

CROG: Coverage-aware Reasoning Optimization on Graphs for Multi-hop Question Answering

Junli Liang¹, Pengfei Zhou², Rui Lv¹, Wangqiu Zhou³, Qi Song^{1*}, Xiang-Yang Li¹

¹University of Science and Technology of China

²The Ohio State University

³Hefei University of Technology

jliang@mail.ustc.edu.cn, zhou.5317@osu.edu, lvrui2018@mail.ustc.edu.cn,

rafazwq@hfut.edu.cn, qisong09@ustc.edu.cn, xiangyangli@ustc.edu.cn

Abstract

Multi-hop question answering requires retrieving and integrating multiple pieces of evidence distributed across documents. Recent graph-based methods improve multi-hop reasoning through graph traversal over relational evidence structures. However, existing methods rely on independent path ranking during evidence organization, which often produces redundant and reasoning-incomplete evidence sets due to substantial structural overlap among selected graph paths. In this work, we argue that graph-based evidence organization may, to some extent, be formulated as a constrained evidence subgraph optimization problem. Based on this perspective, we propose CROG, a coverage-aware reasoning optimization framework that performs marginal-gain-based path selection to favor previously uncovered reasoning steps while suppressing redundant evidence, with node-budget and connectivity constraints to encourage compact and coherent evidence subgraphs. Experiments on multiple benchmarks demonstrate that CROG achieves superior reasoning performance while using fewer evidence units on average, and analysis based on exhaustive optimal search further validates the effectiveness of the greedy optimization strategy.

1 Introduction

Multi-hop question answering requires connecting multiple pieces of evidence dispersed across documents to answer complex queries (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022). This capability is essential for knowledge-intensive applications such as fact-checking, medical consultation, and legal analysis, where reliable decisions depend on integrating information from multiple sources (Gao et al., 2023; Peng et al., 2025; Zhang et al., 2025).

Traditional retrieval-augmented generation (RAG) methods typically retrieve evidence as independently ranked text chunks, which often

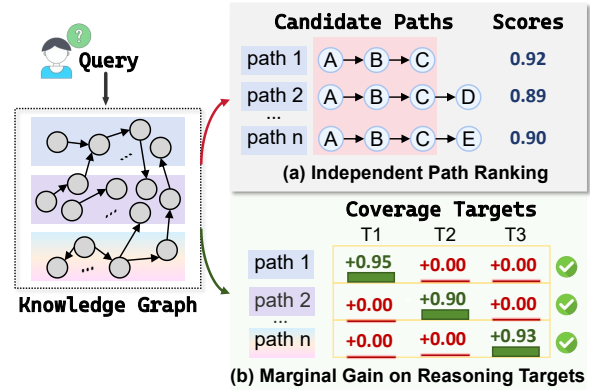


Figure 1: Independent path ranking repeatedly selects highly overlapping paths from locally relevant graph regions, resulting in redundant evidence and insufficient reasoning coverage. CROG instead organizes evidence according to marginal reasoning coverage gain, favoring paths that contribute additional reasoning coverage toward different reasoning targets.

struggle to capture the relational structure required for multi-hop reasoning (Lewis et al., 2020; Mavromatis and Karypis, 2024; Sarthi et al., 2024). To better support structured reasoning across multiple evidence hops, recent work has increasingly adopted graph-based reasoning frameworks that organize evidence as interconnected graph structures. Existing graph-based reasoning methods generally follow a two-stage paradigm: reasoning path exploration and evidence organization. In the first stage, methods progressively explore graph regions from query-relevant anchor nodes to construct candidate multi-hop reasoning paths through strategies such as breadth-first search (BFS) expansion, beam search, or iterative retrieval (Luo et al., 2024; Sun et al., 2024; Trivedi et al., 2023). In the second stage, the generated candidate paths are ranked according to their relevance to the query, and the top-ranked paths are selected as the final reasoning context for the large language model (LLM).

*Corresponding author

Despite substantial progress in reasoning path exploration, evidence organization still largely relies on independently ranking candidate paths by query relevance (Mou et al., 2021; Liang et al., 2026). Such an independent ranking paradigm fails to account for whether different reasoning requirements have already been sufficiently supported, often causing the final evidence set to overemphasize locally explored semantic regions. Once the exploration process reaches a highly relevant local semantic region, iterative exploration tends to repeatedly traverse neighboring nodes within that region, generating multiple highly overlapping candidate paths. As illustrated in Figure 1(a), paths such as $(A \rightarrow B \rightarrow C)$, $(A \rightarrow B \rightarrow C \rightarrow D)$, and $(A \rightarrow B \rightarrow C \rightarrow E)$ all receive high relevance scores because they originate from the same locally relevant region and share substantial structural overlap. As a result, the final evidence set may contain multiple highly similar paths with limited additional reasoning utility.

However, multi-hop question answering inherently requires satisfying multiple reasoning sub-goals, where different reasoning steps may rely on evidence distributed across different semantic regions of the graph. Under a limited evidence budget, repeatedly selecting locally overlapping paths often suppresses evidence supporting other reasoning targets, resulting in incomplete reasoning coverage (Lee et al., 2025; Chen et al., 2026; Tang and Yang, 2024). Effective graph-based reasoning therefore requires coverage-aware evidence optimization, where candidate paths are selected according to the coverage they contribute toward uncovered reasoning sub-goals, rather than being ranked independently by query relevance alone. As illustrated in Figure 1(b), effective evidence organization should favor paths that improve uncovered reasoning sub-goals (e.g., T_1 , T_2 , and T_3), instead of repeatedly selecting highly overlapping paths from the same local semantic region.

Motivated by this perspective, we reformulate graph-based multi-hop reasoning as a constrained evidence subgraph optimization problem that aims to maximize reasoning coverage through compact and complementary reasoning paths. Based on this formulation, we propose CROG, a Coverage-aware Reasoning Optimization method for multi-hop question answering over graphs. Specifically, CROG first performs *Reasoning Coverage Construction* by decomposing the question into reasoning sub-goals and treating them as explicit coverage

targets. It then conducts *Coverage-aware Reasoning Optimization* through marginal-gain-based path selection, favoring paths that improve previously uncovered reasoning sub-goals while suppressing semantically redundant evidence. CROG further incorporates node-budget and graph-connectivity constraints to encourage compact and structurally coherent evidence subgraphs for multi-hop reasoning (Feige, 1998; Nemhauser et al., 1978).

Our contributions are summarized as follows:

- We reformulate graph-based multi-hop reasoning as a constrained evidence subgraph optimization problem and propose CROG, a coverage-aware reasoning optimization framework that optimizes reasoning coverage through marginal-gain-based path selection.
- We introduce coverage-aware evidence organization with node-budget and graph-connectivity constraints, enabling CROG to construct compact, complementary, and structurally coherent evidence subgraphs for multi-hop reasoning. We further prove that the resulting optimization problem is NP-hard via reduction from Maximum Coverage, and establish approximation properties of the proposed greedy optimization strategy through monotone submodular coverage analysis.
- Extensive experiments on multiple multi-hop question answering benchmarks demonstrate that CROG achieves superior reasoning performance while using fewer evidence units. Additional analysis based on exhaustive optimal search further validates the effectiveness of the greedy optimization strategy.

2 Related Work

2.1 Retrieval-based Multi-hop Reasoning

Multi-hop question answering (QA) requires models to integrate evidence distributed across multiple documents and reasoning steps (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022). Early retrieval-based methods retrieve independently ranked passages as supporting evidence (Lewis et al., 2020), but often struggle when answering requires combining complementary evidence across multiple reasoning hops (Gao et al., 2023). Recent work has explored improved retrieval and reasoning strategies, including hierarchical summarization (Sarathi et al., 2024), interleaved retrieval and reasoning (Trivedi

et al., 2023), and set-level evidence refinement (Lee et al., 2025). However, these methods primarily optimize relevance among retrieved evidence units without explicitly modeling complementary reasoning coverage.

2.2 Graph-based Multi-hop Reasoning

Graph-based reasoning frameworks organize evidence as interconnected structures to support structured multi-hop reasoning (Peng et al., 2025; Zhang et al., 2025). One line of work operates over pre-built structured knowledge graphs, where methods such as RoG (Luo et al., 2024) and ToG (Sun et al., 2024) perform graph traversal and iterative reasoning to generate candidate reasoning paths. Another line constructs graphs directly from documents or sentences, including KGP (Wang et al., 2024), LightRAG (Guo et al., 2025), and Sent-Graph (Liang et al., 2026), which organize textual evidence as graph structures to support multi-hop retrieval and reasoning.

Despite improved path discovery and graph traversal strategies, final evidence organization in existing graph-based methods is still largely based on independently evaluating candidate paths according to query relevance or local reasoning utility, often producing overlapping evidence with limited complementary reasoning benefit.

2.3 Submodular Optimization for Information Selection

Submodular optimization has been widely applied to coverage-oriented selection problems due to its diminishing-return property (Nemhauser et al., 1978; Feige, 1998). Lin and Bilmes (Lin and Bilmes, 2011) apply submodular objectives to document summarization for balancing information coverage and redundancy reduction. MMR (Carbonell and Goldstein, 1998) encourages complementary selection in diversified retrieval via marginal relevance scoring. These studies demonstrate the effectiveness of coverage-aware selection under limited budgets.

However, existing submodular selection methods mainly operate on flat document or retrieval sets without modeling reasoning-step coverage or graph-structural constraints in multi-hop reasoning. Different from prior work, CROG formulates evidence organization as a constrained coverage maximization problem over graph reasoning paths, jointly modeling complementary reasoning coverage with node-budget and connectivity constraints.

3 The Proposed CROG Method

3.1 Overview

As illustrated in Figure 2, CROG consists of two key components: *Reasoning Coverage Construction* and *Coverage-aware Reasoning Optimization*. The first component constructs reasoning-aware coverage signals for multi-hop reasoning, while the second component organizes candidate reasoning paths into a compact and complementary evidence subgraph through coverage-aware optimization. The resulting evidence subgraph is finally provided to the LLM for answer generation.

3.2 Problem Formulation

Given a multi-hop question q and a knowledge graph $G = (V, E)$, where nodes in V represent textual or semantic units (e.g., entities, sentences, or passages depending on the graph construction), and edges in E represent semantic or relational connections between nodes.

Starting from query-relevant anchor nodes, graph traversal produces a candidate path set $\mathcal{P} = \{p_1, \dots, p_n\}$, where each path represents a connected reasoning trajectory in the graph.

The goal is to select a subset of candidate paths $\mathcal{S} \subseteq \mathcal{P}$, that maximizes complementary reasoning coverage while satisfying structural constraints. Formally, CROG optimizes:

$$\max_{\mathcal{S} \subseteq \mathcal{P}} \text{Coverage}(\mathcal{S}) \quad (1)$$

subject to:

$$\left| \bigcup_{p \in \mathcal{S}} \text{nodes}(p) \right| \leq B, \quad (2)$$

and

$$\text{Comp}(\mathcal{S}) \leq K, \quad (3)$$

where $\text{Coverage}(\mathcal{S})$ measures the collective reasoning coverage provided by the selected paths, formally defined in Section 3.3, B denotes the evidence budget, K denotes the maximum number of connected components, and $\text{Comp}(\cdot)$ computes the number of connected components in the induced evidence subgraph.

3.3 Reasoning Coverage Construction

CROG first prompts the LLM to generate a structured reasoning plan composed of atomic reasoning targets:

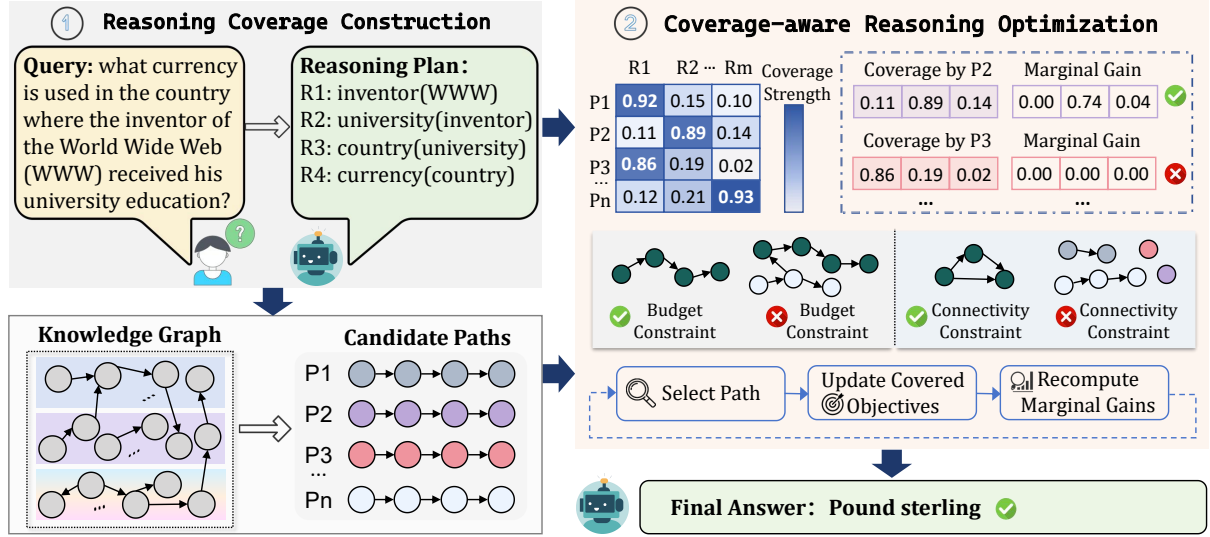


Figure 2: Overview of CROG. The proposed method constructs compact evidence subgraphs through coverage-aware reasoning optimization under budget and connectivity constraints.

$$\text{Plan}(q) = \{r_j\}_{j=1}^m, \quad (4)$$

where each r_j denotes an atomic reasoning target required for answering the question. The detailed prompt template used for reasoning plan construction is provided in Appendix E.

Each reasoning target is represented as a semantic feature vector. The resulting reasoning target set is denoted as:

$$\mathcal{R} = \{\mathbf{u}_1, \dots, \mathbf{u}_m\}, \quad (5)$$

where $\mathbf{u}_j$ denotes the semantic representation of reasoning target r_j .

For a candidate path p_i and reasoning target $\mathbf{u}_j$, CROG measures their semantic alignment through the following coverage function:

$$s(p_i, \mathbf{u}_j) = \max_{\mathbf{v} \in \text{Emb}(p_i)} \max(\cos(\mathbf{v}, \mathbf{u}_j), 0), \quad (6)$$

where $\text{Emb}(p_i)$ denotes the set of node embeddings within path p_i , and $\cos(\cdot, \cdot)$ denotes cosine similarity. Negative similarities are truncated to zero to avoid introducing negative coverage contributions from semantically irrelevant paths.

This formulation reflects that a reasoning target is considered covered if at least one node within the candidate path strongly supports the corresponding reasoning target.

3.4 Coverage-aware Reasoning Optimization

Existing graph-based reasoning methods typically organize evidence by independently ranking candidate paths according to their relevance to the query. However, independently high-scoring paths often contain substantial semantic overlap and repeatedly support similar reasoning targets.

To address this issue, CROG formulates evidence organization as a constrained reasoning coverage optimization problem and iteratively selects paths according to their additional coverage over previously under-covered reasoning targets.

The overall reasoning coverage of a selected evidence set $\mathcal{S}$ is defined as:

$$\text{Coverage}(\mathcal{S}) = \frac{1}{m} \sum_{j=1}^m \max_{p \in \mathcal{S}} s(p, \mathbf{u}_j). \quad (7)$$

This formulation encourages complementary evidence organization by rewarding paths that improve coverage for previously under-covered reasoning targets.

Marginal Coverage Gain. The optimization process is initialized with an empty evidence set $\mathcal{S} = \emptyset$. For each reasoning target r_j , CROG maintains the current best coverage score:

$$c_j = \max_{p \in \mathcal{S}} s(p, \mathbf{u}_j), \quad (8)$$

where c_j represents the current coverage status of reasoning target r_j .

The marginal coverage gain of candidate path p_i is then defined as:

$$\Delta(p_i | \mathcal{S}) = \frac{1}{m} \sum_{j=1}^m \max(s(p_i, \mathbf{u}_j) - c_j, 0). \quad (9)$$

Consequently, semantically overlapping paths that repeatedly support already covered reasoning targets naturally receive lower marginal gains during optimization.

Structural Constraints. Let

$$\mathcal{E} = \bigcup_{p \in \mathcal{S}} \text{nodes}(p) \quad (10)$$

denote the current evidence node set induced by the selected paths.

At each iteration, CROG only considers candidate paths satisfying the following constraints:

1. $|\mathcal{E} \cup \text{nodes}(p_i)| \leq B$;
2. the induced evidence subgraph after adding p_i contains at most K connected components.

The budget constraint encourages compact evidence organization, while the connectivity constraint promotes structurally coherent evidence subgraphs for multi-hop reasoning. Connectivity is efficiently maintained using a Union-Find structure over the induced evidence subgraph.

At each iteration, CROG greedily selects the feasible candidate path with the highest marginal coverage gain and updates the covered reasoning targets until no feasible candidate remains.

3.5 Theoretical Analysis

Theorem 1. *The constrained evidence subgraph optimization problem in CROG is NP-hard.*

Proof. We reduce Maximum Coverage to CROG by constructing one candidate path for each set and one reasoning target for each universe element. Under this construction, selecting paths that maximize reasoning coverage is equivalent to selecting sets that maximize covered elements under a budget constraint. Detailed proof is provided in Appendix B. $\square$

Moreover, we prove that the proposed reasoning coverage objective is monotone submodular under the defined coverage formulation. The greedy optimization procedure therefore follows

the standard greedy approximation framework for monotone submodular maximization (Nemhauser et al., 1978). Detailed analysis is provided in Appendix C.

4 Experiment

4.1 Experiment Setting

Datasets. We evaluate CROG on four widely-used multi-hop question answering benchmarks: HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (2Wiki) (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), and MultiHopRAG (Tang and Yang, 2024). HotpotQA focuses on bridge-style and comparison-based reasoning over Wikipedia documents. 2WikiMultiHopQA contains compositional multi-hop questions involving diverse entity-relation chains and explicit reasoning dependencies across documents. MuSiQue introduces more challenging reasoning compositions with reduced shortcut bias, where answering requires integrating multiple non-trivial reasoning hops. MultiHopRAG consists of retrieval-intensive multi-hop questions designed to require evidence aggregation across multiple semantically related documents and entities.

Baselines. We compare CROG against several representative retrieval and graph-based reasoning methods, including: (1) Retrieval Only (Ram et al., 2023), which directly retrieves passages or sentences without graph reasoning; (2) RankLlama (Ma et al., 2024), a strong dense retrieval reranker; (3) SetR-SelectionOnly, SetR-CoT, and SetR-CoT&IRI (Lee et al., 2025), set-level evidence organization and refinement methods from the SetR framework; (4) LightRAG (Guo et al., 2025), a lightweight graph-based RAG method; (5) KGP (Wang et al., 2024), a knowledge graph guided retrieval framework; and (6) SentGraph (Liang et al., 2026), a sentence-level graph reasoning baseline most related to our work.

Implementation Details. We build CROG on top of sentence-level document graphs following recent graph-based multi-hop QA studies. Sentence-level evidence provides finer-grained reasoning signals and enables more precise evidence composition across reasoning hops, which is particularly beneficial for multi-hop reasoning where supporting evidence is often sparsely distributed across different local contexts.

In the constructed sentence graph, nodes correspond to individual sentences and edges represent

Model	Retrieval Unit	Avg # Units	HotpotQA		2WikiMultiHopQA		MuSiQue		MultiHopRAG Accuracy(%)
			EM(%)	F1(%)	EM(%)	F1(%)	EM(%)	F1(%)	
RETRIEVAL ONLY	Passage	3.00	11.00	18.09	3.80	10.34	2.40	6.86	44.40
RETRIEVAL ONLY	Sentence	3.00	38.60	50.88	30.80	40.01	17.00	28.57	60.20
RankLlama	Passage	5.00	31.88	32.95	32.24	25.78	7.61	11.77	43.51
SetR-Selection only	Passage	3.41	36.68	37.84	34.84	29.40	10.38	15.28	46.20
SetR-CoT	Passage	2.88	36.46	38.20	35.34	30.34	9.76	14.31	45.26
SetR-CoT & IRI	Passage	2.91	36.62	38.11	35.44	30.35	10.79	15.43	47.14
LightRAG	Passage	3.00	27.17	37.75	17.40	26.99	8.60	17.71	20.44
KGP	Passage	3.00	44.00	58.73	36.80	48.20	21.20	34.72	63.40
SentGraph	Sentence	2.70	48.80	62.92	42.00	52.26	26.80	40.36	65.60
CROG	Sentence	2.62	52.60	65.49	47.53	57.00	30.80	43.33	67.40

Table 1: Main results on four multi-hop QA benchmarks using BGE-large-en-v1.5 as the dense retriever. Bold and underlined values denote the best and second-best performance, respectively.

semantic or discourse-level relations between sentences. Candidate evidence paths are extracted via BFS expansion from semantic anchor nodes retrieved by dense retrievers. We use dense embeddings to encode both reasoning targets and sentence representations for reasoning coverage estimation.

During inference, CROG performs reasoning-guided evidence subgraph selection under budget and connectivity constraints to maximize complementary reasoning coverage while maintaining structural coherence of the selected evidence graph.

We evaluate multiple backbone LLMs, including Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen2.5-14B-Instruct, and Qwen2.5-32B-Instruct (Yang et al., 2024). Exact Match (EM) and F1 are reported for HotpotQA, 2WikiMultiHopQA, and MuSiQue, while Accuracy is used for MultiHopRAG.

4.2 Main Results

Dense Retriever Results. Table 1 presents the main results using BGE-large-en-v1.5 (Xiao et al., 2024) as the dense retriever across four multi-hop QA benchmarks.

Overall, CROG consistently achieves the best performance across all datasets and evaluation metrics while using fewer evidence units on average than most baselines. Compared with SentGraph, CROG improves EM from 48.80 to 52.60 and F1 from 62.92 to 65.49 on HotpotQA. On 2WikiMultiHopQA, EM and F1 increase from 42.00 and 52.26 to 47.53 and 57.00, respectively. Similar improvements are observed on MuSiQue, where CROG improves EM from 26.80 to 30.80 and F1 from 40.36 to 43.33.

Sentence-level evidence organization consistently outperforms passage-level retrieval baselines across all benchmarks, indicating the importance of fine-grained evidence composition for multi-hop

reasoning. Building on this observation, the strong performance of CROG further demonstrates that effective multi-hop QA requires not only fine-grained evidence retrieval, but also reasoning-aware evidence organization. While existing methods mainly focus on improving retrieval relevance or refining evidence sets, CROG explicitly models evidence selection as a reasoning coverage optimization problem over sentence graphs. By dynamically prioritizing evidence paths that cover previously uncovered reasoning targets, CROG encourages complementary evidence aggregation while reducing redundant reasoning paths.

Notably, CROG achieves these gains while using fewer evidence units than most competing approaches. CROG retrieves only 2.62 evidence units on average, compared with 2.70 for SentGraph and larger evidence sets used by passage-level retrieval baselines. These results indicate that the proposed coverage-aware subgraph selection strategy improves both reasoning quality and evidence efficiency by organizing compact yet complementary evidence subgraphs.

Robustness Under Lexical Retrieval. We further evaluate CROG using BM25 (Robertson and Zaragoza, 2009) as the retriever, as shown in Table 2. Despite the substantially weaker lexical retrieval setting, CROG still consistently achieves the best performance across all benchmarks.

Compared with SentGraph, CROG improves EM from 43.80 to 47.20 and F1 from 57.13 to 59.75 on HotpotQA. On 2WikiMultiHopQA, EM and F1 increase from 32.20 and 39.48 to 39.80 and 47.32, respectively. Similar improvements are observed on MuSiQue, where CROG improves EM from 16.00 to 20.80 and F1 from 26.83 to 33.04.

These results suggest that the effectiveness of CROG does not rely on a specific dense retriever.

Model	Retrieval Unit	Avg # Units	HotpotQA		2WikiMultiHopQA		MuSiQue		MultiHopRAG Accuracy(%)
			EM(%)	F1(%)	EM(%)	F1(%)	EM(%)	F1(%)	
RETRIEVAL ONLY	Passage	3.00	9.20	15.20	3.40	9.65	1.40	4.27	37.20
RETRIEVAL ONLY	Sentence	3.00	31.20	42.44	23.60	29.29	9.60	16.86	61.60
RankLlama	Passage	5.00	29.48	27.82	30.30	21.91	6.04	9.26	42.09
SetR-Selection only	Passage	2.95	31.61	30.55	32.22	24.20	8.02	11.07	43.62
SetR-CoT	Passage	2.48	30.79	30.12	32.07	24.43	7.03	10.87	41.63
SetR-CoT & IRI	Passage	2.63	32.20	30.57	32.17	24.22	6.62	10.57	44.13
LightRAG	Passage	3.00	28.66	39.06	17.78	27.43	9.83	18.71	26.65
KGP	Passage	3.00	36.82	49.94	22.80	31.21	9.31	18.66	62.00
SentGraph	Sentence	2.57	43.80	57.13	32.20	39.48	16.00	26.83	63.40
CROG	Sentence	2.58	47.20	59.75	39.80	47.32	20.80	33.04	64.60

Table 2: Main results on four multi-hop QA benchmarks using BM25 as the retriever. Bold and underlined values denote the best and second-best performance, respectively.

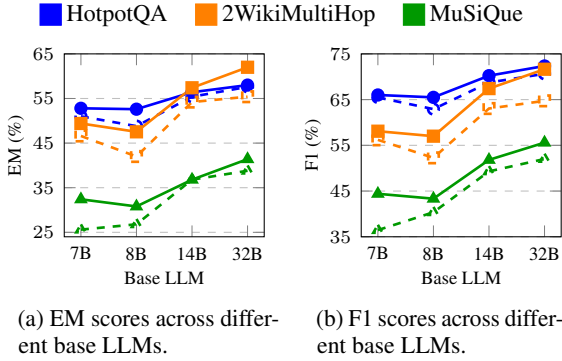


Figure 3: Performance comparison across datasets and base LLMs. Colors denote datasets. Solid lines represent CROG, while dashed lines represent SentGraph.

Instead, the proposed reasoning-guided evidence organization framework generalizes well across different retrieval backbones.

Generalization Across Backbone LLMs. To further evaluate the robustness of CROG under different backbone model scales and architectures, we conduct experiments using multiple backbone LLMs ranging from 7B to 32B parameters. Figure 3 presents the corresponding results on the multi-hop QA benchmarks.

Overall, both CROG and SentGraph generally benefit from larger backbone LLMs, with performance trends improving as model scale increases. Across all evaluated model scales and datasets, CROG consistently outperforms SentGraph, demonstrating that the proposed coverage-aware evidence organization strategy remains effective under diverse backbone configurations. The performance improvements remain stable across different LLM architectures and parameter sizes, indicating that the proposed reasoning coverage optimization operates largely orthogonally to backbone model scaling.

Notably, the gains brought by CROG remain

RCC	CRO	HotpotQA		2Wiki		MuSiQue	
		EM	F1	EM	F1	EM	F1
x	✓	49.80	63.45	41.80	52.03	26.80	40.54
x	x	48.00	61.52	41.00	51.85	25.60	38.40
✓	✓	52.60	65.49	47.53	57.00	30.80	43.33

Table 3: Ablation study on the two core components of CROG: Reasoning Coverage Construction (RCC) and Coverage-aware Reasoning Optimization (CRO).

substantial even for larger backbone LLMs. This suggests that stronger parametric reasoning capabilities alone are insufficient to fully resolve evidence organization challenges in multi-hop QA, and that explicit coverage-aware evidence optimization continues to provide complementary benefits beyond backbone scaling.

4.3 Ablation Study

Effect of Reasoning Coverage Construction and Coverage-aware Reasoning Optimization. Table 3 evaluates the contribution of the two core components in CROG: (1) *Reasoning Coverage Construction* (RCC), which decomposes the question into reasoning-oriented sub-goals, and (2) *Coverage-aware Reasoning Optimization* (CRO), which performs marginal coverage based evidence path selection.

Removing RCC consistently degrades performance across all benchmarks. For example, on 2WikiMultiHopQA, EM decreases from 47.53% to 41.80%, while F1 drops from 57.00% to 52.03%. Similar performance reductions are observed on HotpotQA and MuSiQue. These results indicate that explicitly modeling reasoning-oriented sub-goals provides important structural guidance for evidence organization. Without reasoning coverage construction, the retrieval process lacks explicit signals about which reasoning requirements remain insufficiently supported, making it more difficult to

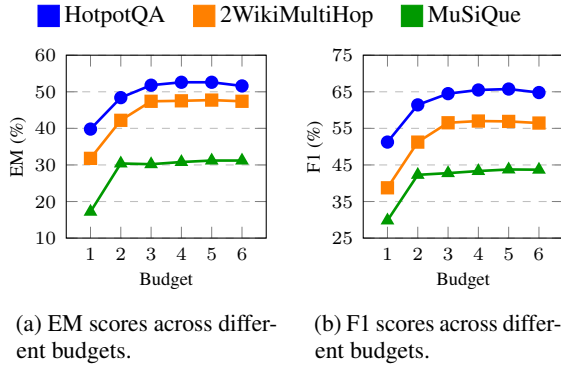


Figure 4: Performance under different evidence budget constraints across multi-hop QA benchmarks.

organize complementary evidence across different reasoning hops.

Further removing CRO leads to an additional performance drop on all datasets. In particular, MuSiQue EM decreases from 26.80% to 25.60%, while HotpotQA F1 drops from 63.45% to 61.52%. This demonstrates the importance of coverage-aware marginal optimization during evidence subgraph construction. Instead of repeatedly selecting semantically similar evidence paths, CRO dynamically prioritizes paths that contribute additional reasoning coverage beyond already selected evidence, thereby encouraging complementary evidence aggregation across different reasoning targets.

Effect of Budget Constraints. Figure 4 presents the performance of CROG under different evidence budget constraints across multiple multi-hop QA benchmarks. Overall, performance improves substantially as the budget increases from 1 to 3, indicating that multi-hop reasoning requires sufficient evidence capacity to support complementary reasoning evidence across different reasoning hops. After the budget reaches approximately 3–5 evidence units, performance becomes relatively stable across all datasets. For example, on HotpotQA, EM increases from 39.80 at budget 1 to 51.80 at budget 3, while larger budgets provide only marginal additional gains. Similar trends are observed on 2WikiMultiHopQA and MuSiQue. These results suggest that CROG can effectively construct compact evidence subgraphs that capture most reasoning requirements without requiring excessively large evidence contexts.

Effect of Connectivity Constraints. We study the effect of the connectivity constraint. As shown in Figure 5, overly strict connectivity constraints degrade performance, as forcing all selected paths

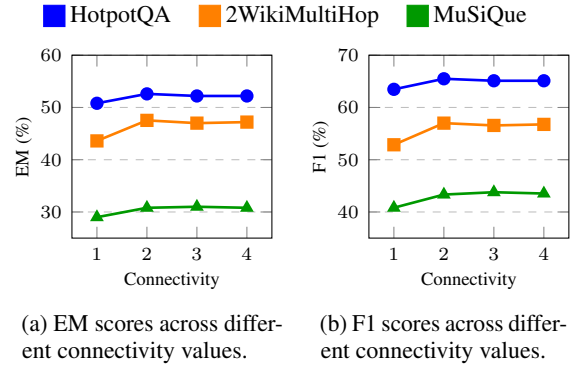


Figure 5: Performance comparison across datasets with varying graph connectivity.

to form a single connected component may overly restrict feasible evidence combinations and limit the aggregation of complementary reasoning evidence across different reasoning hops. Relaxing the connectivity constraint yields relatively stable performance across all datasets, indicating that the proposed method remains robust across a broad range of connectivity settings and does not rely on precise hyperparameter tuning.

4.4 Optimization Analysis

We further compare CROG-Greedy against exhaustive optimal search on small-scale instances containing at most 20 candidate paths. Across 400 randomly sampled instances from four datasets, each containing at most 20 candidate paths, the approximation ratio generally exceeds 0.95, indicating that the proposed greedy strategy achieves reasoning coverage close to the exhaustive optimum on evaluated instances. Detailed theoretical analysis and exhaustive-search evaluation are provided in Appendix C and Appendix D.

5 Conclusion

We propose CROG, a coverage-aware reasoning optimization method for graph-based multi-hop question answering. By formulating evidence organization as a constrained coverage maximization problem, CROG performs marginal-gain-based path selection with node-budget and connectivity constraints to encourage complementary and compact evidence subgraphs. Extensive experiments demonstrate that CROG achieves superior reasoning performance while using fewer evidence units on average, and analysis based on exhaustive optimal search further validates the effectiveness of the proposed greedy optimization strategy.

Limitations

While CROG demonstrates strong performance on multiple multi-hop question answering benchmarks, several limitations remain. First, CROG optimizes evidence organization over the candidate paths produced by the upstream graph construction and retrieval process. Therefore, evidence that is not included in the candidate path set cannot be recovered by the subsequent coverage-aware optimization. Improving the recall and quality of upstream candidate evidence remains an important direction for future work.

Second, CROG defines reasoning coverage based on semantic similarity between evidence and reasoning targets. While this provides an explicit signal for organizing evidence around reasoning requirements, semantic coverage does not fully capture the logical correctness of the subsequent reasoning process. Integrating logic-aware verification into coverage-aware evidence optimization is a promising direction for future work.

Third, the current framework is evaluated on English-language benchmarks, and its generalization to other languages remains unexplored. Experiments are also conducted on general-domain datasets, and the performance of CROG in specialized domains has not been investigated. In future work, we plan to extend CROG to multilingual and domain-specific knowledge-intensive reasoning tasks.

Ethics Statement

This work focuses on improving evidence organization for multi-hop question answering and does not involve the collection of new datasets or human subjects. All datasets used in our experiments are publicly available benchmarks.

Acknowledgments

The research is partially supported by Quantum Science and Technology-National Science and Technology Major Project (QNMP) 2021ZD0302900, China National Natural Science Foundation with No. 92567301, 62132018, 62231015, and Anhui Provincial Science and Technology Innovation Program 20252309050019.

References

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering doc-

uments and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Boyue Chen, Zirui Guo, Zidan Yang, Yuluo Chen, Junze Chen, Zhenghao Liu, Chuan Shi, and Cheng Yang. 2026. [Pathrag: Pruning graph-based retrieval augmented generation with relational paths](#). In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 30183–30191.

Uriel Feige. 1998. A threshold of $\ln n$ for approximating set cover. *Journal of the ACM (JACM)*, 45(4):634–652.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2025. [Lightrag: Simple and fast retrieval-augmented generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 10746–10761.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. [Constructing A multi-hop QA dataset for comprehensive evaluation of reasoning steps](#). In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6609–6625.

Dahyun Lee, Yongrae Jo, Haeju Park, and Moontae Lee. 2025. [Shifting from ranking to set selection for retrieval augmented generation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 17606–17619.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Junli Liang, Pengfei Zhou, Wangqiu Zhou, Wenjie Qing, Qi Zhao, Ziwen Wang, Qi Song, and Xiangyang Li. 2026. [Sentgraph: Hierarchical sentence graph for multi-hop retrieval-augmented question answering](#). *arXiv preprint arXiv:2601.03014*.

- Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 510–520.
- Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2024. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.
- Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2421–2425.
- Costas Mavromatis and George Karypis. 2024. Gnn-rag: Graph neural retrieval for large language model reasoning. *arXiv preprint arXiv:2405.20139*.
- Xiangyang Mou, Mo Yu, Shiyu Chang, Yufei Feng, Li Zhang, and Hui Su. 2021. Complementary evidence identification in open-domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2720–2726.
- George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. 1978. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294.
- Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2025. Graph retrieval-augmented generation: A survey. *ACM Transactions on Information Systems*, 44(2):1–52.
- Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlga, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.
- Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.
- Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations*, volume 2024, pages 32628–32649.
- Jiashuo Sun, Chengjin Xu, Luminyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel Ni, Heung-Yeung Shum, and Jian Guo. 2024. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *International Conference on Learning Representations*, volume 2024, pages 3868–3898.
- Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391*.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Trans. Assoc. Comput. Linguistics*, 10:539–554.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 10014–10037.
- Yu Wang, Nedom Lipka, Ryan A. Rossi, Alexa F. Siu, Ruiyi Zhang, and Tyler Derr. 2024. Knowledge graph prompting for multi-document question answering. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pages 19206–19214.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.
- 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. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380.
- Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Hao Chen, Yilin Xiao, Chuang Zhou, Junnan Dong, et al. 2025. A survey of graph retrieval-augmented generation for customized large language models. *arXiv preprint arXiv:2501.13958*.

A Detailed Optimization Procedure and Complexity Analysis

Algorithm 1 presents the complete optimization procedure of CROG. Given a candidate path set $\mathcal{P}$

and reasoning target set $\mathcal{R}$, CROG iteratively selects paths according to marginal reasoning coverage gain under budget and connectivity constraints.

Algorithm 1 CROG Greedy Optimization

Require: Candidate paths $\mathcal{P}$, reasoning targets $\mathcal{R}$, evidence graph G , node budget B , component limit K
Ensure: Selected path subset $\mathcal{S}$

```

1:  $\mathcal{S} \leftarrow \emptyset$ 
2:  $\mathcal{E} \leftarrow \emptyset$ 
3:  $c_j \leftarrow 0, \forall j \in [m]$ 
4: while True do
5:    $\text{best\_path} \leftarrow \text{None}$ 
6:    $\text{best\_score} \leftarrow -\infty$ 
7:   for each candidate path  $p_i \in \mathcal{P} \setminus \mathcal{S}$  do
8:      $\mathcal{E}' \leftarrow \mathcal{E} \cup \text{nodes}(p_i)$ 
9:     if  $|\mathcal{E}'| > B$  then
10:      continue
11:     end if
12:     if  $\text{Comp}(\mathcal{E}', G) > K$  then
13:       continue
14:     end if
15:     Compute  $\Delta(p_i | \mathcal{S})$  according to Eq. (9)
16:     if  $\Delta(p_i | \mathcal{S}) > \text{best\_score}$  then
17:        $\text{best\_score} \leftarrow \Delta(p_i | \mathcal{S})$ 
18:        $\text{best\_path} \leftarrow p_i$ 
19:     end if
20:   end for
21:   if  $\text{best\_path}$  is None then
22:     break
23:   end if
24:    $\mathcal{S} \leftarrow \mathcal{S} \cup \{\text{best\_path}\}$ 
25:    $\mathcal{E} \leftarrow \mathcal{E} \cup \text{nodes}(\text{best\_path})$ 
26:   for each reasoning target  $u_j$  do
27:      $c_j \leftarrow \max(c_j, s(\text{best\_path}, u_j))$ 
28:   end for
29: end while
30: return  $\mathcal{S}$ 

```

The optimization terminates when no feasible candidate path remains. To efficiently enforce the connectivity constraint, CROG incrementally maintains connected components using a Union-Find structure over the induced evidence subgraph, avoiding repeated recomputation of graph connectivity during greedy selection.

We next analyze the computational complexity of CROG.

Precomputation Stage. For each candidate path and reasoning target pair, CROG computes cosine similarities between path node embeddings and reasoning target embeddings. Suppose the candidate path set contains n paths, the reasoning target set contains m targets, and the average path length is $\bar{l}$. The precomputation complexity is therefore:

$$\mathcal{O}(n \cdot \bar{l} \cdot m). \quad (11)$$

Greedy Optimization Stage. At each greedy iteration, CROG traverses all remaining candidate

paths to compute marginal coverage gains involving m reasoning targets. Over at most n iterations with a shrinking candidate set, the total greedy optimization complexity is:

$$\mathcal{O}(n^2 \cdot m). \quad (12)$$

Each greedy iteration additionally performs connectivity checks using a Union-Find structure over the induced evidence subgraph, requiring amortized

$$\mathcal{O}(\alpha(|V|)) \quad (13)$$

per update, where $\alpha(\cdot)$ denotes the inverse Ackermann function.

Therefore, the overall complexity of CROG is:

$$\mathcal{O}(n \cdot \bar{l} \cdot m + n^2 \cdot m). \quad (14)$$

B NP-Hardness of CROG Optimization

B.1 CROG-Decision Problem

We first define the decision version of the CROG optimization problem.

definition 1 (CROG-Decision). *Given an evidence graph $G = (V, E)$, candidate path set $\mathcal{P}$, reasoning target set $\mathcal{R}$, coverage score function $s : \mathcal{P} \times \mathcal{R} \rightarrow [0, 1]$, node budget B , component limit K , and coverage threshold T , determine whether there exists a subset $\mathcal{S} \subseteq \mathcal{P}$ such that:*

1. $|\bigcup_{p \in \mathcal{S}} \text{nodes}(p)| \leq B$;
2. *the induced subgraph contains at most K connected components*;
3. $F(\mathcal{S}) = \frac{1}{m} \sum_{j=1}^m \max_{p \in \mathcal{S}} s(p, u_j) \geq T$.

B.2 Reduction from Maximum Coverage

Theorem 2. *The CROG-Decision problem is NP-complete.*

Proof. We prove NP-hardness via polynomial reduction from the classical Maximum Coverage problem (Feige, 1998).

Given a Maximum Coverage instance with universe $U = \{u_1, \dots, u_m\}$, set family $\mathcal{F} = \{S_1, \dots, S_n\}$, budget k , and target coverage size t , we construct a CROG instance as follows.

Reasoning targets. Each universe element $u_j \in U$ corresponds to one reasoning target.

Graph construction. Construct an evidence graph containing n isolated nodes, where each node corresponds to one set S_i .

Candidate paths. Each path p_i contains exactly one node associated with set S_i .

Coverage scores. Define binary coverage scores:

$$s(p_i, u_j) = \begin{cases} 1 & u_j \in S_i \\ 0 & \text{otherwise} \end{cases} \quad (15)$$

Constraints. Set node budget $B = k$ and set the component limit sufficiently large so that the connectivity constraint is always satisfied.

Coverage threshold. Set $T = t/m$.

Under this construction, selecting at most k paths achieving coverage at least T is equivalent to selecting at most k sets covering at least t elements in Maximum Coverage. Therefore, $\text{MAXIMUMCOVERAGE} \leq_p \text{CROG-DECISION}$, and since Maximum Coverage is NP-complete, CROG-Decision is NP-hard.

Moreover, feasibility can be verified in polynomial time by checking the node budget, graph connectivity, and coverage score. Therefore, CROG-Decision is NP-complete. $\square$

Remark 1. The connectivity constraint is intentionally set to be non-binding in this reduction to isolate the combinatorial hardness introduced by reasoning coverage maximization under budget constraints.

C Submodularity Analysis

We define the CROG coverage objective as:

$$F(S) = \frac{1}{m} \sum_{j=1}^m \max_{p \in S} s(p, u_j) \quad (16)$$

where S denotes the selected path subset.

Proposition 1. The coverage objective $F(S)$ is monotone.

Proof. For any $\mathcal{A} \subseteq \mathcal{B}$, adding additional paths cannot decrease the maximum coverage score for any reasoning target:

$$\max_{p \in \mathcal{A}} s(p, u_j) \leq \max_{p \in \mathcal{B}} s(p, u_j) \quad (17)$$

Thus $F(\mathcal{A}) \leq F(\mathcal{B})$, and the objective is monotone. $\square$

Proposition 2. The coverage objective $F(S)$ is submodular.

Algorithm 2 Exhaustive Optimal Selection

Require: Candidate paths $\mathcal{P}$, evidence graph G , budget B , component limit K

Ensure: Optimal subset S^*

```

1: best_score  $\leftarrow -\infty$ 
2: best_subset  $\leftarrow \emptyset$ 
3: for each non-empty subset  $\mathcal{S} \subseteq \mathcal{P}$  do
4:   if budget or connectivity constraints violated then
5:     continue
6:   end if
7:   Compute  $F(\mathcal{S})$  according to Eq. (16)
8:   if  $F(\mathcal{S}) > \text{best\_score}$  then
9:     best_score  $\leftarrow F(\mathcal{S})$ 
10:    best_subset  $\leftarrow \mathcal{S}$ 
11:   end if
12: end for
13: return best_subset

```

Proof. For each reasoning target u_j , define $f_j(\mathcal{S}) = \max_{p \in \mathcal{S}} s(p, u_j)$. Consider two path subsets $\mathcal{A} \subseteq \mathcal{B}$ and a candidate path p . The marginal contributions are:

$$f_j(\mathcal{A} \cup \{p\}) - f_j(\mathcal{A}) = \max(s(p, u_j) - f_j(\mathcal{A}), 0) \quad (18)$$

$$f_j(\mathcal{B} \cup \{p\}) - f_j(\mathcal{B}) = \max(s(p, u_j) - f_j(\mathcal{B}), 0) \quad (19)$$

Since $f_j(\mathcal{A}) \leq f_j(\mathcal{B})$, we obtain:

$$f_j(\mathcal{A} \cup \{p\}) - f_j(\mathcal{A}) \geq f_j(\mathcal{B} \cup \{p\}) - f_j(\mathcal{B}) \quad (20)$$

Thus each f_j satisfies the diminishing-return property and is submodular. Since $F(\mathcal{S}) = \frac{1}{m} \sum_{j=1}^m f_j(\mathcal{S})$ is a nonnegative weighted sum of submodular functions, F is also submodular. $\square$

By the classical result of Nemhauser et al. (Nemhauser et al., 1978), greedy optimization achieves a $(1 - 1/e)$ approximation guarantee for monotone submodular maximization under cardinality constraints. This provides useful theoretical intuition for the approximation behavior of the proposed greedy strategy, particularly when the connectivity constraint is non-binding and the optimization is primarily governed by coverage maximization under evidence-budget constraints. Empirically, as shown in Figure 5, performance stabilizes once the connectivity threshold is moderately relaxed, suggesting that the practical optimization behavior is primarily governed by the coverage objective and budget constraint.

Dataset	Avg. #Paths	Avg. Ratio	Min Ratio
HotpotQA	6.2	1.0000	1.0000
2Wiki	5.3	0.9992	0.9395
MuSiQue	5.6	0.9984	0.9607
MultiHopRAG	13.8	0.9977	0.9520

Table 4: Approximation quality of CROG-Greedy compared with exhaustive optimal search.

D Exhaustive Optimal vs CROG-Greedy

D.1 Experimental Setup

To evaluate the empirical approximation quality of CROG-Greedy, we compare it against exhaustive optimal search on instances containing at most 20 candidate paths. For each sample, exhaustive search enumerates all feasible non-empty path subsets satisfying node-budget and connectivity constraints and selects the subset achieving the highest reasoning coverage score.

Unlike the marginal coverage gain used during greedy selection, the following evaluation metric measures the final global reasoning coverage achieved by the selected evidence subset:

$$F(S) = \frac{1}{m} \sum_{j=1}^m \max_{p \in S} s(p, u_j) \quad (21)$$

We further define the approximation ratio as:

$$\text{ApproxRatio} = \frac{F(S_{\text{greedy}})}{F(S_{\text{optimal}})} \quad (22)$$

An approximation ratio close to 1.0 indicates that the greedy solution achieves reasoning coverage comparable to the exhaustive optimum.

D.2 Exhaustive Optimal Search

Algorithm 2 describes the exhaustive optimal selection procedure.

D.3 Discussion

Table 4 shows that CROG-Greedy achieves approximation ratios extremely close to the exhaustive optimum across all datasets.

In particular, the average approximation ratio exceeds 0.99 on all benchmarks, while the minimum ratio remains above 0.93 even in the worst case. On HotpotQA, CROG-Greedy consistently matches the exhaustive optimal solution on all evaluated samples.

These results indicate that although the proposed optimization objective includes additional structural constraints beyond classical submodular maximization settings, the greedy strategy still achieves

Prompt Template
You are a question decomposition assistant for multi-hop question answering. Given a complex question, break it down into atomic, single-hop sub-questions. Each sub-question should be answerable from a single document/passage. Output ONLY valid JSON with the following structure: <pre> {{ "question_type": "yes_no comparison general", "key_entities": ["entity1", "entity2", ...], "reasoning_steps": ["sub-question 1", "sub-question 2", ...] }}</pre> Guidelines: - question_type: * yes_no: Questions asking whether something is true * comparison: Questions comparing two or more things * general: Other factual questions - key_entities: The core named entities mentioned in the question. - reasoning_steps: * Break the question into the smallest possible atomic sub-questions. * Each sub-question should require only ONE piece of information from ONE document. * For bridge questions (e.g., "What is the X of the person who did Y?"), create a chain of sub-questions. * For comparison questions, create sub-questions for each entity AND a sub-question about the comparison relationship itself. * Keep each sub-question concise and self-contained. Question: {question} Output JSON:

Figure 6: Reasoning coverage construction prompt in CROG.

near-optimal reasoning coverage in practice. Empirically, CROG-Greedy attains approximation ratios substantially higher than the classical worst-case $(1 - 1/e)$ bound commonly associated with monotone submodular maximization.

E Prompt

Figure 6 presents the prompt template used for reasoning coverage construction in CROG. Given a multi-hop question, the prompt guides the LLM to generate a structured reasoning plan consisting of atomic reasoning steps used for subsequent coverage-aware evidence optimization. Question type and key entity information are additionally introduced as auxiliary guidance signals to facilitate more effective reasoning plan construction.

F Implementation and Reproducibility Details

F.1 Hyperparameter Settings

Table 5 summarizes the main hyperparameter settings used in CROG.

Hyperparameter	Value
BFS depth	5
anchor nodes	20
Default node budget B	4
Connectivity limit K	2
Max candidate paths per anchor	20
Dense retriever	BGE-large-en-v1.5
Lexical retriever	BM25
Default backbone LLM	Llama-3.1-8B-Instruct
LLM decoding temperature	0.0
Random seeds	42, 123, 456

Table 5: Main hyperparameter settings used in CROG.

F.2 Experimental Environment and Reproducibility

All experiments are conducted on Linux servers equipped with Intel Xeon CPUs, 512 GB RAM, and 8 NVIDIA GeForce RTX 4090 GPUs (24 GB VRAM each). We report all experimental results averaged over three independent runs with random seeds 42, 123, and 456. All LLM inference uses greedy decoding with temperature set to 0.0 to ensure deterministic generation and stable evaluation results.

G Case Study

G.1 Success Case: Complementary Evidence Aggregation

Question: *What screenwriter with credits for “Evolution” co-wrote a film starring Nicolas Cage and Téa Leoni?*

Ground Truth: *David Weissman*

Prediction: *David Weissman*

This example requires jointly satisfying multiple reasoning constraints, including identifying a screenwriter associated with *Evolution* and linking that person to a film starring Nicolas Cage and Téa Leoni. The reasoning planner decomposes the question into multiple atomic reasoning targets corresponding to these constraints.

CROG retrieves complementary evidence paths covering different reasoning targets:

Path 1: “David Weissman: His film credits include *The Family Man* (2000),

Evolution (2001), and *When in Rome* (2010).”

Path 2: “David Weissman: David Weissman is a screenwriter and director.”

The first path links David Weissman to both *Evolution* and *The Family Man*, while the second path explicitly verifies the target occupation as a screenwriter. Together, the selected evidence paths provide complementary support for resolving the multi-constraint question. The marginal-coverage optimization mechanism prioritizes evidence paths according to their additional reasoning contribution, encouraging complementary evidence aggregation instead of repeatedly selecting semantically overlapping paths.

G.2 Failure Case: Answer Disambiguation Error

Question: *What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?*

Ground Truth: *Chief of Protocol*

Prediction: *United States ambassador to Ghana and to Czechoslovakia*

In this example, the retrieved evidence already contains the correct answer:

Path 1: “Kiss and Tell (1945 film): Kiss and Tell is a 1945 American comedy film starring then 17-year-old Shirley Temple as Corliss Archer.”

Path 2: “Shirley Temple: As an adult, she was named United States ambassador to Ghana and to Czechoslovakia and also served as **Chief of Protocol of the United States.**”

Although CROG successfully retrieves evidence covering the required reasoning target, the LLM predicts “United States ambassador to Ghana and to Czechoslovakia” instead of the correct answer “Chief of Protocol.” This failure is therefore caused by ambiguity during the final answer generation stage rather than evidence retrieval. When multiple candidate answers appear in the same evidence sentence, the LLM may fail to select the answer that best matches the intent of the question.