LongAF: A Scalable Evidence-Chain Generation Framework for Long-Document QA Annotation

Anonymous ACL submission

Abstract

Long-context modeling has become a core capability of large language models (LLMs), underpinning applications from multi-turn dialogue to agentic systems. Yet the high-quality annotated data needed to train and evaluate such systems on long-document question answering (QA) remains scarce. Curating such data is uniquely demanding: human annotators must process documents stretching to tens of thousands of tokens, locate sparsely scattered evidence, and reason across widely separated passages, while model-assisted alternatives are undermined by "lost-in-the-middle" errors and length-extrapolation degradation as inputs grow. In this paper, we introduce **LongAF**, a scalable framework that recasts long-document QA annotation as *verification over a compact, evidence-grounded substrate* rather than end-to-end reading. LongAF proceeds in three stages: (1) *multi-path evidence retrieval* that distills candidate evidence from the document; (2) an *MCTS-based reasoning-chain generator* that organizes the evidence into a verifiable reasoning chain; and (3) an *interactive interface* that lets annotators verify and refine the chain. Empirical results show that LongAF matches or slightly exceeds fully manual annotation quality, clearly outperforms LLM-only baselines, and reduces annotation time and cost to **1/8** and **1/4**, respectively.

1 Introduction

Long-context modeling has emerged as a core capability of large language models (LLMs), powering applications from multi-turn dialogue to agentic systems (Tan et al., 2024a; Zhuang et al., 2025). Cultivating this capability depends on training data that is both authentic and diverse—drawn from real-world long documents at sufficient scale. Existing corpora satisfy neither criterion: they are predominantly synthetic and narrow in domain, providing an inadequate foundation for robust long-context models. The natural remedy—carefully annotated, real-world long documents—remains scarce, as such annotation is uniquely labor-intensive and costly (Wang et al., 2024a; Dasigi et al., 2021).

Existing annotation pipelines—whether driven by human annotators (e.g., experts or crowd workers) (Kittur et al., 2008; Snow et al., 2008) or by LLM-assisted approaches (Wang et al., 2024b; Li et al., 2023; Kim et al., 2024)—perform well on short passages but falter on long documents. As illustrated in Figure 1, we attribute this gap to three intrinsic properties of long documents: (*i*) *Information Overload*—the sheer token count imposes a heavy cognitive load; (*ii*) *Low Information Density*—critical evidence is buried in long stretches of irrelevant content; and (*iii*) *Frequent Context Shifts*—resolving a single query requires integrating evidence dispersed across widely separated passages. Together, these properties expose two complementary failure modes, one in each paradigm. (1) **Limited Annotation Efficiency.** For human annotators, processing such documents is cognitively taxing and resource-intensive, rendering large-scale, high-quality dataset construction impractical. (2) **Compromised Annotation Accuracy.** LLM-assisted methods are themselves unreliable on long inputs: the *lost-in-the-middle* effect (Huang, 2024; Zhang et al., 2025, 2024) and *length-extrapolation* failures (Chen et al., 2023a; Qin et al., 2024) degrade answer quality as documents approach or exceed the effective context window. The sparse, dispersed distribution of evidence further compounds these issues, propagating hallucinations and factual inconsistencies into the final annotations. Neither paradigm alone can therefore yield real-world long-document QA data at scale, motivating an annotation framework.

To address these challenges, we introduce **LongAF**, a three-stage framework that recasts long-document QA annotation as *verification over a*

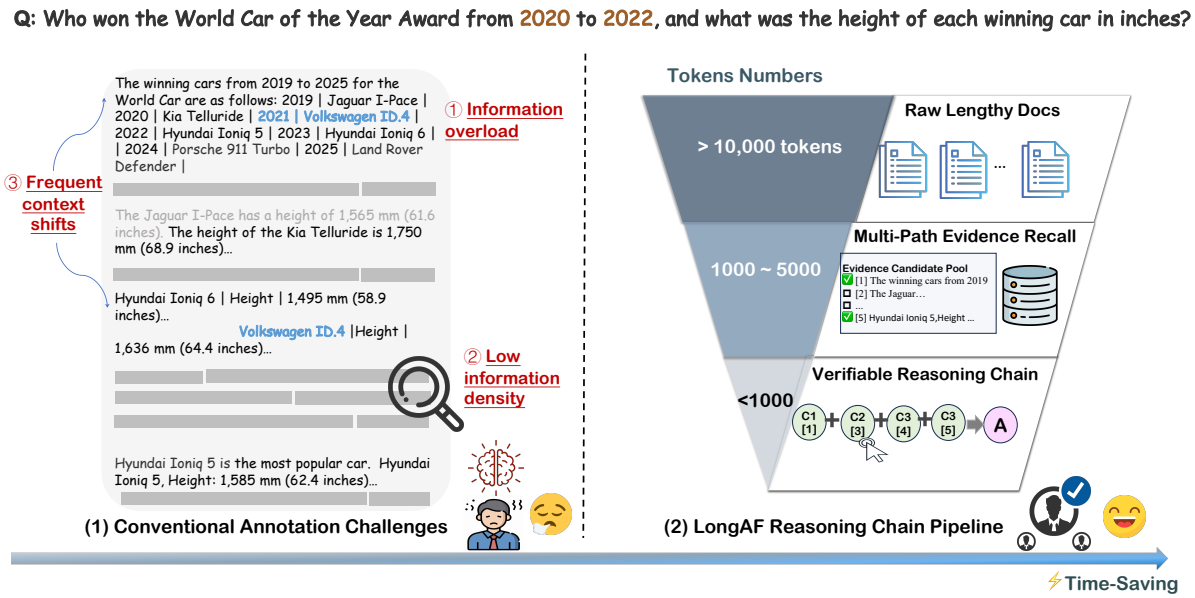


Figure 1: Conventional annotation vs. LongAF. **Left:** annotators face cognitive overload from evidence sparsely scattered across >10,000-token documents. **Right:** LongAF distills the document into a multi-path evidence pool and organizes key evidence into a verifiable reasoning chain (<1,000 tokens) for human verification.

compact, evidence-grounded substrate rather than end-to-end reading. (1) *Multi-Path Evidence Retrieval* couples query expansion with parallel sparse, dense, and LLM-augmented retrievers, distilling the document into a high-recall pool of candidates and narrowing the search space. (2) *MCTS-based Reasoning Chain Generator* uses Monte Carlo Tree Search to organize the retrieved evidence into a verifiable reasoning chain, replacing open-ended reasoning with structured steps that annotators can verify. (3) *Interactive Human Verification* offers a lightweight interface that lets annotators inspect, refine, and complete each evidence–reasoning step against the source document. Across seven long-document QA benchmarks, LongAF matches or slightly exceeds fully manual annotation quality, clearly outperforms LLM-only baselines, and reduces annotation time and cost to **1/8** and **1/4**, respectively. This work makes three contributions:

- **A new annotation paradigm.** We reformulate long-document QA annotation as localized verification over a compact, evidence-grounded substrate, replacing whole-document reading and brittle end-to-end LLM annotation while preserving human control over grounding.
- **The LongAF framework.** We realize this paradigm in a three-stage pipeline that integrates multi-path retrieval, MCTS-guided reasoning-chain construction, and an interactive verification interface, released as an open-source toolkit.
- **Empirical validation and a new resource.** On seven long-document QA benchmarks, LongAF attains fully manual annotation quality at **1/8** the time and **1/4** the cost; we further release the resulting fine-grained, verifiable evidence chains to augment widely-used long-document QA benchmarks.

2 Related Work

Data annotation methodologies for long-document QA fall into three paradigms: *purely manual*, *model-based*, and *human–LLM collaborative*.

2.1 Manual Annotation Paradigms

Manual annotation has long served as the primary mechanism for constructing high-quality QA datasets (Kittur et al., 2008; Snow et al., 2008). For long documents, prior work has documented substantial cognitive load on annotators (Lenzner et al., 2010; Mitchell et al., 2021). Reported costs are non-trivial: constructing DetectiveQA required, on average, 3.5 hours of annotation per book, with some cases reaching 12 hours (Xu et al., 2024b; Wang et al., 2024a); LongBench_v2 reports 53.7% unaided annotation accuracy (Bai et al., 2024); and ChatterEVAL reports

an inter-annotator agreement of Krippendorff's $\alpha = 0.448$ (Baruah and Narayanan, 2024).

2.2 Model-Based Annotation Paradigms

A growing line of work uses LLMs as automated annotators. With suitable prompting, modern LLMs have been reported to approach or match crowd workers on tasks ranging from zero-shot sentiment analysis to complex reasoning (Ding et al., 2022; Törnberg, 2023; Tan et al., 2024b). On long inputs, however, two failure modes have been documented: LLMs miss evidence located in the middle of long contexts—the "lost-in-the-middle" effect (Liu et al., 2023; Huang, 2024; Zhang et al., 2025, 2024)—and exhibit length-extrapolation degradation once inputs exceed their training horizon (Chen et al., 2023a; Qin et al., 2024). Hallucination in open-ended generation has also been widely reported. Extending the context window beyond 100K tokens has not been shown to eliminate these issues, while compute and latency costs grow substantially with input length.

2.3 Human–LLM Collaboration Paradigms

Human–LLM collaboration combines model scalability with human judgment (Wang et al., 2024b). Representative systems span a range of tasks: CoAnnotating allocates instances between humans and LLMs according to model uncertainty (Li et al., 2023); MEGAnno+ provides customizable tasks and label schemas for general-purpose annotation (Kim et al., 2024); SQLSynth supports interactive schema visualization for text-to-SQL workflows (Tian et al., 2025). To our knowledge, no existing collaborative system is specifically designed for multi-hop evidence selection over long documents.

3 Methodology

As shown in Figure 2, LongAF reduces the cognitive burden on human annotators by compressing long documents into concise reasoning chains. The framework comprises three stages: evidence candidate retrieval, reasoning path generation, and human-in-the-loop verification.

3.1 Task Formulation and Motivation

Given a query Q and a long document $\mathcal{D} = \{s_1, s_2, \ldots, s_N\}$ where s_i denotes a sentence, the objective is to generate an answer A supported by a set of evidence spans $\mathcal{E} = \{e_1, e_2, \ldots, e_k\}$, where each e_i is a contiguous span of $\mathcal{D}$. To support multi-hop reasoning, we represent the reasoning path as a sequence of triplets $\mathcal{P} = \big[(sq_1, e_1, sa_1), \ldots, (sq_k, e_k, sa_k)\big]$, where sq_i and sa_i denote the sub-query and sub-answer.

Long-document answer annotation overwhelms annotators along three axes: *information overload* from repeated search, *contextual dependency* from resolving long-range references, and *attention decay* from sustained focus. Together these strain the limited capacity of human working memory (Mitchell et al., 2021), and the resulting fatigue further amplifies *automation bias* in LLM-assisted settings, where annotators uncritically accept model suggestions (Jacovi et al., 2024). We mitigate this by compressing long documents into concise *reasoning chains* that filter redundancy and surface long-range dependencies (Pang et al., 2025), guided by two objectives: (i) maximizing recall of relevant snippets so that annotation accuracy is not bottlenecked by missing evidence; and (ii) synthesizing the recalled evidence into structured reasoning paths to reduce annotators' search effort and cognitive burden.

3.2 Multi-Path Evidence Retrieval

This stage condenses a long document into a high-recall pool of evidence candidates, dramatically narrowing the search space for downstream reasoning, which is built through four steps.

Document Splitting. We partition the document into overlapping chunks; the overlap prevents relevant information from being fragmented across chunk boundaries.

Query Expansion. We use an LLM to extract key terms and decompose complex queries into sub-questions, broadening retrieval coverage on multi-faceted queries.

Multi-Source Parallel Retrieval. For broad coverage, we run three retrievers in parallel: (i) *Sparse Retrieval (BM25)* (Robertson et al., 2009) for lexical matching against chunks; (ii) *Dense Retrieval (BGE)* (Chen et al., 2024) for semantic matching via dense embeddings; and (iii) *LLM-Augmented Retrieval*, where an LLM extracts evidence snippets directly from the document.

Evidence Ranking. Retrieved chunks are clustered and aggregated into the final "Evidence Bank". We compute pairwise similarity $\mathrm{Sim}_{\mathrm{sem}}(C_i, C_j)$ via Sentence-Transformers (all-MiniLM-L6-v2) (Yin and Zhang, 2024); chunks with similarity above τ (e.g., τ=0.95) are merged

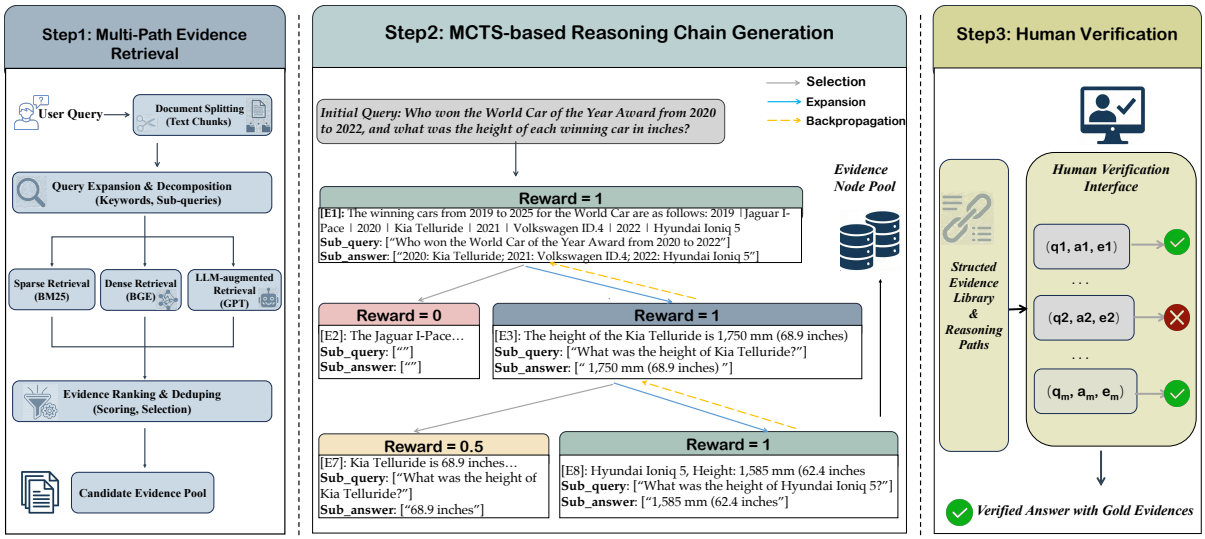


Figure 2: Overview of the LongAF framework: (1) Multi-Path Evidence Retrieval collects candidate evidence; (2) MCTS-based Reasoning Chain Generator organizes it into verifiable reasoning paths; and (3) Interactive Human Verification validates and refines the chain.

into a group $G_k = \{C_1, \ldots, C_m\}$, whose aggregated score averages normalized retrieval scores:

$$S_{G_k} = \frac{1}{|G_k|} \sum_{C \in G_k} S_C, \tag{1}$$

where S_C is the relevance score of chunk C.

3.3 MCTS-based Reasoning Chain Generator

The combinatorial space of evidence permutations grows exponentially, making exhaustive search intractable. We therefore use Monte Carlo Tree Search (MCTS) (Bagaria et al., 2021; Xie et al., 2024) to identify high-quality reasoning chains.

Node Selection. Starting from the root s_0, we iteratively select the child with the highest Upper Confidence Bound for Trees (UCT) score (Kocsis and Szepesvári, 2006; Mern et al., 2021):

$$\mathrm{UCT}(s_l) = V(s_l) + w\sqrt{\frac{\ln N(s_p)}{N(s_l)}}, \tag{2}$$

where s_l and s_p denote the candidate child and its parent, $V(\cdot)$ the expected value, $N(\cdot)$ the visit count, and w a tunable exploration weight.

Evidence Expansion. From the selected node, an unvisited evidence chunk is drawn from the candidate pool by relevance score. To accelerate the search, we expand multiple child nodes in parallel.

Reward Modeling. Each node maintains an unsolved sub-query set Q_U and a solved set Q_S. The reward for adding evidence e_{t+1} at step $t+1$ is

$$r_{t+1} = \begin{cases} +1, & \text{resolves a sub-query in } Q_U, \\ +0.5, & \text{supports a sub-query in } Q_S, \\ 0, & \text{irrelevant to any sub-query}, \\ -1, & \text{conflicts with an answer in } Q_S. \end{cases} \tag{3}$$

Q_S is also presented to annotators as a reference.

Backpropagation. The reward r is propagated along the path from the root to the current node s_t (Silver et al., 2016), updating its visit count and value:

$$\begin{aligned} N_{\text{new}}(s_t) &= N_{\text{old}}(s_t) + 1, \\ V_{\text{new}}(s_t) &= \frac{V_{\text{old}}(s_t)\, N_{\text{old}}(s_t) + r}{N_{\text{new}}(s_t)}. \end{aligned} \tag{4}$$

3.4 Interactive Human Verification

A final human pass turns the automated reasoning chain into a reliable annotation. Annotators verify each (sq_i, e_i, sa_i) triplet through two operations.

Verifying Evidence Nodes. The presence of e_i does not guarantee that it supports sq_i or sa_i. Annotators click any snippet to view its source-document context (located via BM25 matching) and confirm whether it supports the corresponding sub-question and sub-answer, removing hallucinated evidence.

Completing the Reasoning Chain. When the chain is judged incomplete, annotators add missing nodes without rereading the document, using two supplements: (i) the *Auxiliary Evidence*

Table 1: **Dataset Statistics**. *Avg. Input* is the average input length in tokens; *#Samples* is the number of evaluation samples; *Avg. Evidence* is the average number of evidence spans per sample.

Dataset	LongBench					FanoutQA	LongBench_v2
	HotpotQA	MuSiQue	NarrativeQA	Qasper	2WikiMQA		
Avg. Input	10,603	12,975	71,787	4,236	11,765	89,823	108,478
#Samples	200	200	200	200	200	200	300
Avg. Evidence	2.11	1.37	1.45	1.18	2.35	4.86	2.67

Node Pool, which retains MCTS-explored but unselected candidates; and (ii) *Rapid Retrieval from the Evidence Bank*, where the high-recall design of the first stage makes missing snippets likely available for direct search or browsing. Together, the two operations ensure that every evidence node is reliable and the reasoning path is complete.

4 Experimental Protocol

We evaluate LongAF along three axes: annotation *quality*, *efficiency*, and *cost*.

4.1 Datasets

We evaluate LongAF on three benchmarks. From LongBench (Bai et al., 2023), we use five multi-hop QA subsets that require reasoning across multiple passages: HotpotQA (Yang et al., 2018), MuSiQue (Trivedi et al., 2022), NarrativeQA (Kočiskỳ et al., 2018), Qasper (Dasigi et al., 2021), and 2WikiMQA (Ho et al., 2020). We further include FanoutQA (Zhu et al., 2024), which requires aggregating evidence from at least four sources, and LongBench_v2 (Bai et al., 2024), a real-world long-context benchmark covering academic, legal, financial, and government domains. For diagnostic analysis, we manually annotate the supporting evidence segments for all datasets, as summarized in Table 1.

4.2 Baselines and Experimental Settings

We compare three annotation paradigms: human-only, model-only, and human-LLM collaboration. We recruit 30 university-educated annotators with strong English proficiency and split them into two groups of 15 to prevent information leakage. Detailed protocols are provided in Appendix A.1.

Human-based Method. To establish a high-quality upper bound and simulate conventional workflows, this group performs annotations without AI assistance, working within a standard document editor (Microsoft Word) with only basic keyword search available.

Model-based Baselines. We evaluate two families. *End-to-end LLMs* ingest the entire document with Chain-of-Thought (CoT) prompting (Wei et al., 2022; Gao et al., 2024): GPT-4o-mini (Achiam et al., 2023), DeepSeek-R1 (Guo et al., 2025), LLaMA-3.1-8B-Instruct (Dubey et al., 2024), Qwen-2.5-7B-Instruct (Team, 2024), and Qwen3-4B-Thinking.[1] *Chunk-based approaches* retrieve relevant segments to bypass context-window limits, including standard RAG (Li et al., 2024; Jiang et al., 2024; Zhao et al., 2024) and two reading-inspired agents: MemWalker (Chen et al., 2023b) (hierarchical summary trees) and ReadAgent (Lee et al., 2024) (gist-based compression).

Human-LLM Collaboration Method. This group uses our **LongAF** framework. Unlike the *human-only* group, annotators interact with a specialized interface to verify and refine the reasoning chains, combining the efficiency of automated retrieval with the precision of expert oversight.

4.3 Evaluation Metrics

We evaluate annotation along three axes: (i) *Quality*—accuracy computed by GPT-4o-mini as the proportion of correctly identified answer spans, handling queries with multiple valid answers (Xu et al., 2024a), with the judge validated by a meta-evaluation on a stratified sample of 150 instances against independent human experts (96.45% agreement); (ii) *Efficiency*—the total annotation time per sample, summing model inference latency and human time as applicable; and (iii) *Cost*—human time priced at $5.56/hour (Wang et al., 2024a; Li et al., 2023), with model cost based on GPT-4o-mini API pricing of $0.005 per 1K prompt tokens.

4.4 Hyper-parameter Settings

We use a chunk size of 256 tokens and retrieve the top k=20 chunks. The LLM backbone is GPT-4o-

[1] https://qwen.readthedocs.io/en/latest/

Table 2: Performance comparison on LongBench. We report Accuracy (Acc, %), Expert Time (Min), and Total Cost (USD). Best results are **bolded** and second-best are underlined.

Method	Acc (↑)					Avg Acc. (↑)	Time (Min ↓)	Cost (USD ↓)
	HotpotQA	MuSiQue	NarrativeQA	Qasper	2WikiMQA			
HUMAN-BASED METHODS								
Full_Human	95.43	**93.50**	**94.86**	92.64	93.35	93.96	9.45	783.02
MODEL-BASED METHODS								
End-to-End LLMs								
GPT-4o-mini	64.81	40.42	33.29	48.84	56.33	48.74	0.53	25.22
DeepSeek-R1	69.99	62.66	21.51	40.00	78.23	54.48	0.64	*
LLaMA-3.1-8B-Instruct	55.22	29.77	26.67	43.80	42.69	39.63	0.42	*
Qwen-2.5-7B-Instruct	56.24	29.53	33.33	44.04	47.74	42.18	0.48	*
Qwen3-4B-Thinking	62.45	38.13	36.15	46.35	52.15	47.05	0.56	*
Chunk-Based Approaches								
RAG (BM25)	38.98	28.86	18.86	28.40	26.98	28.42	0.85	*
MemWalker (Chen et al., 2023b)	69.62	56.37	53.32	64.44	57.34	60.21	1.88	49.21
ReadAgent (Lee et al., 2024)	73.32	53.87	46.36	60.24	64.14	59.78	1.01	44.12
HUMAN-LLM COLLABORATION								
LongAF (Ours)	**96.24**	92.68	93.56	**93.24**	**96.52**	**94.45**	1.32	185.05
w/o Human Check	78.05	62.74	68.58	71.57	70.45	70.88	0.89	58.86

* indicates that we employ locally deployed open-source models, which incur no API-based inference cost.

mini with a context window of 2048 tokens.

5 Experimental Results

We report main results, ablations, generalization across LLM backbones, and error analysis.

5.1 Main Results

Tables 2 and 3 show that LongAF reconciles the trade-off between annotation quality and scalability. LongAF reaches 93.45% accuracy on LongBench_v2, the most challenging benchmark, surpassing both automated and human baselines. End-to-end models suffer from the "Lost-in-the-middle" phenomenon (e.g., GPT-4o-mini: 46.30%), and chunk-based RAG struggles on multi-hop queries (e.g., 28.86% on MuSiQue); by constructing explicit reasoning chains with MCTS, LongAF combines high retrieval recall with logical coherence. Notably, it also marginally exceeds the Full Human baseline (92.05%), suggesting that offloading retrieval mitigates the cognitive fatigue and attention decay that typically degrade human performance on ultra-long documents.

Beyond accuracy, LongAF substantially advances the cost-quality Pareto frontier. Average annotation time per sample drops from 20.85 to 2.68 minutes (a nearly 8× speedup), and total cost drops by 77% (from $468.71 to $106.76). To quantify objectivity, we further conduct a blind inter-annotator agreement (IAA) study on 150 randomly sampled LongBench_v2 instances with three independent expert annotators. Unassisted annotation yields Fleiss' $\kappa = 0.78$ (substantial), whereas LongAF improves consistency to $\kappa = 0.92$ (near-perfect). We attribute this gain to a paradigm shift from "unstructured search" to "structured verification", where anchoring attention on model-generated evidence paths decouples annotation quality from individual annotator variance.

5.2 Ablation Study

We ablate the two key components of LongAF: multi-path evidence retrieval and the MCTS-based reasoning chain generator.

5.2.1 Effect of Multi-Path Evidence Retrieval

Table 4 shows that removing any component of our multi-path retrieval significantly degrades recall. Disabling **Query Expansion** causes the largest drop (↓37.21%), confirming its role in broadening the search scope. Relying on a single retriever also hurts substantially (LLM: ↓30.13%; BGE: ↓17.04%; BM25: ↓13.27%), showing that **multi-source coverage** is essential. Finally, removing **Evidence Ranking** yields a ↓25.76% decline, demonstrating that effective evidence ordering matters beyond random top-k selection.

Table 3: Performance comparison on FanoutQA and LongBench_v2 (Real-world). We report Accuracy (Acc, %), Expert Time (Min), and Cost (USD). Domains in LongBench_v2 are abbreviated as: **Aca**demic, **Leg**al, **Fin**ancial, and **Gov**ernment. Best results are **bolded** and human performance is underlined.

Method	FanoutQA			LongBench_v2 (Real-world)						
	Acc (↑)	**Time** (↓)	**Cost** (↓)	Aca.	Leg.	Fin.	Gov.	**Avg Acc.** (↑)	**Time** (↓)	**Cost** (↓)
HUMAN-BASED METHODS										
Full_Human	92.86	15.64	471.33	90.86	92.54	91.56	93.23	92.05	20.85	468.71
MODEL-BASED METHODS										
End-to-End LLMs										
GPT-4o-mini	49.46	1.01	18.86	39.46	46.90	48.55	50.28	46.30	1.21	23.61
DeepSeek-R1	53.86	1.32	*	42.48	52.90	56.24	52.64	49.30	1.32	*
LLaMA-3.1-8B-Instruct	39.68	0.67	*	34.68	38.25	41.53	42.00	39.12	0.75	*
Qwen-2.5-7B-Instruct	40.24	0.72	*	36.23	40.35	43.34	44.38	41.08	0.82	*
Qwen3-4B-Thinking	46.25	0.89	*	38.45	45.23	46.12	47.98	44.45	0.96	*
Chunk-Based Approaches										
RAG (BM25)	13.53	0.89	*	23.53	26.74	29.75	32.12	28.04	1.45	*
MemWalker (Chen et al., 2023b)	53.45	2.24	37.42	44.31	49.54	53.64	56.98	51.12	2.82	48.98
ReadAgent (Lee et al., 2024)	59.38	2.03	30.45	48.46	59.34	57.24	61.45	56.62	2.45	44.90
HUMAN-LLM COLLABORATION										
LongAF (Ours)	**93.64**	2.56	106.72	**92.74**	**93.98**	**92.22**	**94.86**	**93.45**	2.68	106.76
w/o Human Check	68.24	1.89	46.64	69.86	72.62	73.46	76.42	73.09	2.06	54.90

* indicates that we employ locally deployed open-source models, which incur no API-based inference cost.

Table 4: Ablation on multi-path retrieval (R@20 on LongBench). Best is **bolded**, second-best is underlined.

	HQA	MSQ	NQA	Qsp	2WMQ	Avg.
LongAF (Full)	**97.75**	**91.96**	**91.24**	**89.64**	**98.81**	**93.88**
w/o Query Exp.	58.55	54.42	57.35	59.13	53.42	56.57
w/o Evid. Rank	74.45	68.34	65.34	64.13	68.34	68.12
Single Retriever (w/o Multi-path)						
w/ BM25	85.25	78.86	77.42	78.45	83.08	80.61
w/ BGE	82.00	72.40	74.56	73.45	81.83	76.84
w/ LLM	74.33	55.65	53.56	58.63	76.58	63.75

HQA: HotpotQA, MSQ: MuSiQue, NQA: NarrativeQA, Qsp: Qasper, 2WMQ: 2WikiMQA.

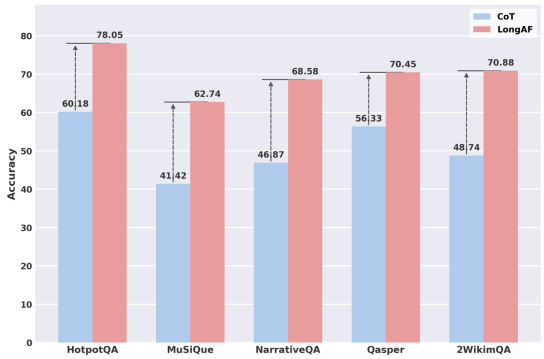


Figure 3: Reasoning-chain accuracy of MCTS-based LongAF vs. vanilla CoT across LongBench.

5.2.2 Effect of Reasoning Chain Generator

We compare our evidence-based MCTS against a vanilla CoT baseline that generates chains directly from retrieved evidence. As shown in Figure 3, MCTS yields a 19.43% improvement. Unlike vanilla CoT, which often fails at evidence selection in multi-hop settings, MCTS systematically explores the search space to construct evidence-grounded reasoning paths, substantially improving the reliability of the generated chains.

5.3 Analysis of Generalization Capabilities

To assess the adaptability of LongAF, we integrate it with a range of open-source LLMs. As shown in Table 5, LongAF consistently improves performance from Qwen2.5-7B to Qwen2.5-72B, surpassing both the vanilla LLM and the ReadAgent baseline on every model. When paired with DeepSeek-R1, LongAF reaches 76.34% accuracy, approaching the proprietary GPT-4o-mini result (78.05%). This indicates that LongAF enables cost-effective open-source models to deliver competitive annotation quality.

5.4 Hyper-parameter Analysis

We analyze the sensitivity of k in multi-path retrieval and the rollout count in MCTS. Figure 4a shows that k=20 balances recall and efficiency, as gains diminish beyond this threshold. Figure 4b

Table 5: Generalization of LongAF across LLMs on LongBench.

Model	Method	HotpotQA	MuSiQue	NarrativeQA
GPT-4o-mini	Vanilla	64.81	40.42	33.29
	ReadAgent	68.94	56.32	45.34
	LongAF	**78.05**	**62.74**	**68.58**
DeepSeek-R1	Vanilla	69.99	62.66	21.51
	ReadAgent	73.46	68.43	34.23
	LongAF	**76.34**	**72.45**	**65.34**
Qwen2.5-7B	Vanilla	56.24	29.53	33.33
	ReadAgent	60.23	38.12	40.15
	LongAF	**63.18**	**45.35**	**46.76**
Qwen2.5-72B	Vanilla	61.42	49.53	35.42
	ReadAgent	65.23	52.42	44.23
	LongAF	**70.52**	**58.28**	**59.64**

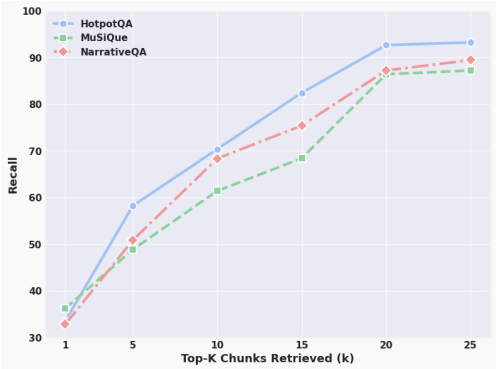
shows that 30 rollouts maximize accuracy while keeping token usage and latency manageable.

5.5 Error Analysis

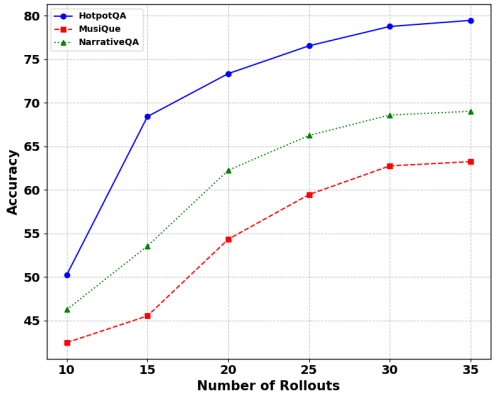
To diagnose the residual cognitive burden on annotators, we conduct a fine-grained taxonomy of LongAF-generated reasoning chains (Figure 5). (*Overall distribution*) 70.88% of chains are *fully correct* and can be verified without manual modification, demonstrating the robustness of the MCTS engine; (*Soft vs. hard failures*) the most frequent error, *missing evidence in pool* (17.43%), is a "soft failure"the missing snippet remains in the Stage-1 candidate pool, allowing annotators to repair the chain through low-effort selection rather than rereading the source, while truly "hard" failures with gold evidence absent from the pool are confined to 6.66%, implying that LongAF eliminates exhaustive document reading in over 93% of instances; (*Grounding*) hallucinated evidence is virtually absent (0.38%) and logical errors are limited (4.65%), confirming that strict citation requirements effectively constrain the generation space. Detailed cases are provided in Appendix A.3.

6 Conclusion

We presented **LongAF**, which reformulates long-document QA annotation as *localized verification over a compact, evidence-grounded substrate* rather than whole-document reading. By coupling multi-path retrieval, MCTS-based reasoning-chain generation, and an interactive verification interface, LongAF shifts the annotator's role from open-ended reasoning to step-wise verification, so that cognitive cost scales with evidence rather than document length. Across seven long-document QA benchmarks, LongAF matches or slightly exceeds fully manual annotation quality at **1/8** the time and **1/4** the cost; we release the toolkit and resulting evidence chains as a foundation for long-context data construction.



(a) Recall vs. chunk count k.



(b) Accuracy vs. rollout count.

Figure 4: Hyper-parameter analysis on HotpotQA, MuSiQue, and NarrativeQA.

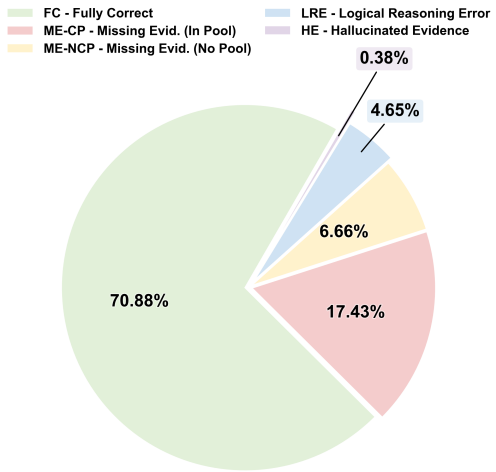


Figure 5: Distribution of LongAF reasoning chains across five categories: fully correct (FC), missing evidence in pool (ME-CP), missing evidence outside pool (ME-NCP), logical reasoning errors (LRE), and hallucinated evidence (HE).

Limitation

While our proposed LLM proxy framework demonstrates strong performance in evidence recall, it also presents notable limitations, including high computational cost and slow interaction speed due to the need for multiple model queries per interaction. Future work will focus on addressing these limitations by optimizing efficiency.

Ethical Considerations

Our research utilizes publicly available datasets, which we have verified to be free of personally identifiable information, thus ensuring full compliance with privacy standards. A key ethical consideration stems from our reliance on large language models for text generation. As with many applications of generative AI, this methodology carries an inherent risk: the potential to produce content that is harmful, biased, or reflects societal stereotypes. We acknowledge these challenges and recognize our responsibility to transparently report any such limitations observed during our study.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Vivek Bagaria, Tavor Z Baharav, Govinda M Kamath, and David N Tse. 2021. Bandit-based monte carlo optimization for nearest neighbors. *IEEE Journal on Selected Areas in Information Theory*, 2(2):599–610.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, and 1 others. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, and 1 others. 2024. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*.

Sabyasachee Baruah and Shrikanth Narayanan. 2024. Chatter: A character attribution dataset for narrative understanding. *arXiv preprint arXiv:2411.05227*.

Guanzheng Chen, Xin Li, Zaiqiao Meng, Shangsong Liang, and Lidong Bing. 2023a. Clex: Continuous length extrapolation for large language models. *arXiv preprint arXiv:2310.16450*.

Howard Chen, Ramakanth Pasunuru, Jason Weston, and Asli Celikyilmaz. 2023b. Walking down the memory maze: Beyond context limit through interactive reading. *arXiv preprint arXiv:2310.05029*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. *arXiv preprint arXiv:2105.03011*.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Shafiq Joty, Boyang Li, and Lidong Bing. 2022. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. 2024. How to train long-context language models (effectively). *arXiv preprint arXiv:2410.02660*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.

Jerry Huang. 2024. How well can a long sequence model model long sequences? comparing architechtural inductive biases on long-context abilities. *arXiv preprint arXiv:2407.08112*.

Alon Jacovi, Yonatan Bitton, Bernd Bohnet, Jonathan Herzig, Or Honovich, Michael Tseng, Michael Collins, Roee Aharoni, and Mor Geva. 2024. A chain-of-thought is as strong as its weakest link: A benchmark for verifiers of reasoning chains. *arXiv preprint arXiv:2402.00559*.

Ziyan Jiang, Xueguang Ma, and Wenhu Chen. 2024. Longrag: Enhancing retrieval-augmented generation with long-context llms. *arXiv preprint arXiv:2406.15319*.

Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. 2024. Meganno+: A human-llm collaborative annotation system. *arXiv preprint arXiv:2402.18050*.

Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Levente Kocsis and Csaba Szepesvári. 2006. Bandit based monte-carlo planning. In *European conference on machine learning*, pages 282–293. Springer.

Kuang-Huei Lee, Xinyun Chen, Hiroki Furuta, John Canny, and Ian Fischer. 2024. A human-inspired reading agent with gist memory of very long contexts. *arXiv preprint arXiv:2402.09727*.

Timo Lenzner, Lars Kaczmirek, and Alwine Lenzner. 2010. Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied cognitive psychology*, 24(7):1003–1020.

Huayang Li, Pat Verga, Priyanka Sen, Bowen Yang, Vijay Viswanathan, Patrick Lewis, Taro Watanabe, and Yixuan Su. 2024. Alr ^{2}: A retrieve-then-reason framework for long-context question answering. *arXiv preprint arXiv:2410.03227*.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy F Chen, Zhengyuan Liu, and Diyi Yang. 2023. Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation. *arXiv preprint arXiv:2310.15638*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

John Mern, Anil Yildiz, Zachary Sunberg, Tapan Mukerji, and Mykel J Kochenderfer. 2021. Bayesian optimized monte carlo planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11880–11887.

Benjamin R Mitchell, Marion C Cohen, and Stanley Cohen. 2021. Dealing with multi-dimensional data and the burden of annotation: easing the burden of annotation. *The American Journal of Pathology*, 191(10):1709–1716.

Rock Yuren Pang, KJ Feng, Shangbin Feng, Chu Li, Weijia Shi, Yulia Tsvetkov, Jeffrey Heer, and Katharina Reinecke. 2025. Interactive reasoning: Visualizing and controlling chain-of-thought reasoning in large language models. *arXiv preprint arXiv:2506.23678*.

Zhen Qin, Yiran Zhong, and Hui Deng. 2024. Exploring transformer extrapolation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18897–18905.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, and 1 others. 2016. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.

Rion Snow, Brendan Oconnor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Cheng Tan, Dongxin Lyu, Siyuan Li, Zhangyang Gao, Jingxuan Wei, Siqi Ma, Zicheng Liu, and Stan Z Li. 2024a. Peer review as a multi-turn and long-context dialogue with role-based interactions. *arXiv preprint arXiv:2406.05688*.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024b. Large language models for data annotation and synthesis: A survey. *arXiv preprint arXiv:2402.13446*.

Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Yuan Tian, Daniel Lee, Fei Wu, Tung Mai, Kun Qian, Siddhartha Sahai, Tianyi Zhang, and Yunyao Li. 2025. Text-to-sql domain adaptation via human-llm collaborative data annotation. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 1398–1425.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Cunxiang Wang, Ruoxi Ning, Boqi Pan, Tonghui Wu, Qipeng Guo, Cheng Deng, Guangsheng Bao, Qian Wang, and Yue Zhang. 2024a. Novelqa: A benchmark for long-range novel question answering. *arXiv e-prints*, pages arXiv–2403.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024b. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yuxi Xie, Anirudh Goyal, Wenyue Zheng, Min-Yen Kan, Timothy P Lillicrap, Kenji Kawaguchi, and Michael Shieh. 2024. Monte carlo tree search boosts reasoning via iterative preference learning. *arXiv preprint arXiv:2405.00451*.

Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2024a. Search-in-the-chain: Interactively enhancing large language models with search for knowledge-intensive tasks. In *Proceedings of the ACM Web Conference 2024*, pages 1362–1373.

Zhe Xu, Jiasheng Ye, Xiangyang Liu, Tianxiang Sun, Xiaoran Liu, Qipeng Guo, Linlin Li, Qun Liu, Xuanjing Huang, and Xipeng Qiu. 2024b. Detectiveqa: Evaluating long-context reasoning on detective novels. *arXiv preprint arXiv:2409.02465*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Chen Yin and Zixuan Zhang. 2024. A study of sentence similarity based on the all-minilm-l6-v2 model with same semantics, different structure after fine tuning. In *2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, pages 677–684. Atlantis Press.

Junhao Zhang, Richong Zhang, Fanshuang Kong, Ziyang Miao, Yanhan Ye, and Yaowei Zheng. 2025. Lost-in-the-middle in long-text generation: Synthetic dataset, evaluation framework, and mitigation. *arXiv preprint arXiv:2503.06868*.

Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, Zhangyang Wang, and 1 others. 2024. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. *Advances in Neural Information Processing Systems*, 37:60755–60775.

Qingfei Zhao, Ruobing Wang, Yukuo Cen, Daren Zha, Shicheng Tan, Yuxiao Dong, and Jie Tang. 2024. Longrag: A dual-perspective retrieval-augmented generation paradigm for long-context question answering. *arXiv preprint arXiv:2410.18050*.

Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. 2024. Fanoutqa: A multi-hop, multi-document question answering benchmark for large language models. *arXiv preprint arXiv:2402.14116*.

Yufan Zhuang, Xiaodong Yu, Jialian Wu, Ximeng Sun, Ze Wang, Jiang Liu, Yusheng Su, Jingbo Shang, Zicheng Liu, and Emad Barsoum. 2025. Self-taught agentic long context understanding. *arXiv preprint arXiv:2502.15920*.

A Appendixes

A.1 Annotation Details

To assess the diversity and expertise of our annotators, we collect information on their majority and degree, ensuring the protection of personal privacy. All annotators are either current graduate students or have completed their graduate studies. The fields are diverse, with the most common being Computer Science (53.33%), followed by Law (23.33%), Economics (13.33%), and Sociology (10.00%).

A.2 Annotation Platform

The platform is architected to streamline the human-in-the-loop verification process by decoupling evidence retrieval from logical reasoning. The interface comprises four functional modules (see Figure 6):

1. Global Context and Navigation: Located at the top, this module establishes situational awareness. It displays the Dataset Metadata and provides a Document Viewer containing the raw, ultra-long context. This ensures that annotators maintain access to the ground-truth text for ad-hoc verification while navigating the dataset via sequential controls.

2. Interactive Multi-Path Evidence Retrieval (Left Panel): This component visualizes the Stage-1 retrieval process.

   - Query Expansion & Decomposition: To address the semantic gap, the system displays LLM-generated expanded keywords and sub-question decompositions (e.g., splitting a complex multi-hop query into atomic sub-queries).
   - Selectable Evidence Pool: A critical interactive feature where evidence candidates retrieved via hybrid methods (BM25, BGE, LLM) are presented with their relevance scores and source provenance. Annotators can efficiently filter noise by toggling checkboxes ("Select ALL Relevant"), refining the evidence pool for the downstream reasoning stage without reading the full document.

3. Structured Reasoning Chain Visualization (Top-Right Panel): This module provides interpretability for the MCTS-based reasoning engine.

    - Hierarchical Chain Display: The reasoning path is rendered as a structured JSON object or tree, breaking down the decision process into discrete Steps.
    - Traceable Logic Nodes: Each step explicitly links a sub_query (reasoning goal), sub_answer (intermediate conclusion), and evidence_used (citation ID). This transparency allows annotators to rapidly validate the logical soundness of the model's deduction trajectory.

4. Human Verification and Gold Annotation (Bottom-Right Panel): The final decision module.

    - Gold Data Management: Annotators synthesize the Gold Answer and curate the final Gold Evidence set. Evidence Augmentation: The interface supports adding missing evidence directly from the candidate pool or manual entry, ensuring the completeness of the annotation before saving.

A.3 Case Study

To rigorously diagnose the framework's limitations and quantify the residual cognitive burden on annotators, we propose a fine-grained taxonomy of the generated reasoning chains. Let $\mathcal{D}$ denote the global document corpus, $\mathcal{C} \subset \mathcal{D}$ be the Stage-1 candidate pool retrieved by the coarse-grained retriever, and $\mathcal{N} \subset \mathcal{C}$ be the refined node pool used by the MCTS engine. We define E_{gold} as the set of necessary evidence required to answer a specific query correctly.

- **Fully Correct (Ideal Success):** This category represents the optimal scenario where the system successfully retrieves all necessary evidence ($E_{gold} \subseteq \mathcal{N}$) and synthesizes a logically sound reasoning path R to derive the correct answer. These instances substantiate the robustness of the MCTS engine, allowing for *Direct Verification* without manual modification. (see Figure 7).
- **Missing Evidence in Pool (Soft Failure):** A retrieval anomaly characterized by the presence of gold evidence in the coarse-grained candidates but its exclusion from the reasoning graph, formalized as $\exists e \in E_{gold}$ such that $e \in \mathcal{C} \setminus \mathcal{N}$. This often results from aggressive filtering thresholds. We classify this as a "Soft Failure" because the evidence remains retrievable; annotators can repair the chain via *Low-Effort Selection* from $\mathcal{C}$ rather than re-reading source texts (see Figure 8).
- **Missing Evidence in Recall (Hard Failure):** A fundamental retrieval deficit where critical evidence is absent from the system's purview ($e \notin \mathcal{C}$). This "Hard Failure" represents the sole scenario necessitating *Exhaustive Document Reading*. The low prevalence of this error implies that the framework effectively obviates the need for full-text review (see Figure 9).
- **Hallucination (Grounding Failure):** A rare failure mode where the generated chain contains claims c unsupported by the retrieved context, formally $P(c|\mathcal{N}) \approx 0$. The negligible occurrence rate confirms that our strict citation constraints effectively restrict the generation space, virtually ensuring that outputs are grounded in the provided evidence (see Figure 10).
- **Logical Error (Reasoning Misalignment):** Instances where the retrieval is accurate ($E_{gold} \subseteq \mathcal{N}$), yet the generation module G fails to synthesize the information correctly, resulting in internal contradictions. Unlike retrieval errors, these cases require Correction, where annotators intervene to realign the reasoning process with the correctly retrieved facts (see Figure 11).

A.4 Prompt Template

This section details the prompt engineering employed in two long-context reasoning systems: ReadAgent (Lee et al., 2024) and LongAF. For ReadAgent, we delineate the structured prompt instructions guiding its core functionalities: gist collection, gist querying, and answer inference. For LongAF, we present the prompt content for LLM-based retrieval within its multi-path system, along

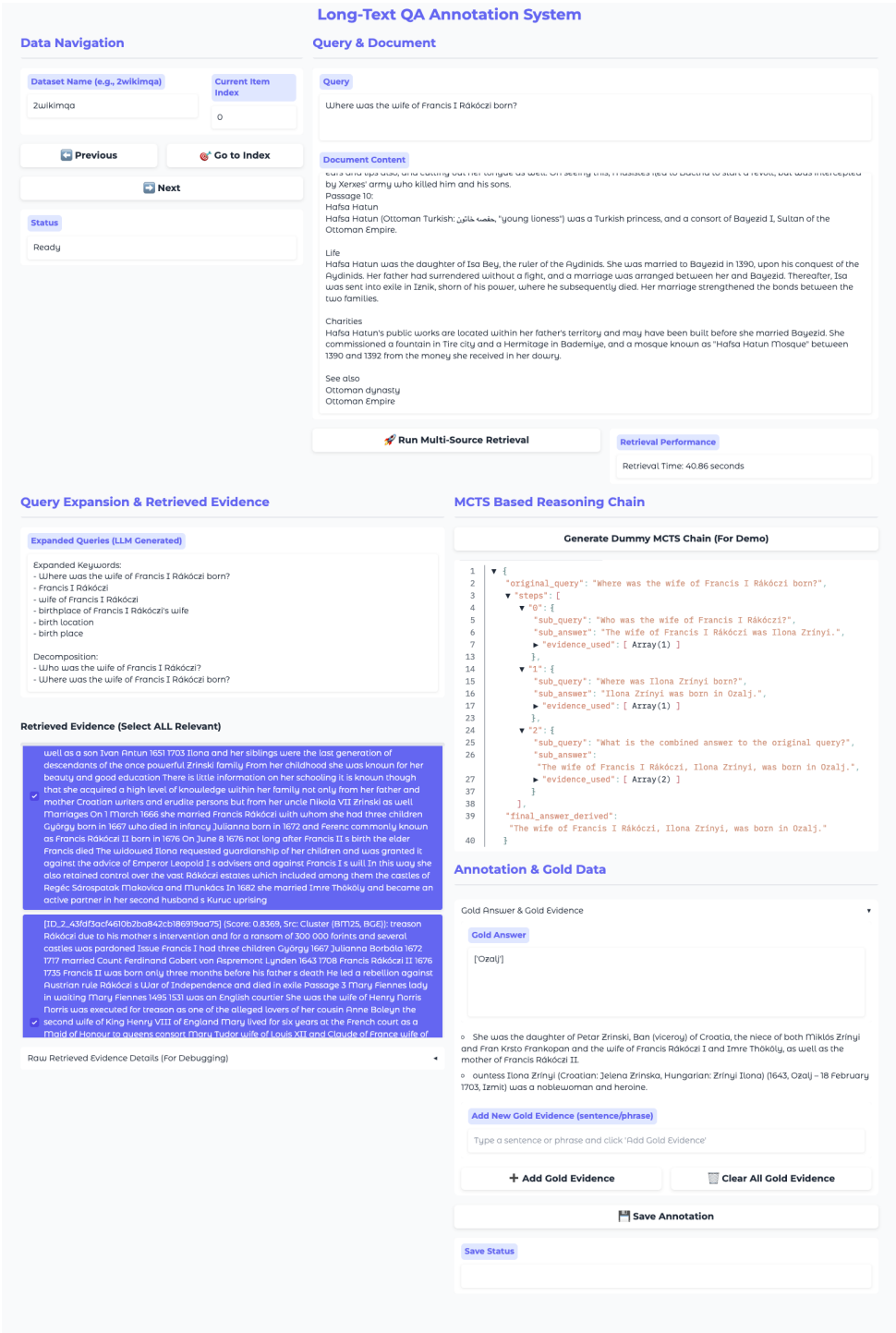


Figure 6: The LongAF Annotation Interface. The Left Panel streamlines evidence retrieval by displaying expanded queries and an interactive candidate pool, enabling rapid filtering without exhaustive reading. The Right Panel visualizes the MCTS-driven reasoning chain as traceable triplets (sq_i, e_i, sa_i) and provides a dedicated workspace for validating logic, refining gold answers, and augmenting evidence.

Multi Path Evidence Recall

Q: Do both Beauty And The Bad Man and Wild Child (Film) films have the directors from the same country?

[ID=1 Score=1.0 Source=BM25] …as China Bandit then War Lord then The Adventures of Fang Filming took place in February 1937…
[ID=2 Score=0.9727 Source=BM25, BGE] …tutor He is currently a Clinical Tutor at the European School of Osteopathy…
[ID=3 Score=0.8539 Source=BM25, BGE, LLM] …Passage 9 Beauty and the Bad Man Beauty and the Bad Man is a 1925 American silent Western film directed by William Worthington and starring Mabel Ballin…
[ID=5 Score=0.7703 Source=BM25, BGE, LLM] …Passage 3 Wild Child film Wild Child is a 2008 teen comedy film directed by Nick Moore and starring Emma Roberts Natasha Richardson Shirley Henderson Alex Pettyfer and Aidan Quinn with Georgia King Kimberley Nixon Juno Temple Linzey Cocker and Sophie Wu…
[ID=4 Score=0.8243 Source=BM25, BGE, LLM] …Nick Moore film director Nick Moore is a British film director and editor He directed the 2011 British Movie Horrid Henry The Movie…
[ID=6 Score=0.7138 Source=BM25] …William Worthington actor William Worthington April 8 1872 in Troy New York April 9 1941 in Beverly Hills California was an American silent film actor and director…
…

Evidence Nood Pool

sub_query: What is the director of Beauty And The Bad Man?
sub_answer: The director of Beauty And The Bad Man, William Worthington.
evidence_id: 3 ✅

sub_query: What is the nationality or country of origin of the director of Wild Child (Film)?
sub_answer: The director of Wild Child (Film), Nick Moore, is British.
evidence_id: 4 ✅

sub_query: What is the director of Wild Child (Film)?
sub_answer: The director of Wild Child (Film), Nick Moore, is British.
evidence_id: 5 ✅

sub_query: What is the nationality or country of origin of the director of Beauty And The Bad Man?
sub_answer: The director of Wild Child (Film), William Worthington, in New York.
evidence_id: 6 ✅

sub_query: What is the the director of Bad Man?
sub_answer: The director of Bad Man, Mabel Ballin.
evidence_id: 9
…

MCTS-based Reasoning Chain Generation

sub_query: What is the director of Wild Child (Film)?
sub_answer: The director of Wild Child (Film), Nick Moore, is British.
evidence_id: 5

sub_query: What is the nationality or country of origin of the director of Wild Child (Film)?
sub_answer: The director of Wild Child (Film), Nick Moore, is British.
evidence_id: 4

sub_query: What is the director of Beauty And The Bad Man?
sub_answer: The director of Beauty And The Bad Man, William Worthington.
evidence_id: 3

sub_query: What is the nationality or country of origin of the director of Beauty And The Bad Man?
sub_answer: The director of Wild Child (Film), William Worthington, in New York.
evidence_id: 6

final answer: No, The director of Wild Child, Nick Moore, is British. The director of Beauty And The Bad Man, in New York.

Figure 7: Case Study: **Fully Correct**. In these instances, the framework effectively retrieves all necessary gold evidence and synthesizes the correct answer, permitting direct verification by annotators and eliminating the need for manual modification.

Multi Path Evidence Recall

Q: Who is the mother-in-law of Louise Of Mecklenburg-Güstrow?

[ID=1 Score= 0.9127 Source=BM25] … VI was anxious to secure the throne of England He sent his ambassadors to his wife s family and allies to discuss his right to Elizabeth…
[ID=2 Score=0.9047 Source=BM25]… 26 February 1696 Zörbig 26 March 1696 Stillborn twin sons 1707 Hedwig died in Zörbig aged 69 She was buried …
[ID=3 Score=0.8500 Source=BM25, BGE] … Passage 4 Frederick IV of Denmark Frederick IV Danish Frederik 11 October 1671 12 October 1730 was King of..✅ Denmark and Norway from 1699 until his death Frederick was the son of Christian V of Denmark Norway and his wife Charlotte Amalie of Hesse Kassel…✅
[ID=4 Score=0.8243 Source= BM25, BGE, LLM] … Passage 8 Louise of Mecklenburg Güstrow Louise of Mecklenburg Güstrow 28 August 1667 15 March 1721 was Queen of Denmark and Norway as the first spouse of King Frederick IV of Denmark In 1708 09 she…
[ID=5 Score=0.8012 Source=BM25, BGE, LLM] …Nick Moore film director Nick Moore is a British film director and editor He directed the 2011 British Movie Horrid Henry The Movie…
[ID=6 Score=0.7138 Source=BM25] … Passage 1 Hedwig of Mecklenburg Güstrow Hedwig of Mecklenburg Güstrow Hedwig Eleonore 12 January 1666 9 …
…

Evidence Nood Pool

sub_query: What is the director of Wild Child (Film)?
sub_answer: The director of Wild Child (Film), Nick Moore, is British.
evidence_id: 2

sub_query: What is the nationality or country of origin of the director of Beauty And The Bad Man?
sub_answer: The director of Wild Child (Film), William Worthington, in New York.
evidence_id: 1

sub_query: What is the the director of Bad Man?
sub_answer: The director of Bad Man, Mabel Ballin.
evidence_id: 9
…

MCTS-based Reasoning Chain Generation

sub_query: Who did Louise Of Mecklenburg-Güstrow marry?
sub_answer: Louise Of Mecklenburg-Güstrow married Duke Christian Ulrich I of Württemberg Oels.
evidence_id: 2

sub_query: Who is the mother of Louise Of Mecklenburg-Güstrow's spouse?

sub_answer: Elizabeth of Denmark is the mother of Duke Christian Ulrich I of Württemberg Oels, who is Louise Of Mecklenburg-Güstrow's spouse.
evidence_id: 1

final answer: The mother-in-law of Louise of Mecklenburg-Güstrow is Elizabeth of Denmark. This is because Louise of Mecklenburg-Güstrow married Duke Christian Ulrich I of Württemberg Oels, whose mother was Elizabeth of Denmark.

Figure 8: Case Study: **Missing Evidence in Pool (Soft Failure)**. An instance where the MCTS engine fails to locate necessary evidence in the Node Pool, interrupting the chain generation. However, this constitutes a "soft failure" as the gold evidence remains retrievable in the Stage-1 candidate pool. Annotators can repair such chains via low-effort selection rather than re-reading source texts, maintaining high efficiency.

Multi Path Evidence Recall

Q: Don Holmes played for the football club that is based in what eastern suburb of Perth?

[ID=1 Score=1.0 Source=BM25] …as China Bandit then War Lord then The Adventures of Fang Filming took place in February 1937…
[ID=2 Score=0.9727 Source=BM25, BGE] …tutor He is currently a Clinical Tutor at the European School of Osteopathy…
[ID=3 Score=0.8539 Source=BM25, BGE, LLM] …Passage 9 Beauty and the Bad Man Beauty and the Bad Man is a 1925 American silent Western film directed by William Worthington and starring Mabel Ballin…
[ID=5 Score=0.7703 Source=BM25, BGE, LLM] …Passage 3 Wild Child film Wild Child is a 2008 teen comedy film directed by Nick Moore and starring Emma Roberts Natasha Richardson Shirley Henderson Alex Pettyfer and Aidan Quinn with Georgia King Kimberley Nixon Juno Temple Linzey Cocker and Sophie Wu…
[ID=4 Score=0.8243 Source=BM25, BGE, LLM] …Nick Moore film director Nick Moore is a British film director and editor He directed the 2011 British Movie Horrid Henry The Movie…
[ID=6 Score=0.7138 Source=BM25] …William Worthington actor William Worthington April 8 1872 in Troy New York April 9 1941 in Beverly Hills California was an American silent film actor and director…
…

Evidence Nood Pool

sub_query: Which football club did Don Holmes play for?
sub_answer: Don Holmes played for East Perth Football Club..
evidence_id: 2

sub_query: In which eastern suburb of Perth is that football club based?
sub_answer: Leederville.
evidence_id: 4
…

MCTS-based Reasoning Chain Generation

sub_query: Which football club did Don Holmes play for?
sub_answer: Don Holmes played for East Perth Football Club..
evidence_id: 2

No gold evidence

final answer:
Don Holmes played for the East Perth Football Club, which is based in Leedervill e, an eastern suburb of Perth.

Figure 9: Case Study: **Missing Evidence in Recall (Hard Failure)**. A minority case where the gold evidence is entirely absent from both the Node Pool and the Stage-1 candidate pool. These "hard failures" represent the only scenarios necessitating exhaustive document reading by humans.

Multi Path Evidence Recall

Q: What is the capital of the county that shares a border with the county where KRSU is licensed to broadcast?

[ID=1 Score=1.0 Source=BM25] …February 4 1861 in Montgomery Alabama and remained there unti…
[ID=2 Score=0.8955 Source=BM25, LLM] … Passage 6 Chol Kiri District Chol Kiri Khmer is a district srok …
[ID=3 Score=0.8068 Source=BM25, BGE, LLM] … Passage 1 Baranya County Baranya Hungarian Baranya vármegye pronounced …
[ID=4 Score=0.7910 Source=BM25, BGE] …
On July 2 1923 President Warren G Harding commented in a speech or as a historic marker tells it reporters noted that the little town of Meacham Oregon was th e nation s capital …
[ID=5 Score=0.7761 Source=BM25, BGE, LLM] … while 22 of the population lives in villages that have less than 1000 inhabitants Cities …
…

Evidence Nood Pool

sub_query:
Which counties share a bord er with the county where K RSU is licensed to broadcas t?
sub_answer:
Chol Kiri district shares a b order with Kampong Thom and Kampong Cham provin ces.
evidence_id: 3
…

MCTS-based Reasoning Chain Generation

sub_query: Which counties share a border with the county where KRSU is licensed to broadcast?
sub_answer: Chol Kiri district shares a border with Kampong Thom and Kampong Cham provinces.
evidence_id: 3

Cannot be inferred

final answer:
KRSU is licensed to broadcast in Chol Kiri district, which is in Kampong Chhnang province. Chol Kiri district shares a border with Ka mpong Thom and Kampong Cham provinces. Therefore, the counties (provinces) that share a border with the county where KRSU is licensed are Kampong Thom and Kampong Cham.\n\nThe original question asks for the capital of the county that shares a border w ith the county where KRSU is licensed. Since Kampong Thom and Kampong Cham both share a border with Kampong Chhnang, we need the capitals of these provinces.\n\n- The capital of Kampong Thom province is Stung Saen.\n- The capital of Kampong Cham province is Kampong Cham city.\n\nFinal Answer: The capitals of the counties that share a border wit h the county where KRSU is licensed to broadcast are Stung Saen (Kampong Thom) and Kampong Cham city (Kampong Cham).

Figure 10: Case Study: **Hallucination**. An extremely rare failure mode where the generated reasoning chain includes content or claims not supported by the retrieved evidence.

Multi Path Evidence Recall

Q: Which composer was wrote his music most recently, Michael Tippett or Luigi Cherubini?

[ID=1 Score=0.9890 Source=BM25] …
The Piano Concerto 1955 was declared unplayable by its scheduled soloist Julius Katchen who had to be replaced before the première by Louis Kentner The De nnis …
[ID=2 Score=0.9273 Source=BM25, LLM] … The music staff at Morley was augmented by the recruitment of refugee musicians from Europe …
[ID=3 Score=0.8034 Source=BM25, BGE, LLM] … BBC Symphony Orchestra under Boult in a live broadcast from the Royal Festival Hall on 5 February 1958 …
[ID=5 Score=0.7013 Source=BM25, BGE, LLM] … Nancy 1878 Stratonice introduction to the edition of Hoffman and Méhul …
[ID=4 Score=0.8243 Source=BM25, BGE, LLM] …Nick Moore film director Nick Moore is a British film director and editor He directed the 2011 British Movie Horrid Henry The Movie…
…

Evidence Nood Pool

sub_query: When did Luigi Cherubini die?
sub_answer: Luigi Cherubini died on 15 March 1842.
evidence_id: 3

sub_query: When did Michael Kemp Tippett born?
sub_answer:
Sir Michael Kemp Tippett (2 January 1905 – 8 January 1998) .
evidence_id: 2
…

MCTS-based Reasoning Chain Generation

sub_query: When did Luigi Cherubini die?
sub_answer: Luigi Cherubini died on 15 March 1842.
evidence_id: 3

sub_query: When did Michael Kemp Tippett born?
sub_answer: Sir Michael Kemp Tippett (2 January 1905 – 8 January 1998) .
evidence_id: 2

Logical Error

final answer:
Luigi Cherubini died on 15 March 1842. Since Michael Tippett lived and composed in the 20th century (1905–1998). Therefore, Luigi Cherubini was the composer who wrote his music most recently.

Figure 11: Case Study: **Logical Error**. An error case where the framework successfully retrieves the correct evidence but fails to synthesize it logically. In these instances, the reasoning chain contains invalid deduction steps or contradictions, requiring human intervention to correct the logic despite the presence of accurate source information.

with the prompts associated with the MCTS algorithm used for refining retrieved evidence, as illustrated in Fig 12, 13, 14.

Prompt for Method: ReadAgent

```
### Example Pagination :

You are given a passage that is taken from a larger text (article, book, ...) and
some numbered labels between the paragraphs in the passage.
Numbered labels are in angle brackets. For example, if the label number is 19, it shows as 19
in text.
Please choose a label where it is natural to break reading.
The label can be a scene transition, the end of a dialogue, the end of an argument, a narrative
transition, etc.
Please answer with the break point label and explain.
For example, if 57 is a good point to break, answer with Break point: 57
Because ...
Passage:
...
[PARAGRAPH 5] TEXT
[PARAGRAPH 6] TEXT
[PARAGRAPH 7] TEXT
...

### Memory Gisting :

Please shorten the following passage.
Just give me a shortened version. DO NOT explain your reason.

Passage:
[PAGE TEXT]

### Example Parallel Lookup :

The following text is what you remember from reading an article and a multiple-choice question
related to it. You may read 1 to 5 pages of the article again to refresh your memory to prepare
yourself for the question. Please respond with which page(s) you would like to read.

For example, if you only need to read Page 8, respond with I want to look up Page [8] to ...;
if you would like to read Page 7 and 12, respond with I want to look up Page [7, 12] to ...; if
you would like to read Page 2, 3, 7, 15 and 18, respond with I want to look up Page [2, 3, 7,
15, 18] to ....
DO NOT select more pages if you dont need to.
You dont need to answer the question yet.
Text:
[GIST MEMORY]
Question:
[QUESTION]
Relpy with the passage number as your action
You MUST choose one summary number and you should reply with the following format:


Reasoning: ...
Action: 0 / 1 / 2, ...


### Navigation (leaf):

Read the text in triple quotes and answer a question:
Story background information: [WORKING MEMORY]
[QUERY]
[OPTIONS]
If the answer CAN be inferred from the text above, reply with action -2, and also provide your
reasoning, and the final answer.
You are ONLY allowed to reply with action -2 or -1
Your should reply with the following format:
Reasoning: ...
Action: -2 or -1
Answer: ...
```

Figure 12: ReadAgent: Prompts used for pagination, memory gisting, parallel lookup, and the navigation stage.

Prompt for LLM to Recall Evidence

Leader Instruction Generation:

You need to generate instructions for all team members. You can ask them to answer the Task objective based on their respective documents. If the question is too complex, you can consider breaking it down into some sub-questions.
Task Description:
{task_description}

Your output must follow the JSON format: {{"type": "instruction", "content": "your_instruction_content"}}

Reader Evidence Extraction:

You will receive a current source text, a previous source text, and a question. If you cannot answer the question based on the information in the current source text, please write "Cannot answer". If there is multiple relevant information in the article that can help answer the question, please clearly indicate them and put the original information in <ref></ref> before giving your answer, and generate a summary that includes both.
Your output must follow the JSON format:{{"type": "response", "content": "your_response_content", "summary": "your summary for the current source text and previous source text"}}

Leader Decision Making:

Here are the responses from all the members. Each member sees different segments of a document, and these segments do not intersect. The correct answer may appear in any one or several members' responses. Note that if a minority of members find information relevant to the question while the majority reply that the document does not contain information relevant to the question, you should pay attention to the replies from those members who found relevant information.
Member Response
{member_response}

Task Description
{task_description}

Task Objective
{task_objective}

Determine
Based on the above information, you need to determine whether the task objective can be solved. You have two options:
1. You need to reflect on whether the clues provided by the member are sufficient to answer the task objective, whether there are sub-questions that are not fully resolved, are missing important information, or contain contradictory answers. Please provide new instructions to complete their responses.
2. Otherwise, if the task objective can be solved, organize the relevant clues, give a clue chain that derives the result, provide the supporting fragments of the original text in <ref></ref>, and finally give an answer, using one phrase as much as possible. Do not provide any explanation. Your output must following the JSON format: {{"type": "instruction", "content": "your_instruction_content", "reason": "your reason for the instruction"}} or {{"type": "answer", "content": "your_answer_content", "reason": "your reason for the instruction"}}

Leader Answer Generation:

You have reached the maximum number of conversation rounds, please ignore the last instruction.
Generate Final Answer
Now, you need to generate final answer to achieve the task objective. You need to organize the relevant clues, give a clue chain that derives the result, provide the supporting fragments of the original text in <ref></ref>, and finally give an answer.
Your output must following the JSON format: {{"type": "answer", "content": "your_answer_content", "reason": "your reason for the instruction"}}

Figure 13: Prompts for the *leader* and *reader* roles used to recall evidence snippets in the Multi-Path Evidence Retrieval stage.

Prompt for Task to Reasoning Chain Generator

```
### Reward Generation:

You are an expert reasoning engine designed to evaluate the utility of new evidence in solving a
complex, multi-hop question.
Your task is to determine how a 'New Evidence' snippet advances or hinders the progress toward
answering the 'Original Question', given the 'Current Reasoning Path'.

Original Question: {{original_question}}

Current Reasoning Context:

{{context}}

Sub-queries Addressed So Far: {{solved_sub_queries_ref}}
Sub-queries Still Needed to Answer the Original Question: {{needed_sub_queries_ref}}

New Evidence to Evaluate: {{new_evidence_text}}

Instructions:
Carefully analyze the 'New Evidence' in relation to the 'Original Question' and the 'Current
Reasoning Context'. Provide your evaluation as a JSON object with the following fields:

1.'reward': A float representing the impact of this 'New Evidence' (from -1.0 to 1.0):
- '1.0': Direct Solution/Major Advance. The evidence directly and clearly helps solve one or more
unsolved sub-queries listed in 'Sub-queries Still Needed'. This is the primary goal.
- '0.5': Supporting/Redundant. The evidence confirms or provides additional support for an already
'Sub-query Addressed So Far'.
- '0.0': Irrelevant/No Value. The evidence is not relevant, provides no new useful information,
or does not fit any other category.
- '-1.0': Minor Contradiction/Misleading. The evidence subtly contradicts or misleads, but isn't
a direct negation of established facts.

2. 'newly_addressed_sub_query': A list of strings. If 'reward' is '1.0', state the specific
sub-query from 'Sub-queries Still Needed' that this evidence helps solve or significantly advance.
If it helps multiple, choose the most significant one. Otherwise, an empty string.

3. 'derived_answer_for_sub_query': A list of strings. If 'newly_addressed_sub_query' is not
empty, provide a concise answer to that sub-query based only on the 'New Evidence' and relevant
'Current Reasoning Context'. Otherwise, an empty string.

4. 'is_contradictory': A boolean ('true' or 'false'). Set to 'true' if the 'New Evidence'
contradicts prior information (even minor contradiction), 'false' otherwise. If 'true', the
'reward' must be negative.

5. 'updated_solved_sub_queries': A list of strings representing all sub-queries considered solved
after evaluating this new evidence. This list should be the union of 'solved_sub_queries_ref' and
any 'newly_addressed_sub_query' that is now considered fully solved.

6. 'updated_needed_sub_queries': A list of strings representing all sub-queries still needed to
be solved after evaluating this new evidence. This list should be 'needed_sub_queries_ref' minus
any 'newly_addressed_sub_query' that is now considered fully solved.

Ensure your output is a single, valid JSON object.
```

Figure 14: Prompt for scoring evidence relevance (reward signal) in the MCTS-based Reasoning Chain Generator.