where that end-to-end score comes from. This decomposition directly aligns with our motivation: for trustworthy generation, high-level answer fluency is insufficient unless each sentence is supported by correctly localized evidence and an appropriate relation type.

Overall, the table shows that end-to-end differences are driven mainly by provenance rather than surface answer quality. For competitive systems, Text Quality scores concentrate in a relatively narrow range, whereas Overall Prov. and relation-specific provenance scores vary substantially. As a result, models with comparable Text Quality can still diverge sharply in Avg, which indicates that traceability correctness is the primary bottleneck captured by this benchmark.

Statistic	Value
PROVE-tag statistics	
Total [PROVE] tags	11,467
[PROVE] tags per answer (avg / median)	3.96 / 4
[PROVE] tags per answer (min / max)	1 / 14
Provenance-triple statistics	
Total provenance triples	22,138
Triples per answer (avg / median)	7.64 / 7
Triples per answer (min / max)	1 / 46
Triples per [PROVE] tag (avg / median)	1.98 / 2
Triples per [PROVE] tag (min / max)	1 / 18
Answer length (words, without [PROVE] tags)	
Answer length (avg / median)	94.38 / 96
Answer length (min / max)	10 / 303
Source document statistics	
Sentences per document (avg / median)	4.52 / 4
Sentences per document (min / max)	1 / 19
Words per document (avg / median)	88.61 / 97
Words per document (min / max)	25 / 103

Table 3: Corpus-level statistics of ReFInE, detailing provenance tag density, provenance aggregation, and source document granularity.

Across model groups, open-source systems exhibit the largest dispersion. Earlier baselines such as Baichuan2-7B (Yang et al., 2023) and Vicuna-7B-v1.5 (Zheng et al., 2023) show low provenance scores, consistent with the weaker schema-following and attribution behavior observed in the main results. Stronger open-source instruction-tuned models, including Qwen3 (Yang et al., 2025) and GLM-4 (GLM et al., 2024), achieve much higher provenance scores and therefore higher Avg. Among non-GenProve systems, GLM-4.5 (GLM et al., 2024) provides the strongest overall baseline by Avg and overall provenance. Closed-source models are competitive: Gemini 2.5 Pro (Comanici et al., 2025) yields the strongest closed-source Avg and provenance, while GPT-5 (Achiam et al., 2024) and Kimi (Team et al., 2025b) show weaker provenance breakdowns despite strong text scores.

From the metric perspective, the relation-specific columns reveal a stable difficulty pattern. Quotation receives the highest scores for most models, which indicates that direct reuse attribution is relatively easy to judge and to satisfy. In contrast, Compression and Inference scores remain low for many systems, even when their Quotation scores are strong, and these two relations largely deter-

Benchmark	Year	#Rel	Avg. citations	Sent.-level	Simul.	Generative	Structured tags	Typed relations
Explicit dataset (Xing et al., 2020)	2020	1	1.00	✗	✗	✓	✗	✗
FEVER (Chen et al., 2022)	2022	3	1.86	✓	✗	✗	✗	✗
ASQA (Gao et al., 2023)	2023	1	0	✓	✓	✓	✗	✗
WikiRetr (Li et al., 2024)	2024	2	1	✓	✗	✓	✗	✗
SCiFi (Cao and Wang, 2024)	2024	1	1.86	✗	✓	✓	✗	✗
ELI5 (Hsu et al., 2024)	2024	1	0	✓	✗	✓	✗	✗
FRONT (Huang et al., 2024)	2024	1	4.40	✓	✓	✓	✗	✗
MDS (Slobodkin et al., 2024)	2024	1	3.00	✓	✗	✓	✗	✗
CG (Li and Chen, 2025)	2025	3	1.00	✓	✗	✓	✗	✓
TROVE (Zhu et al., 2025)	2025	4	1.97	✓	✗	✗	✓	✓
ReFInE (Ours)	–	3	7.64	✓	✓	✓	✓	✓

Table 4: Comparison between ReFInE and representative benchmarks. We use ✓/✗ to denote Yes/No. "Avg. citations" denotes the average number of provenance triples or their equivalents in the reference answers. "Structured tags" indicates metadata-rich provenance tags (DocID, SentID, Relation) rather than plain indices. ReFInE distinguishes itself by enforcing generation-time, sentence-level provenance with explicit relation typing.

Model	Text Quality	Prov.	Quotation	Compression	Inference	Avg
Open-Source						
Baichuan2-7B (Yang et al., 2023)	3.04	0.37	0.39	0.05	0.05	0.78
Vicuna-7b-v1.5 (Zheng et al., 2023)	3.24	0.84	0.86	0.35	0.19	1.10
InternLM2.5-7B (Cai et al., 2024)	3.83	1.69	1.71	1.10	0.03	1.67
Hunyuan-7B (Zheng et al., 2025)	3.05	2.11	2.64	0.09	0.69	1.72
Yi-1.5-9B (Liu et al.)	3.84	2.12	2.62	0.23	0.17	1.80
Llama-3.1-8B (Grattafiori et al., 2024)	3.78	2.20	2.56	1.05	0.42	2.00
GLM-4-9B (GLM et al., 2024)	4.00	2.67	3.51	0.14	0.48	2.16
Qwen3-8B (Yang et al., 2025)	4.02	2.67	3.42	0.66	0.47	2.25
Gemma-3-12B (Team et al., 2025a)	3.97	2.89	4.03	1.06	0.42	2.47
Qwen3-14B (Yang et al., 2025)	4.10	3.03	4.06	1.56	0.21	2.59
GLM-4.5-355B (GLM et al., 2024)	**4.29**	3.02	4.13	0.91	0.79	2.63
Closed-Source						
Kimi (Team et al., 2025b)	4.13	2.70	3.60	0.44	0.13	2.20
GPT-5 (Achiam et al., 2024)	4.14	2.34	2.39	1.97	0.29	2.23
Gemini 2.5 Pro (Comanici et al., 2025)	4.24	3.02	4.33	1.09	0.19	2.57
Ours						
GenProve	3.98	**3.42**	**4.43**	**2.45**	**1.40**	**3.14**

Table 5: Breakdown of LLM-as-a-Judge evaluation. GenProve achieves superior overall ratings primarily through significant gains in complex *Compression* and *Inference* relations.

mine differences in Overall Prov. and thus Avg. This breakdown clarifies that improvements in Avg mainly coincide with better handling of abstraction and reasoning-based provenance, rather than with changes in Text Quality alone.

C.5 Full Ablation Results

Table 6 clarifies how each training component affects different dimensions of performance. GRPO primarily improves end-to-end utility under joint requirements, as reflected by a higher judge score compared with the SFT-only model. The provenance reward directly strengthens sentence-level attribution, and its removal leads to a broad collapse in provenance precision, recall, and F1. In contrast, the similarity reward provides an explicit pressure toward content faithfulness and wording alignment; removing it yields high provenance F1 but reduces answer-quality metrics and lowers the judge score, which suggests that provenance matching alone can be satisfied by outputs that are less faithful to the reference answer. The full model balances these pressures and achieves the best overall trade-off, which matches the intended objective of generation-time fine-grained provenance: accurate answers with localized and correctly typed evidence.

C.6 Human Evaluation Results

We conduct human evaluation to validate the judge-based results and to provide a fine-grained view of answer quality and provenance quality. For each model, raters score (i) answer quality while ignoring provenance tags and (ii) provenance quality,

Inference Prompt

You are a rigorous AI assistant specializing in traceable Question Answering. Your task is to generate an accurate, fluent, and factual answer based ONLY on the provided Source Documents.
CORE INSTRUCTIONS:
1. **Evidence-Based Generation**: Every sentence containing factual information must be supported by the Source Documents.
2. **0-Based Indexing**: Always use 0-based indexing for Document IDs (`Doc[0] → "0"`) and Sentence IDs exactly as they appear in the input.
3. **Strict Citation Format**: Append a citation tag at the end of every factual sentence. Format: `[PROVE: ("doc_id", "sent_id", "relation")]`

If multiple sources support the same sentence, merge them inside a *single* PROVE tag: Correct: `[PROVE: ("0","1","Quotation"), ("1","3","Inference")]` Incorrect: `[PROVE: ("0","1","Quotation")] [PROVE: ("1","3","Inference")]`
—
ATTRIBUTION RELATIONS
Choose exactly one relation for each citation tuple:
- **Quotation**: The answer sentence partially or fully copies the wording from the source (even with small rewrites). - **Compression**: The answer sentence condenses information from one or more source sentences. - **Inference**: The answer sentence expresses a fact implied but not explicitly stated by the source.
—

EXCLUSION RULES (Do NOT cite):
• Transition phrases (e.g., "To summarize", "In conclusion"). • Sentences that describe document structure without factual content.
—
ONE-SHOT EXAMPLE
User Input:
Question: What was the 2018 All-Ireland Senior Hurling Championship Final?
Source Documents:

Doc[0]: [0] "2018 All-Ireland Senior Hurling Championship Final The 2018 All-Ireland Senior Hurling Championship Final, the 131st event of its kind and the culmination of the 2018 All-Ireland Senior Hurling Championship, was played at Croke Park in Dublin on 19 August 2018.[1] The final was shown live in Ireland on RTÉ Two ... Match commentary was provided by Marty Morrissey ... The game was also shown live on Sky Sports ..."
Doc[1]: [0] "2018 All-Ireland Minor Hurling Championship ... ended on 19 August 2018 ..."
Doc[2]: [0] "2018 All-Ireland Senior Hurling Championship ... 131st staging ..."
—
Model Output:

The 2018 All-Ireland Senior Hurling Championship Final was the 131st event of its kind and the culmination of the 2018 All-Ireland Senior Hurling Championship. `[PROVE: ("0","0","Quotation"), ("2","0","Compression")]`
It was played at Croke Park in Dublin on August 19, 2018, and was shown live in Ireland on RTÉ Two as part of "The Sunday Game" live program... `[PROVE: ("0","0","Quotation"), ("0","1","Quotation")]`
Match commentary was provided by Marty Morrissey with analysis by Michael Duignan. `[PROVE: ("0","2","Quotation")]`
The game was also shown live on Sky Sports, presented by Rachel. `[PROVE: ("0","3","Quotation")]`

Figure 12: Prompt used for model inference with structured provenance.

Model	ROUGE-L	BLEU	METEOR	BERTScore	MoverScore	Prec.	Rec.	F1	Format	Judge
GenProve	57.25	42.22	59.39	61.98	51.04	54.96	51.26	51.21	99.85	3.14
w/o Provenance Reward	36.85	11.94	25.06	46.96	30.04	27.10	24.08	24.67	96.66	2.20
w/o Similarity Reward	38.72	24.60	54.97	44.50	35.49	67.27	58.60	60.32	95.74	2.71
w/o GRPO	56.08	41.82	59.96	60.70	49.79	53.22	51.65	50.48	99.70	2.62

Table 6: Full ablation results on ReFInE (EVAL).

including an overall provenance score and relation-specific provenance scores for Quotation, Compression, and Inference. Table 7 reports the averaged scores across models.

Overall. Human scores broadly track the main results: models that achieve higher judge scores also receive higher overall human scores (Avg), which supports the use of judge-based evaluation for this task.

Model groups. Open-source models show a wide spread in provenance-related scores, ranging from near-failing provenance (e.g., low Prov. and relation scores) to much stronger traceability among recent instruction-tuned systems. Closed-source models are generally strong on answer quality, while provenance remains uneven across relation types. GenProve achieves the highest Avg and the strongest overall provenance score, indicating that improvements are not limited to fluency but extend to evidence attribution.

Metric dimensions. Answer scores are relatively high for many models, while provenance scores are substantially lower and vary more by relation type. In particular, Quotation tends to score higher than Compression and Inference, consistent with the increasing difficulty of abstraction and reasoning under sentence-level evidence constraints. GenProve improves all three relation types and shows especially large gains on Compression and Inference, which aligns with the goal of generation-time fine-grained provenance.

D Error Analysis

We present a qualitative error analysis to illustrate representative failure modes in sentence-level, typed provenance generation. Although GenProve significantly improves attribution accuracy, strict generation-time provenance supervision still poses challenges. The following examples highlight four common error patterns observed across models.

Unsynchronized Provenance Generation. In this failure mode, the model produces answer con-

Text Quality Evaluation Prompt

You are a content quality evaluation expert. Your task is to evaluate the text quality of a Q&A model's response.
Input Data:

- `question`: The user's query.
- `documents`: The source material provided to the model.
- `labels`: The standard reference answer (Ground Truth).
- `response`: The model's generated answer.

Objective: Evaluate the natural language answer. Compare the model's `response` against the `labels` and `documents`.
Scoring Criteria (0–5):

Focus on **Accuracy**, **Fluency**, and **Completeness**.
5 (Perfect): Accurate, fluent, complete. No hallucinations.
4 (Good): Basically accurate. Covers main points.
3 (Acceptable): Captures core answer, minor slips.
2 (Poor): Misses key info, hallucinations, or poor grammar.
1 (Very Poor): Barely relevant or severe errors.
0 (Useless): Completely wrong or empty.

—

Output Format (JSON only):

```
{
  "id": "<id>",
  "question": "<question>",
  "text_quality_score": <integer 0-5>,

  "text_quality_reasoning": "<Concise explanation>"
}
```

—

Reference Example:
Reference input:

```
{
  "id": "e318e8cf-cfa9-4889-8a2e-b37b18b64ac7",
  "question": "What happened to the Milwaukee Brewers in the 2008
  National League Division Series?",
  "documents": { ... },
  "response": "...",
  "labels": "..."
}
```

Reference output:

```
{
  "id": "e318e8cf-cfa9-4889-8a2e-b37b18b64ac7",
  "question": "What happened to the Milwaukee Brewers in the 2008
  National League Division Series?",
  "text_quality_score": 4,
  "text_quality_reasoning":
    "Response accurately and fluently states the Brewers played
     and were eliminated by the Phillies in the 2008 NLDS.
     However, it omits the context of clinching a wild card spot
     with a 90-72 record, making it slightly less complete."
}
```

Figure 13: Prompt used for evaluating the text generation quality of model responses.

tent and provenance tags in an unsynchronized manner. Provenance annotations are delayed or structurally detached from the sentences they are intended to support, resulting in partially traceable outputs. This error reflects the difficulty of tightly coupling natural language generation with structured attribution decisions at token level during decoding.

Incomplete Provenance Coverage. Here the model generates factually plausible answer sentences but omits provenance tags for some of them. Such errors break sentence-level verifiability even when the content itself is supported by the source documents. This pattern indicates a recall failure in provenance generation, where the model underestimates the need for explicit attribution under strict coverage requirements.

Incorrect Provenance Localization. In this case, the model emits well-formed provenance tags, but the referenced document or sentence indices do not actually contain the supporting evidence. Although the cited source is often topically related, the precise sentence-level grounding is incorrect. This error highlights the challenge of fine-grained evidence localization under multi-document settings.

Incorrect Provenance Type. In this error pattern, the cited evidence is relevant, but the predicted relation type (Quotation, Compression, or Inference)

Model	Text Quality	Prov.	Quotation	Compression	Inference	Avg
Open-Source						
Baichuan2-7B (Yang et al., 2023)	3.65	0.41	0.51	0.20	0.10	0.98
Vicuna-7b-v1.5 (Zheng et al., 2023)	2.82	0.62	0.75	0.16	0.41	0.95
InternLM2.5-7B (Cai et al., 2024)	3.91	1.49	1.51	1.16	0.10	1.64
Hunyuan-7B (Zheng et al., 2025)	3.45	2.38	3.12	0.61	1.08	2.13
Yi-1.5-9B (Liu et al.)	3.98	2.05	2.44	0.10	0.20	1.76
Llama-3.1-8B (Grattafiori et al., 2024)	3.82	2.40	2.74	1.16	1.20	2.27
GLM-4-9B (GLM et al., 2024)	4.02	2.45	2.85	0.33	0.92	2.11
Qwen3-8B (Yang et al., 2025)	4.03	2.63	3.36	0.82	1.26	2.42
Gemma-3-12B (Team et al., 2025a)	3.90	3.03	3.69	1.14	0.96	2.54
Qwen3-14B (Yang et al., 2025)	3.99	2.91	3.25	1.20	0.47	2.36
GLM-4.5-355B (GLM et al., 2024)	4.15	2.90	3.59	0.99	0.30	2.39
Closed-Source						
Kimi (Team et al., 2025b)	4.21	2.44	3.85	0.38	0.21	2.22
GPT-5 (Achiam et al., 2024)	4.24	3.17	3.22	2.33	0.63	2.72
Gemini 2.5 Pro (Comanici et al., 2025)	**4.36**	3.23	3.95	1.28	0.16	2.60
Ours						
GenProve	4.29	**3.83**	**4.16**	**2.57**	**1.94**	**3.36**

Table 7: Human evaluation results. The manual ratings corroborate automatic metrics, confirming GenProve's superiority in generating verifiable answers with accurate fine-grained provenance.

does not match the reference annotation. Such mistakes arise from ambiguity between surface copying, abstraction, and reasoning, especially when evidence is partially transformed. This illustrates the intrinsic difficulty of relation-type discrimination in fine-grained provenance generation.

E Potential Risks

While generation-time provenance improves transparency and auditability, it also introduces several potential risks that merit careful consideration.

Over-reliance on provenance signals. Typed, sentence-level provenance may give users a strong sense of trust in generated answers. However, correct provenance does not guarantee that a statement is fully accurate or appropriate for a given context. Provenance should therefore be interpreted as an aid for inspection rather than a definitive validation of correctness, especially in high-stakes domains.

False sense of completeness. Our framework focuses on identifying supporting evidence for generated sentences, but it does not ensure that all relevant evidence has been considered. A model may provide plausible and correctly typed provenance while still omitting counter-evidence or alternative interpretations present in the source documents.

Annotation and evaluation bias. PROVE-ASQA relies on LLM-assisted annotation followed by expert validation. Although we apply multi-stage screening and human verification, residual biases from annotator judgment or model priors may affect relation labeling, particularly for subjective cases such as *Inference*. These biases could influence both training and evaluation outcomes.

Computational and deployment considerations. Generating sentence-level provenance alongside answers increases output length and computational cost, which may limit applicability in latency-sensitive or resource-constrained settings. Careful system design is required to balance transparency with efficiency.

We emphasize that GenProve is intended as a research step toward more accountable text generation. It should be deployed as part of broader human-in-the-loop or verification workflows rather than as a standalone authority on factual correctness.

Traceability Evaluation Prompt

You are a rigorous citation evaluation expert. Your task is to evaluate the **Traceability** of a Q&A model's response. You must verify the model's `[PROVE]` tags against the provided `documents` and the standard ground-truth `labels`.
Input Data:

- `id`: Data item ID.
- `question`: The user's query.
- `documents`: The source material.
- `labels`: The standard reference answer (Ground Truth).
- `response`: The model's generated answer (containing `[PROVE]` tags).

The definitions of attribution relationships in the `[PROVE]` tags are as follows:
Quotation: The answer sentence partially or fully copies sentences from the source document.
Compression: The answer sentence condenses information from one or more sentences.
Inference: The answer sentence is based on information implied by the source document.
Evaluate the correctness and completeness of the `[PROVE]` tags by comparing the `response` against the `labels`.
CRITICAL SCORING LOGIC:
1. Check for Missing Types (Recall): If a relationship type exists in `labels` but is **NOT** in `response`, the score for that type **MUST** be **0**.
2. Check for Unused Types: If a relationship type is **NOT** in `labels` AND **NOT** in `response`, the score **MUST** be `null`.
3. Check for Accuracy (Precision): If the type exists in `response`, score it based on correctness (0–5) relative to the documents:
5: Most tags for this type are correct and match the standard labels logic.
4: More than half for this type are correct; minor issues with boundaries.
3: Mixed accuracy; only about one-third for this type are correct.
2: Only about one-eighth of them are correct.
1: Few tags are correct; vast majority are hallucinations.
0: All tags are hallucinations, irrelevant, OR the type is required by labels but missing in response.
2. Overall Citation Score (0–5)
Provide a holistic score for the model's citation performance.
5: Perfect. Captures all relationships required by `labels` with accurate citations.
4: Good. Captured all required relationships but with minor inaccuracies in index or boundaries.
3: Acceptable but flawed. Missed one relationship type or has several inaccurate citations.
2: Poor. Missed multiple required relationships or most citations are wrong.
1: Very Poor. Citations are mostly hallucinated or irrelevant.
0: No valid citations or complete failure to follow instructions.
Special Note on Indexing:
Source `documents` are 0-indexed. If the answer text refers to "Document 1" but the `[PROVE]` tag uses index `"0"`, this is considered correct and should not be penalized.
Output Format:
Strictly output in the following JSON format. Do not add any extra explanations.

```
{
  "id": "<id of the current data item>",
  "question": "<question of the current data item>",
  "relationship_scores": {
    "Quotation": <integer from 0 to 5 OR null>,
    "Compression": <integer from 0 to 5 OR null>,
    "Inference": <integer from 0 to 5 OR null>
  },
  "overall_citation_score": <integer from 0 to 5>,
  "citation_reasoning":
    "<Explain the scores. Explicitly mention if a type required
      by 'labels' was missed (Recall error) or if the generated
      tags were inaccurate (Precision error).>"
}
```

Reference Example (abridged):

```
Reference input: {...}

Reference output:
{
  "relationship_scores": {
    "Quotation": 3,
    "Compression": null,
    "Inference": 0
  },
  "overall_citation_score": 3,
  "citation_reasoning":
    "Recall error: the required 'Inference' relationship is missing.
     Precision error: one 'Quotation' tag is misaligned with the
     referenced document sentence."
}
```

Figure 14: Prompt used for evaluating the correctness and completeness of provenance annotations (`[PROVE]` tags), with explicit modeling of recall and precision errors.

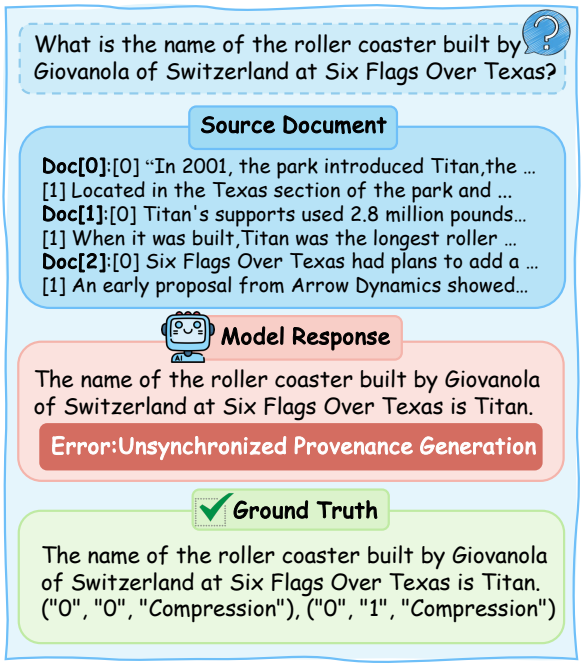

Figure 15: Unsynchronized provenance generation, where answer content and provenance tags are misaligned in generation.

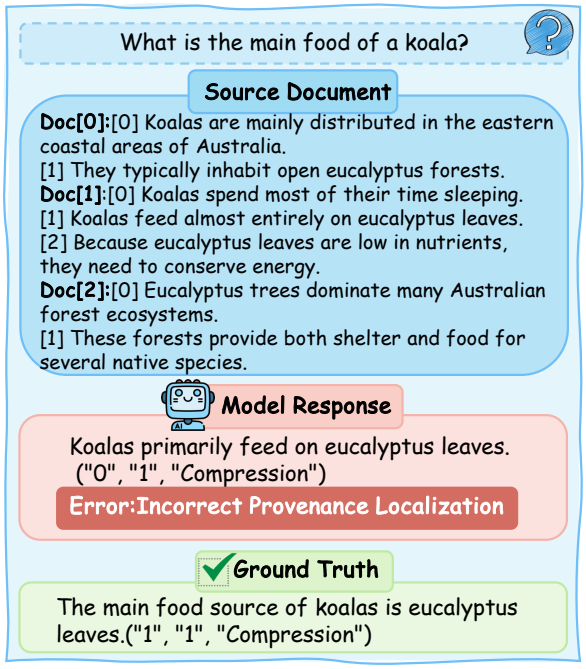

Figure 17: Incorrect provenance localization, where cited document or sentence indices do not support the answer.

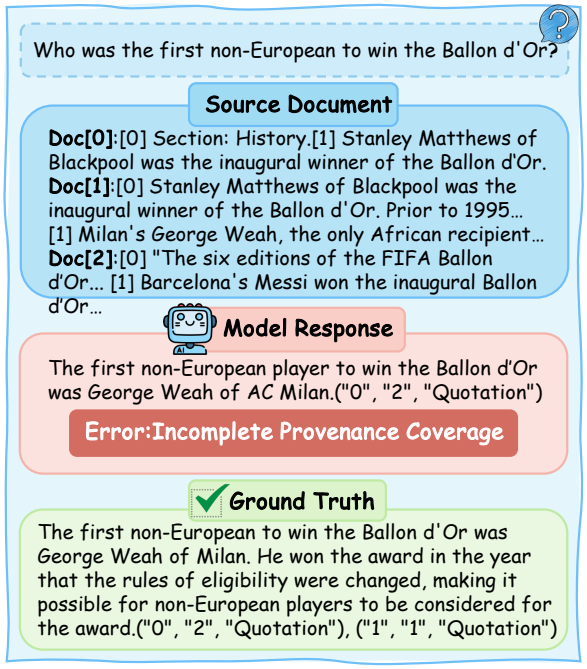

Figure 16: Incomplete provenance coverage, where some factual sentences lack required provenance tags.

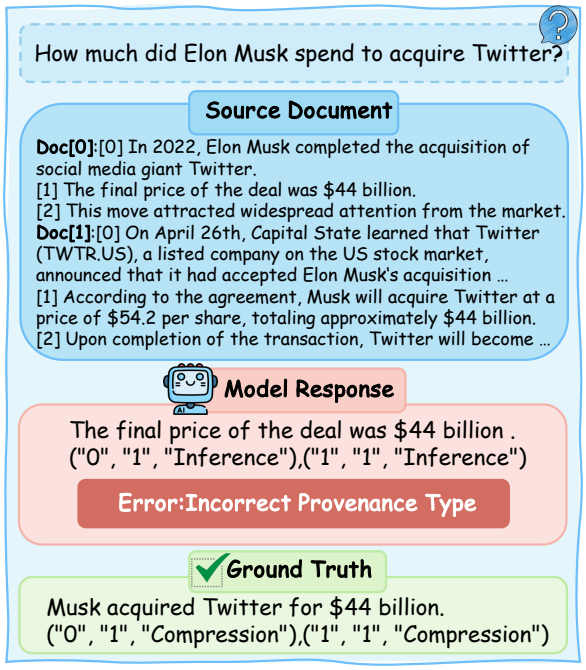

Figure 18: Incorrect provenance type, where the relation label does not match the reference annotation.