

## Article

# Structure-Aware Graph-RAG for Small Language Models: Reducing Hallucinations and Improving Multi-Step Reasoning in Specialized Domains

Ali Asghari <sup>1,\*</sup>, Mohammad Mojahedivaraki <sup>1</sup> and Abbas Barzegarinezhad <sup>2</sup><sup>1</sup> Department of Computer Engineering, Shafagh Institute of Higher Education, Tonekabon 683165363, Iran; mohammad.mojahedi@shafagh.ac.ir<sup>2</sup> Department of Mathematics, Shafagh Institute of Higher Education, Tonekabon 683165363, Iran; barzegari@shafagh.ac.ir

\* Correspondence: asghari\_ali@aut.ac.ir

## Abstract

Retrieval-augmented generation (RAG) has been widely used to improve factual grounding in small language models (SLMs). However, many existing retrieval methods rely mainly on shallow semantic similarity between queries and text chunks. As a result, the retrieved evidence is often fragmented or weakly connected, which can limit multi-hop reasoning and sometimes lead to hallucinated answers. In this work, we propose Hallucination-Aware Multi-Objective Ant Colony Optimization (HA-MOACO), a structure-aware Graph-RAG framework that models relationships between pieces of evidence and searches for useful reasoning paths in a graph structure. The framework first constructs a domain knowledge graph and identifies several candidate entry nodes related to the query. Starting from these nodes, an ant colony optimization (ACO) strategy is used to explore possible evidence paths. During this process, multiple signals are considered simultaneously, including semantic relevance, structural connectivity between nodes, and reliability indicators that help reduce contradictory or low-confidence evidence. The selected evidence paths are then used to build a compact and grounded context for the language model. This process helps keep important reasoning connections while reducing the influence of irrelevant information. Experimental results on reasoning-focused benchmarks show that the proposed framework improves answer accuracy by about 6%, increases F1 by roughly 5–6%, and improves exact-match scores by around 7% compared with strong graph-based retrieval baselines. At the same time, the hallucination rate is reduced by about 40%, while retrieval latency is roughly 20% lower. These results suggest that combining structure-aware retrieval with reliability-oriented optimization can improve both reasoning quality and factual consistency in SLMs, while still remaining efficient for practical deployment.

Academic Editor: Renato Bruni

Received: 18 May 2026

Revised: 9 July 2026

Accepted: 10 July 2026

Published: 12 July 2026

**Copyright:** © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

**Keywords:** structure-aware Graph-RAG; HA-MOACO; multi-hop reasoning; hallucination robustness; evidence path optimization

**MSC:** 68T01; 68W01; 68W50; 68U07; 68U03

## 1. Introduction

In recent years, SLMs [1] have gained increasing attention as a practical solution for building intelligent language systems in resource-limited environments. Compared with large language models (LLMs) [2], which typically require substantial computational resources, memory capacity, and cloud-based infrastructure, SLMs offer several practical advantages. They are generally more efficient, less expensive to deploy, faster in response time, and often more suitable for privacy-sensitive or edge-based applications. Because of these properties, SLMs are increasingly adopted in enterprise systems, industrial environments, embedded devices, and other domain-specific scenarios where efficiency and data confidentiality are important. Despite these advantages, SLMs still face important limitations that affect their reliability in specialized tasks [3]. Two major challenges are hallucination [4], where the model produces fluent but unsupported information, and weak multi-step reasoning [5], particularly when answering a question requires integrating multiple related facts across documents or concepts. These issues become more critical in knowledge-intensive domains such as healthcare, law, finance, engineering, and scientific knowledge management. In such settings, answers are expected to be supported by clear evidence and logically consistent reasoning. An incorrect but confident response may lead to a misleading interpretation or poor decision-making. Moreover, many real-world domain questions are inherently multi-hop in nature and require reasoning over connected concepts rather than retrieving a single isolated fact. Since SLMs have more limited parametric capacity than larger models, their performance depends strongly on the quality, organization, and relevance of the external evidence provided to them. RAG [6] has been widely used to mitigate some of these limitations by grounding model outputs in external knowledge sources. Instead of relying only on internal model parameters, a RAG system retrieves relevant documents or text segments and then generates an answer conditioned on this retrieved context. This paradigm is especially beneficial for SLMs because it extends practical access to knowledge without increasing the model's parametric size. However, most conventional RAG pipelines rely on linear retrieval mechanisms, such as dense vector similarity search or keyword matching, and treat retrieved evidence as independent text fragments. While such approaches may work reasonably well for simple factual queries, they often struggle when the task requires relational understanding, combining information across sources, or multi-step reasoning. The top-ranked chunks may individually appear relevant, yet fail to form a coherent and sufficient reasoning path when considered together.

To address these limitations, recent studies have introduced Graph-based RAG (Graph-RAG) frameworks [7,8], where knowledge is organized as a structured graph composed of entities, concepts, documents, and their semantic relations. By explicitly modeling structural dependencies, Graph-RAG provides a richer representation of the knowledge space and can better support reasoning over interconnected information. This structure allows retrieval to consider not only semantic similarity, but also relational context between pieces of evidence. Nevertheless, although the graph structure is available, many existing Graph-RAG methods still rely on heuristic or locally guided strategies for selecting evidence paths. In other words, the traversal of the graph is often not formulated as a principled optimization problem. As a result, the system may select nodes that are locally related but globally suboptimal, overlook more informative reasoning chains, or include noisy or weakly connected evidence. For SLMs, which are highly sensitive to context quality, such suboptimal path selection can significantly affect answer faithfulness and reasoning completeness. Therefore, the key challenge is not only how to construct a structured retrieval space, but also how to search this space in an adaptive and principled manner so that the selected evidence improves relevance, multi-step reasoning coverage, and factual consistency, while controlling retrieval cost.

To address this challenge, we propose HA-MOACO, a structure-aware Graph-RAG framework that integrates ACO [9,10] into the evidence path selection process. In the proposed framework, evidence path discovery is formulated as a multi-objective optimization problem over a knowledge graph. Instead of ranking isolated text chunks, the system searches for coherent reasoning paths that jointly optimize several criteria, including semantic relevance to the query, structural connectivity, evidence coverage, diversity, reliability indicators, and retrieval cost. Inspired by the collective foraging behavior of ants, the ACO mechanism performs iterative exploration of the graph space, where artificial ants construct candidate paths and update pheromone trails based on path quality. Through repeated exploration and adaptive pheromone refinement, more informative and reliable reasoning paths are gradually strengthened, while weaker or contradictory paths are suppressed. In this way, retrieval becomes a guided and structure-aware search process rather than a static similarity-based ranking step.

The proposed framework operates in several stages. First, domain knowledge is converted into a structured graph in which nodes represent entities, concepts, passages, or documents, and edges capture semantic, contextual, or citation-based relations. Second, the input query is analyzed to identify key concepts and determine candidate entry nodes for graph traversal. Third, the ACO-based module performs a runtime search over the graph to identify globally coherent evidence paths that are suitable for multi-hop reasoning. Finally, the optimized subgraph or selected path set is provided to the SLM as grounded context for answer generation. By combining graph structure with optimization-based search, the framework aims to reduce unsupported outputs, improve reasoning faithfulness, and enhance overall reliability in specialized domains.

To evaluate the effectiveness of the proposed framework, we compare HA-MOACO with four state-of-the-art baselines: BGPT, IC-HRAG, RuleRAG, and MaRAG. These methods represent different retrieval and reasoning strategies in RAGs.

The comparison, therefore, includes four baselines in addition to the proposed HA-MOACO method. The evaluation considers not only answer accuracy, but also evidence faithfulness, hallucination rate, reasoning coverage, and retrieval efficiency. In addition, ablation studies are performed to analyze the contribution of graph structuring, ACO-based optimization, and grounding strategies.

The main contributions of this work are summarized as follows:

- Formulation of evidence path selection in Graph-RAG as a multi-objective optimization problem, jointly considering semantic relevance, structural connectivity, reasoning coverage, reliability signals, and retrieval cost.
- Introduction of HA-MOACO, a structure-aware training free Graph-RAG framework that integrates ACO to discover coherent multi-hop reasoning paths for SLMs at inference time.
- A hallucination-aware path evaluation mechanism that incorporates reliability and structural consistency signals to improve factual grounding and reduce unsupported generation.
- Comprehensive experimental evaluation on domain-specific and multi-hop reasoning benchmarks, comparing the proposed framework with several baseline methods using accuracy, faithfulness, hallucination, and efficiency metrics.
- The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed methodology. Section 4 describes the experimental setup and evaluation results. Finally, Section 5 concludes the paper and discusses future directions.

## 2. Related Works

Recent studies indicate that RAG can significantly improve the reliability of language models in knowledge-intensive domains. By integrating external knowledge sources into the generation process, RAG frameworks help reduce hallucination and improve factual grounding. However, conventional RAG approaches typically rely on dense vector similarity or keyword matching for document retrieval. Such strategies often treat retrieved information as independent text fragments and do not fully capture the structural relationships between pieces of knowledge. In many domain applications where information is distributed across multiple documents or concepts, semantic similarity alone may not be sufficient for identifying coherent reasoning paths. As a result, several recent studies have explored hybrid retrieval architectures and knowledge-graph-based reasoning frameworks to improve evidence selection and reasoning capability.

A first line of research focuses on hybrid RAG architectures that combine vector retrieval with structured knowledge representations. For example, the BuildingGPT framework (BGPT) [11] integrates vector search, graph-based retrieval, and automatic SPARQL generation to support natural language queries over semantic building data. Similar ideas were explored in manufacturing systems through a hybrid RAG architecture that incorporates domain knowledge graphs to improve reasoning in complex industrial scenarios [12]. The IC-HRAG framework [13] further enhances retrieval by projecting user queries into an ontology space, allowing the retrieval process to be guided by structured domain semantics. Other hybrid approaches combine metadata-enriched knowledge graphs with vector retrieval mechanisms in order to balance relational reasoning capability and retrieval scalability [14]. In construction management applications, RAG4CM [15] introduces hierarchical document parsing and preference-based retrieval to improve evidence selection in domain-specific queries. Similarly, Schema-Guided RAG [16] integrates knowledge graphs to support text-to-SQL query generation over multi-relational databases. In the medical domain, optimized RAG pipelines have also been developed that combine data cleaning, prompt engineering, and hybrid sparse–dense retrieval to reduce hallucination and improve answer reliability [17].

A second research direction investigates graph-based retrieval and reasoning frameworks, where knowledge graphs play a central role in representing domain knowledge. For instance, the MAKG framework [18] introduces a multi-agent knowledge graph architecture for intelligent maintenance reasoning in industrial environments. In the context of failure analysis, a knowledge graph-enhanced RAG system was developed for Failure Mode and Effects Analysis (FMEA), where standardized failure knowledge is encoded into graph embeddings to support semantic retrieval [19]. The GRAG ProSafe QAS framework [20] constructs a dynamic knowledge graph for industrial process safety and combines multi-hop retrieval with chain-of-thought reasoning to improve interpretability and answer reliability. In geospatial applications, GeoGraphRAG [21] models spatial relations within a graph structure and performs subgraph selection for geospatial reasoning tasks. Multimodal and educational knowledge retrieval has also been explored through frameworks such as MDKAG [22], which integrates multimodal disciplinary knowledge graphs with answer verification mechanisms to reduce hallucinations. Structure-aware retrieval representations have further been studied through Graph Codes and Multimedia Graph Codes [23,24], which encode entities and relations into graph-based feature matrices to preserve semantic topology and improve retrieval efficiency.

Another group of studies focuses on domain-specific knowledge integration and reasoning enhancement. For example, BuildingGPT2 [25] extends the original BGPT architecture by combining fine-tuning, vector retrieval, graph-RAG, and chain-of-thought prompting to improve SPARQL query generation over building knowledge graphs. Engineering knowledge extraction from patent documents has also been explored through

structured triple extraction methods that build knowledge bases for retrieval-augmented reasoning [26]. RuleRAG [27] integrates temporal logical rules with RAG to support reasoning over temporal knowledge graphs, while the DR-RAG framework [28] incorporates domain knowledge graphs, rule mining, and digital-twin feedback for aviation design decision support. In maritime accident analysis, the MaRAG framework [29] combines ontology-guided knowledge graphs with optimal transport-based feature alignment and multi-paradigm retrieval to enable causal reasoning and evidence tracing. In addition, industrial asset management studies have explored retrieval optimization strategies that combine Asset Administration Shell (AAS) representations with contrastive selection mechanisms to improve evidence filtering and reduce distractor information during inference [30].

Although these studies demonstrate that knowledge graphs and hybrid retrieval strategies can improve reasoning and factual grounding in specialized domains, several limitations remain. Many existing approaches still rely primarily on vector similarity during the retrieval stage, while the knowledge graph mainly acts as a supplementary component. Consequently, the retrieval process is often not fully structure-aware, particularly when multi-hop reasoning paths must be discovered across interconnected entities. In addition, most current frameworks are designed for LLMs and may not operate efficiently when applied to SLMs, which have more limited reasoning capacity and context memory. As a result, issues such as incomplete evidence paths, weak reasoning coverage, and hallucinated responses can still occur. To address these challenges, this study proposes a Structure-Aware Graph-RAG framework for SLMs. The proposed approach leverages the structural properties of domain knowledge graphs to guide the retrieval process and identifies evidence paths that support coherent multi-step reasoning. By moving beyond purely similarity-based retrieval and emphasizing structure-aware evidence exploration, the framework aims to improve reasoning consistency and reduce hallucinations in specialized domain applications.

To provide a concise overview of existing approaches, Table 1 compares representative graph-based and hybrid RAG frameworks reported in recent studies. The comparison focuses on key characteristics such as the use of knowledge graphs, hybrid retrieval strategies, multi-hop reasoning capability, and the mechanism used for evidence path selection. As shown in Table 1, most existing frameworks rely on heuristic ranking or similarity-based retrieval to select evidence from knowledge graphs. These strategies often treat the graph as a supporting resource rather than explicitly optimizing reasoning paths across it. In contrast, the proposed HA-MOACO framework formulates evidence path discovery as a hallucination-aware multi-objective optimization problem over the knowledge graph. By employing an ACO strategy to explore candidate reasoning paths while considering relevance, reasoning coherence, and hallucination risk, the framework aims to identify more reliable evidence chains for SLMs.

**Table 1.** Qualitative comparison of state-of-the-art graph-based RAG frameworks and their structural characteristics.

| Method                    | Knowledge Graph | Hybrid Retrieval | Multi-Hop Reasoning | Optimization-Based Path Search | Target Model |
|---------------------------|-----------------|------------------|---------------------|--------------------------------|--------------|
| BGPT [11]                 | Yes             | Yes              | Limited             | No                             | LLM          |
| MAKG [12]                 | Yes             | No               | Yes                 | No                             | LLM          |
| Hybrid KG-Vector RAG [17] | Yes             | Yes              | Limited             | No                             | LLM          |
| IC-HRAG [14]              | Yes             | Yes              | Limited             | No                             | LLM          |
| GeoGraphRAG [18]          | Yes             | No               | Yes                 | No                             | LLM          |
| RAG4CM [19]               | Partial         | Yes              | Limited             | No                             | LLM          |
| MDKAG [24]                | Yes             | Yes              | Yes                 | No                             | LLM          |

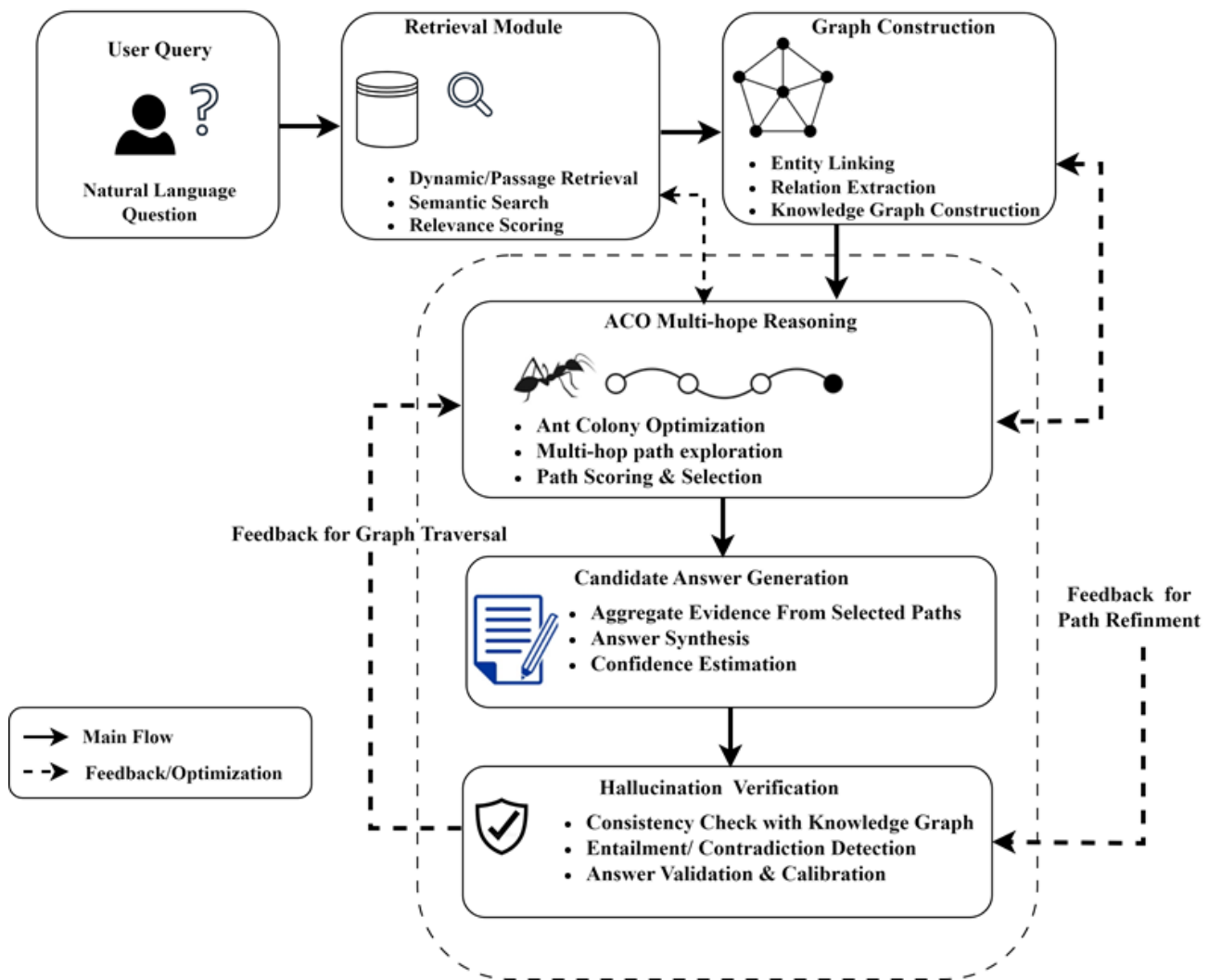
|                     |     |         |     |            |     |
|---------------------|-----|---------|-----|------------|-----|
| MaRAG [23]          | Yes | Yes     | Yes | No         | LLM |
| RuleRAG [29]        | Yes | Partial | Yes | Rule-based | LLM |
| DR-RAG [30]         | Yes | Yes     | Yes | No         | LLM |
| HA-MOACO (Proposed) | Yes | Yes     | Yes | Yes        | SLM |

### 3. HA-MOACO Framework

#### 3.1. Framework Overview

RAG has significantly improved the factual grounding of language models by incorporating external knowledge during generation. However, most existing RAG frameworks rely on similarity-based retrieval mechanisms that treat retrieved passages as independent textual units. While this paradigm works reasonably well for LLMs, it often becomes inefficient for SLMs, which operate under constrained context windows and weaker reasoning capabilities. In such settings, retrieving isolated passages frequently leads to incomplete reasoning chains and increased hallucination risk. To address these limitations, this work introduces a structure-aware Graph-RAG framework based on Hallucination-Aware Multi-Objective Ant Colony Optimization (HA-MOACO). The key idea is to reformulate evidence retrieval as a reasoning path optimization problem over a heterogeneous evidence graph. Instead of selecting individual passages solely based on semantic similarity, the framework searches for coherent reasoning paths that connect entities, concepts, and supporting evidence through meaningful semantic and structural relations. The retrieval space is modeled as an evidence graph  $G = (V, E)$ , where nodes represent entities, concepts, and evidence fragments, and edges encode semantic similarity, citation relations, and logical dependencies. Given an input query, relevant evidence candidates are first retrieved and integrated into this graph structure. The system then explores the graph to identify reliable reasoning paths that provide strong support for answering the query. To discover such paths, the framework employs a hallucination-aware multi-objective optimization strategy based on ACO. Multiple agents explore the graph and construct candidate reasoning trajectories while jointly optimizing several criteria, including semantic relevance to the query, structural connectivity between evidence nodes, factual grounding strength, and the risk of contradiction propagation. This process enables the system to identify evidence chains that are both semantically relevant and logically consistent. The optimized reasoning paths are subsequently converted into structured contextual evidence that guides response generation in the language model. By preserving the reasoning structure discovered during graph exploration, the framework encourages the model to follow evidence-supported inference trajectories. The overall architecture of the proposed HA-MOACO framework is illustrated in Figure 1.

To ensure methodological transparency and reproducibility, the framework implements query decomposition by parsing input queries into structured sub-concepts to identify key entities and entry nodes. Entities and concepts are extracted and mapped to graph nodes based on contextual relevance, while the knowledge graph is constructed by establishing weighted edges that reflect semantic similarity, citation links, and logical relations. Furthermore, factual consistency is maintained by identifying contradiction relations and integrating them into edge weights. This entire verification procedure is governed by our multi-objective fitness function, which proactively minimizes contradiction propagation and maximizes grounding reliability throughout the optimization process.



**Figure 1.** Overall Architecture of the Proposed HA-MOACO Framework.

### 3.2. Graph Construction and Query-Aware Candidate Retrieval

To capture structural dependencies among retrieved evidence, the proposed framework represents the retrieval space as a heterogeneous evidence graph rather than a flat collection of independent passages. Conventional retrieval methods typically rank passages based solely on semantic similarity, which often fails to capture the relational dependencies required for multi-hop reasoning. By modeling retrieved information as a graph structure, the framework explicitly represents semantic, logical, and contextual relationships between entities, documents, and evidence fragments, enabling more coherent reasoning path discovery. The evidence graph is defined in Equation (1).

$$G = (V, E) \quad (1)$$

where  $V$  denotes the set of nodes and  $E$  represents relations between nodes. The node set consists of heterogeneous components, shown by Equation (2).

$$V = \{V_e, V_c, V_d, V_q\} \quad (2)$$

where  $V_e$  denotes entity nodes,  $V_c$  represents concept and evidence-chunk nodes,  $V_d$  corresponds to document nodes, and  $V_q$  represents query-related nodes. Entity and concept nodes are extracted using domain-specific entity recognition and semantic parsing techniques, while evidence-chunk nodes correspond to retrieved textual passages from the domain corpus. Document nodes preserve document-level context and maintain structural relations between related evidence fragments.

Relations between nodes are represented by the edge set in Equation (3).

$$E = \{E_{sem}, E_{sup}, E_{cit}, E_{con}\} \quad (3)$$

where  $E_{sem}$  denotes semantic similarity relations,  $E_{sup}$  represents support relations between claims and evidence,  $E_{cit}$  corresponds to citation links between documents, and  $E_{con}$  captures contradiction relations. Each edge is associated with a reliability weight that reflects the strength and consistency of the connection between two nodes. The edge weight is defined by Equation (4).

$$w(i, j) = \alpha_w S_{sem}(i, j) + \beta_w S_{sup}(i, j) + \gamma_w S_{cit}(i, j) - \delta_w S_{con}(i, j) \quad (4)$$

where  $S_{sem}$ ,  $S_{sup}$ ,  $S_{cit}$ , and  $S_{con}$  denote semantic similarity, support strength, citation relevance, and contradiction score, respectively, and  $\alpha_w$ ,  $\beta_w$ ,  $\gamma_w$ ,  $\delta_w$  control the contribution of each component. After constructing the initial graph, low-confidence edges and weakly connected nodes are pruned in order to reduce noise and improve the efficiency of downstream reasoning.

To detect contradiction relations between evidence fragments during graph construction, we employ a natural language inference (NLI) model that evaluates the semantic relationship between pairs of candidate statements. The model predicts the likelihood of entailment, neutrality, or contradiction for each pair. The resulting contradiction probability is used to compute a contradiction score that contributes to the reliability weighting mechanism defined in Equation (4). In this way, edges connecting evidence fragments that exhibit stronger contradiction signals receive lower reliability weights, which helps suppress structurally inconsistent evidence during graph construction.

Given an input query  $Q$ , the framework performs semantic decomposition to identify relevant entities, concepts, and latent reasoning dependencies. Instead of representing the query as a single embedding vector, the method models it as a structured query graph by Equation (5).

$$Q_G = (V_Q, E_Q) \quad (5)$$

where  $V_Q$  denotes the set of extracted query concepts and  $E_Q$  represents semantic relations between these concepts. This representation allows the system to explicitly capture potential multi-hop reasoning requirements embedded in complex domain queries. For each query concept  $q_i \in V_Q$ , the retriever searches the domain corpus and identifies a candidate evidence set by Equation (6).

$$\mathcal{C}(q_i) = \{v_1, v_2, \dots, v_k\} \quad (6)$$

where each  $v_j$  corresponds to a candidate evidence node in the graph. To estimate the relevance between query concepts. Although the retriever identifies an initial candidate evidence set  $\mathcal{C}(q_i)$  used to initialize the reasoning process, the knowledge graph itself is constructed from the entire indexed document collection. Therefore, the ACO-based reasoning mechanism is able to explore neighboring nodes beyond the initial top-k retrieval results through graph connectivity. This design allows the framework to discover additional supporting evidence that may not appear in the initial retrieval stage. However, if the relevant information is absent from the indexed graph, it cannot be recovered during reasoning, which remains a general limitation of RAGs. To estimate the relevance between query concepts and candidate evidence nodes, the framework employs a hybrid semantic-structural relevance function, using Equation (7).

$$Rel(q_i, v_j) = \lambda_1 S_{emb}(q_i, v_j) + \lambda_2 S_{graph}(q_i, v_j) + \lambda_3 S_{ctx}(q_i, v_j) \quad (7)$$

where  $S_{emb}$  measures embedding-based semantic similarity,  $S_{graph}$  captures graph neighborhood compatibility, and  $S_{ctx}$  evaluates contextual consistency between

neighboring evidence nodes. The coefficients  $\lambda_1, \lambda_2$ , and  $\lambda_3$  control the contribution of each component.

To model the multi-step reasoning dependencies, the framework estimates the transition probability between evidence nodes during graph exploration by Equation (8).

$$P(v_j | v_i, Q) = \frac{\exp(\phi(v_i, v_j, Q))}{\sum_{v_k \in N(v_i)} \exp(\phi(v_i, v_k, Q))} \quad (8)$$

where  $N(v_i)$  denotes the neighborhood of the node  $v_i$ . The compatibility function is defined by Equation (9).

$$\phi(v_i, v_j, Q) = \alpha_r \text{Rel}(Q, v_j) + \beta_r w(v_i, v_j) - \gamma_r \text{Con}(v_i, v_j) \quad (9)$$

where  $w(v_i, v_j)$  represents the edge strength between nodes and  $\text{Con}(v_i, v_j)$  denotes the contradiction score between evidence fragments. The parameters  $\alpha_r$ ,  $\beta_r$ , and  $\gamma_r$  balance the influence of semantic relevance, structural connectivity, and contradiction penalty.

Based on these relevance and transition signals, the candidate retrieval stage aims to identify evidence subgraphs that simultaneously maximize semantic relevance and reasoning connectivity while minimizing contradiction propagation. This objective is formulated by Equation (10).

$$G^* = \arg \max_{G' \subseteq G} [\mu_1 R(G', Q) + \mu_2 C(G') - \mu_3 H(G')] \quad (10)$$

where  $R(G', Q)$  measures the overall relevance between the query and the retrieved subgraph,  $C(G')$  represents reasoning connectivity across evidence paths, and  $H(G')$  estimates the hallucination and contradiction risk within the selected subgraph.

The parameters of the proposed HA-MOACO framework are determined through a validation-based tuning procedure. In general, these parameters control the relative importance of different components involved in evidence retrieval and reasoning, such as semantic relevance, structural connectivity, and contradiction suppression. For example, the coefficients in Equation (4) regulate the balance between semantic compatibility and structural relations during graph traversal, while the objective in Equation (10) combines query–subgraph relevance, reasoning connectivity, and contradiction risk, subject to the token-budget constraint in Equation (11). The parameter values are selected via grid search on the validation set and then fixed for all experiments to ensure consistent and fair comparisons.

To ensure efficiency for SLMs with limited context capacity, the framework additionally enforces a token-budget constraint by Equation (11).

$$\sum_{v_i \in P_k} \text{Tok}(v_i) \leq B \quad (11)$$

where  $\text{Tok}(v_i)$  denotes the token cost associated with the node  $v_i$  and  $B$  represents the maximum context budget of the language model. This constraint encourages the selection of compact evidence structures that preserve essential reasoning information while remaining compatible with the limited context windows of SLMs. The resulting candidate evidence subgraphs define a structured reasoning space that is subsequently explored by the proposed HA-MOACO optimization module to identify the most reliable reasoning paths for grounded answer generation. This candidate evidence graph defines the reasoning search space that will be explored by the HA-MOACO optimization algorithm described in the next section.

To ensure the logical integrity of the constructed evidence graph, we specifically employ the DeBERTa-v3-large model [31] as our core Natural Language Inference (NLI) engine for contradiction detection. For each candidate edge representing a logical dependency, the NLI model computes probabilities for *entailment*, *neutrality*, and *contradiction*. An

edge is pruned or penalized if the contradiction probability exceeds a threshold of  $\tau = 0.7$ , ensuring that the reasoning path remains logically consistent.

The overall procedure for constructing the heterogeneous evidence graph and retrieving query-aware candidate evidence subgraphs is summarized in Algorithm 1.

---

**Algorithm 1.** Evidence Graph Construction and Query-Aware Candidate Retrieval

---

Input: Query  $Q$ , corpus  $D$ , retrieval budget  $k$

Output: Candidate evidence subgraph  $G^*$

1: Perform query decomposition to extract query concepts  $V_Q$

2: Initialize empty graph  $G = (V, E)$

3: **For** each concept  $q_i \in V_Q$  **do**

4: Retrieve top-k candidate passages  $C(q_i)$  from corpus  $D$

5: Extract entities and concepts from each passage

6: Add corresponding nodes to  $V$

7: Construct graph edges between nodes using semantic similarity, citation relations, support links, and contradiction signals

8: Compute edge weight

$$w(i, j) = \alpha_w S_{sem}(i, j) + \beta_w S_{sup}(i, j) + \gamma_w S_{cit}(i, j) - \delta_w S_{con}(i, j)$$

9: Prune weak edges and isolated nodes

10: Estimate query-evidence relevance

$$Rel(q_i, v_j) = \lambda_1 S_{emb} + \lambda_2 S_{graph} + \lambda_3 S_{ctx}$$

11: Construct candidate evidence subgraphs maximizing

$$G^* = \arg \max_{G' \subseteq G} [\mu_1 R(G', Q) + \mu_2 C(G') - \mu_3 H(G')]$$

12: Apply token budget constraint for SLM context window

13: **Return** candidate subgraph  $G^*$

---

### 3.3. Hallucination-Aware Evidence Path Optimization

Traditional ACO algorithms are primarily designed for shortest-path discovery and combinatorial optimization problems. However, reasoning over retrieved evidence in Graph-RAG environments introduces additional challenges that are not addressed by conventional ACO formulations. In particular, evidence exploration must account for semantic consistency between evidence fragments, contradiction propagation across reasoning steps, and the limited context capacity of SLMs. To address these challenges, the proposed framework introduces the HA-MOACO algorithm for discovering reliable reasoning paths in the evidence graph.

Given the heterogeneous evidence graph  $G = (V, E)$ , each ant incrementally constructs a reasoning trajectory.

$$P_k = (v_1, v_2, \dots, v_n) \quad (12)$$

where nodes represent evidence units and edges encode semantic and structural relations. Unlike traditional ACO formulations that rely only on pheromone strength and heuristic distance, the proposed algorithm incorporates additional signals that reflect semantic compatibility and factual reliability during evidence exploration.

To maintain reasoning coherence across multi-hop exploration, each ant maintains a semantic memory state, which records the semantic trajectory accumulated during exploration.

$$M_k^t = \{s_1, s_2, \dots, s_t\} \quad (13)$$

This memory mechanism allows the algorithm to discourage transitions that introduce semantic conflicts with previously visited evidence nodes. The semantic compatibility between a candidate node and the existing reasoning trajectory is defined by Equation (14).

$$\Psi(v_j, M_k^t) = \frac{1}{|M_k^t|} \sum_{s_i \in M_k^t} \text{Sim}(v_j, s_i) \quad (14)$$

where  $\text{Sim}(\cdot)$  measures semantic similarity between evidence representations. Transitions that produce low semantic compatibility are penalized to prevent reasoning drift and hallucination propagation. The probabilistic transition rule guiding ant exploration is defined by Equation (15).

$$p_{ij}^k = \frac{(\tau_{ij})^\alpha (\eta_{ij})^\beta (\Psi_{ij})^\gamma (\Gamma_{ij})^\delta}{\sum_{v_l \in N(v_i)} (\tau_{il})^\alpha (\eta_{il})^\beta (\Psi_{il})^\gamma (\Gamma_{il})^\delta} \quad (15)$$

where  $\tau_{ij}$  denotes pheromone intensity on edge  $(i, j)$ ,  $\eta_{ij}$  represents heuristic relevance to the query,  $\Psi_{ij}$  captures semantic consistency with the current reasoning trajectory, and  $\Gamma_{ij}$  measures factual grounding reliability between evidence nodes. To adjust the probability of selecting specific paths based on their cumulative quality, the pheromone concentration ( $\tau$ ) is evaluated for edges on the selected paths using the following rule:

$$\tau_{i,j} \leftarrow (1 - \rho) \cdot \tau_{i,j} + \sum_{k=1}^m \Delta \tau_{i,j}^k \quad (16)$$

where  $\rho \in (0, 1]$  is the evaporation rate,  $m$  is the number of ants in each iteration, and  $\Delta \tau_{i,j}^k$  is the amount of pheromone added by ant  $k$ , which is proportional to the quality score of the path.

To explicitly model hallucination risk during reasoning, the framework introduces a contradiction propagation function defined by Equation (17).

$$\Omega(P_k) = \sum_{(v_i, v_j) \in P_k} \text{Con}(v_i, v_j) e^{\kappa d_{ij}} \quad (17)$$

where  $\text{Con}(v_i, v_j)$  represents the contradiction intensity between evidence nodes and  $d_{ij}$  denotes their reasoning distance within the path. This formulation reflects the observation that contradictions appearing early in a reasoning chain may propagate and amplify hallucinations in later stages of answer generation.

The overall reasoning-path optimization objective integrates multiple criteria reflecting semantic relevance, structural coherence, factual grounding, contradiction risk, and contextual efficiency:

$$F(P_k) = \omega_1 R(P_k) + \omega_2 C(P_k) + \omega_3 G(P_k) - \omega_4 \Omega(P_k) - \omega_5 T(P_k) \quad (18)$$

where  $R(P_k)$  measures semantic relevance to the query,  $C(P_k)$  evaluates structural connectivity along the reasoning chain,  $G(P_k)$  represents factual grounding strength,  $\Omega(P_k)$  models contradiction propagation risk, and  $T(P_k)$  penalizes excessive token cost. The coefficients  $\omega_1, \omega_2, \omega_3, \omega_4$ , and  $\omega_5$  are weighting factors that determine the relative importance of semantic relevance, structural connectivity, factual grounding, contradiction risk, and token consumption, respectively, allowing the framework to balance reasoning quality with computational constraints.

The overall reasoning path optimization procedure of the proposed HA-MOACO framework is summarized in Algorithm 2.

---

**Algorithm 2.** HA-MOACO Reasoning Path Optimization
 

---

Input: Candidate evidence graph  $G = (V, E)$ , query  $Q$

Output: Optimal reasoning path set  $G^*$

- 1: Initialize pheromone matrix  $\tau_{ij}$
  - 2: **For** iteration = 1 to MaxIter **do**
  - 3:   **For** each ant  $k$  **do**
  - 4:     Initialize reasoning path  $P_k$
  - 5:     Initialize semantic memory  $M_k^t$
-

---

```

6:   While stopping condition not reached do
7:     Select next node  $v_j$  according to the transition probability

$$p_{ij}^k = \frac{(\tau_{ij})^\alpha (\eta_{ij})^\beta (\Psi_{ij})^\gamma (\Gamma_{ij})^\delta}{\sum_{v_l \in N(v_i)} (\tau_{il})^\alpha (\eta_{il})^\beta (\Psi_{il})^\gamma (\Gamma_{il})^\delta}$$

8:   Update the  $\tau_{i,j}$ 

$$\tau_{i,j} \leftarrow (1 - \rho) \cdot \tau_{i,j} + \sum_{k=1}^m \Delta \tau_{i,j}^k$$

9:     Update semantic memory  $M_k^t$ 
10:    Append node to reasoning path
11:  End While
12:  Evaluate path quality using

$$F(P_k) = \omega_1 R(P_k) + \omega_2 C(P_k) + \omega_3 G(P_k) - \omega_4 \Omega(P_k) - \omega_5 T(P_k)$$

13: End for
14: Update pheromone trails based on path quality
15: Apply pheromone evaporation
16: End for
17: Return the top-ranked reasoning paths

```

---

During optimization, multiple ants explore the graph in parallel and iteratively construct candidate reasoning paths. To control search complexity, exploration is restricted to the top-ranked neighboring nodes within a bounded beam width and hop depth. After each iteration, pheromone trails are updated according to the quality of the discovered reasoning paths, reinforcing trajectories that achieve higher multi-objective scores while gradually evaporating weaker paths.

Through iterative exploration and pheromone-based path refinement, the HA-MO-ACO algorithm effectively guides the search towards a set of high-confidence reasoning paths that provide semantically coherent and factually grounded evidence structures. These optimized reasoning trajectories are subsequently forwarded to the grounded response generation module to guide answer construction. The resulting reasoning paths form the structured evidence context used by the grounded generation module.

### 3.4. Evidence-Grounded Response Generation for SLMs

After the HA-MOACO optimization stage identifies reliable reasoning paths within the evidence graph, the framework performs grounded response generation using a reasoning-aware decoding strategy designed for SLMs. Unlike conventional RAG methods that simply concatenate retrieved passages into the prompt, the proposed framework transforms optimized reasoning paths into structured reasoning trajectories that guide the generation process.

Let  $G^* = \{P_1, P_2, \dots, P_m\}$  denote the set of high-quality reasoning paths selected by the HA-MOACO optimization stage. Each reasoning path  $P_i$  is transformed into a structured contextual representation through a semantic mapping function.

$$C_Q = \bigcup_{i=1}^m \Phi(P_i) \quad (19)$$

where  $\Phi(P_i)$  converts a reasoning path into an ordered evidence sequence suitable for language model conditioning. Instead of treating retrieved evidence as independent textual fragments, this transformation preserves the logical and semantic dependencies between evidence units discovered during graph exploration. Each reasoning path is linearized into an ordered reasoning chain.

$$\Phi(P_i) = (e_1 \rightarrow e_2 \rightarrow \dots \rightarrow e_n) \quad (20)$$

where the transition operators encode semantic support, causal relations, or citation dependencies between evidence fragments. This structure-aware representation encourages the language model to follow coherent reasoning trajectories during answer generation rather than performing unconstrained token prediction over disconnected evidence chunks.

To reduce hallucination during decoding, the framework introduces a grounding-aware token generation mechanism. Let  $y_t$  denote the token generated at the decoding step  $t$ . The probability of generating a token is adjusted by Equation (21).

$$P(y_t | y_{<t}, Q, C_Q) \propto P_{LM}(y_t) \cdot \Lambda(y_t, C_Q) \quad (21)$$

where  $P_{LM}(y_t)$  denotes the native language model probability and  $\Lambda(y_t, C_Q)$  represents a grounding compatibility function that evaluates whether the candidate token is semantically supported by the optimized evidence context. The grounding compatibility function is defined by Equation (22).

$$\Lambda(y_t, C_Q) = \exp(\theta_1 S_{ctx}(y_t, C_Q) - \theta_2 H(y_t)) \quad (22)$$

where  $S_{ctx}$  measures contextual support between the generated token and the evidence graph, while  $H(y_t)$  estimates hallucination likelihood caused by unsupported semantic expansion. To further improve efficiency for SLMs with limited context windows, the framework applies adaptive evidence compression before decoding. Instead of forwarding all retrieved evidence nodes, the system selects the smallest subset of evidence that preserves reasoning completeness:

$$C_Q^* = \arg \min_{C' \subseteq C_Q} |C'| \text{ s.t. } Cov(C') \geq \epsilon \quad (23)$$

where  $Cov(C')$  measures reasoning coverage over the optimized evidence graph, and  $\epsilon$  denotes the minimum acceptable completeness threshold.

Finally, the generated response is validated through a post-generation verification step that evaluates the factual alignment between the generated answer and the optimized evidence subgraph. Let  $A$  denote the generated response. Its verification score is defined by Equation (24)

$$Ver(A) = \frac{\sum_{f_i \in A} Sup(f_i, G^*)}{|A|} \quad (24)$$

where  $Sup(f_i, G^*)$  measures whether a generated factual claim  $f_i$  is supported by the selected evidence subgraph. Responses with low verification scores can be refined through iterative decoding or rejected to prevent unsupported outputs. By integrating reasoning-aware context construction, grounding-constrained decoding, and post-generation verification, the proposed framework enables SLMs to produce responses that are both semantically relevant and strongly grounded in retrieved evidence.

### 3.5. Joint Inference-Time Refinement Strategy

Although the proposed framework is composed of several interacting modules—including heterogeneous graph construction, query-aware evidence retrieval, hallucination-aware reasoning optimization, and grounded response generation—optimizing these components independently may lead to fragmented behavior and suboptimal reasoning quality. To address this limitation, the framework adopts a unified inference-time refinement that aligns retrieval quality, reasoning reliability, and generation fidelity under a common search-optimization objective.

Formally, let  $\theta$  denote the set of configuration coefficients and hyperparameters across the framework and  $J$  the overall multi-objective search function. The total optimization objective combines retrieval alignment, reasoning-path quality, and response grounding. Each module is configured to improve not only its local performance but also

its contribution to global factual reasoning. In particular, the retrieval objective encourages the selection of evidence nodes that are both semantically relevant to the query and structurally compatible with the evidence graph, while the reasoning objective reinforces evidence paths that exhibit high coherence and low contradiction risk.

To improve optimization stability, the framework introduces adaptive weighting across objectives, allowing the relative contribution of each component to vary according to its specific search difficulty and convergence behavior. This prevents dominant objectives from suppressing weaker but important signals during the optimization process. In addition, the framework integrates heuristic-based factual feedback into path selection: evidence trajectories that lead to factually verified answers receive higher scores during the ACO phase, whereas trajectories associated with unsupported or contradictory outputs are penalized.

Overall, this joint optimization strategy enables the system to perform self-refinement across the retrieval, reasoning, and generation stages at runtime, thereby improving multi-hop factual reasoning under the constrained context windows of SLMs. By coupling these selection signals across modules, the framework reduces hallucination risk and promotes more consistent evidence-grounded responses without requiring any gradient-based updates or fine-tuning of the underlying language model parameters. The framework operates entirely as a non-parametric, inference-time heuristic search, ensuring a strictly training-free deployment.

Figure 1 illustrates the overall workflow of the proposed framework and shows how retrieval, graph-based reasoning, path optimization, and answer verification interact during inference. Rather than repeating the full pipeline description, this section emphasizes the joint refinement mechanism: evidence retrieval provides candidate graph nodes, HA-MOACO optimizes reasoning paths over these nodes, and the verification signal is used to guide the selection of more reliable evidence paths. This interaction allows the framework to refine the reasoning process at inference time without updating the parameters of the underlying language model. All hyperparameter configurations and prompt engineering templates used in this study are provided in Appendix A to facilitate reproducibility.

Table 2 shows the Summary of Notation and Variables Used in this paper.

**Table 2.** Summary of Notation and Variables Used in HA-MOACO.

| Symbol                                  | Description                                          |
|-----------------------------------------|------------------------------------------------------|
| $G = (V, E)$                            | Evidence graph with nodes $V$ and edges $E$          |
| $Q$                                     | Input query                                          |
| $v_i, v_j$                              | Evidence nodes in the graph                          |
| $P_k$                                   | Evidence path constructed by ant $k$                 |
| $P^*$                                   | Final selected evidence path set                     |
| $\tau_{ij}$                             | Pheromone value on edge $(v_i, v_j)$                 |
| $\eta_{ij}$                             | Heuristic desirability of moving from $v_i$ to $v_j$ |
| $Sim(Q, v)$                             | Semantic similarity between the query and the node   |
| $M_k^t$                                 | Semantic memory of ant $k$ at step $t$               |
| $\Psi(v_j, M_k^t)$                      | Semantic consistency between the node and memory     |
| $p_{ij}^k$                              | Transition probability for ant $k$                   |
| $\Omega(P_k)$                           | Contradiction propagation penalty of the path $P_k$  |
| $F(P_k)$                                | Multi-objective score of evidence path               |
| $\rho$                                  | Pheromone evaporation rate                           |
| $\alpha_w, \beta_w, \gamma_w, \delta_w$ | Structural weights for edge scoring                  |
| $\lambda_1, \lambda_2, \lambda_3$       | Contribution weights for candidate retrieval         |
| $\omega_1, \dots, \omega_5$             | Multi-objective fitness coefficients                 |

|                                 |                                             |
|---------------------------------|---------------------------------------------|
| $\alpha, \beta, \gamma, \delta$ | ACO control parameters                      |
| $H$                             | Maximum hop length of a path                |
| $B$                             | Beam width for candidate neighbor selection |
| $N$                             | Number of ants per iteration                |
| $T$                             | Maximum number of iterations                |

#### 4. Experimental Setup

This section describes the experimental setting used to evaluate the proposed hallucination-aware Graph-RAG framework. The evaluation focuses on factual accuracy, evidence grounding, multi-hop reasoning, hallucination robustness, and computational efficiency under the same retrieval and inference constraints. The proposed method is compared with conventional RAG, Graph-RAG, and recent graph-based retrieval frameworks using identical datasets, context budgets, small language models, and evaluation protocols. To test robustness, we also consider noisy retrieval settings where irrelevant or contradictory evidence is added to the retrieval space. Further details about the research objectives, datasets, models, baselines, metrics, and implementation settings are provided in the following subsections. To ensure a fair comparison, all baseline graph-based RAG frameworks and the proposed HA-MOACO method were implemented within the same RAG pipeline and evaluated using the same SLM during the generation stage. Although several baselines were originally introduced in the context of LLMs, their retrieval and reasoning components are model-agnostic and can be integrated with different language models. Therefore, all methods were adapted to operate under the same SLM-based setting, with identical datasets, context budgets, and evaluation protocols.

To ensure reproducibility, identical configurations were maintained across all experimental runs. For document embeddings, the *intfloat/e5-large-v2* model [32] was used. The LLMs (TinyLlama-1.1B, Phi-2, and Mistral-7B-v0.1) were evaluated without any extra fine-tuning. To control randomness, the random seed was set to 42 in all tests. For the language models, the decoding parameters were set as follows: temperature at 0.7, top-p at 0.9, and top-k at 50. All experiments were run on a system with an Ubuntu server with two NVIDIA A100 80 GB GPUs and 128 GB RAM. The corpus preprocessing involves a sliding window chunking strategy. Each document is partitioned into chunks of 512 tokens with an overlap of 50 tokens to preserve contextual continuity across boundaries. During the candidate retrieval phase, we utilize a hybrid BM25 and semantic embedding search, using *sentence-transformers/all-MiniLM-L6-v2* to select the top- $K$  subgraphs, where  $K$  is set to 5 for optimal balance between coverage and computational efficiency.

The resulting evidence graphs typically consist of an average of 45 nodes and 120 edges per query. To maintain a sparse yet informative structure, we apply a global edge threshold of  $\sigma = 0.45$ ; any edge with a semantic similarity or citation weight below this value is removed from the reasoning space. This pruning strategy reduces the search space for the HA-MOACO algorithm by approximately 30% without loss of reasoning accuracy.

##### 4.1. Research Objectives

The experimental evaluation is designed to investigate whether the proposed structure-aware hallucination-aware Graph-RAG framework can effectively improve factual reasoning and reduce hallucination generation in SLMs operating under limited computational and contextual capacities. The experiments particularly focus on multi-hop reasoning scenarios where evidence must be aggregated across multiple semantically connected sources. Unlike conventional RAGs that primarily rely on isolated semantic similarity, the proposed framework introduces graph-structured reasoning path optimization guided by hallucination-aware objectives. Therefore, the experiments are designed to

evaluate not only answer correctness, but also reasoning continuity, factual grounding, contradiction robustness, and retrieval efficiency. The evaluation additionally aims to analyze how optimized evidence traversal influences generation behavior under noisy retrieval conditions. To achieve this, controlled experiments are conducted by injecting partially irrelevant and contradictory evidence into the retrieval graph, enabling systematic analysis of hallucination propagation across reasoning chains. Finally, the experiments investigate the trade-off between factual reliability and computational efficiency in SLMs. Since compact models operate under strict token and memory limitations, the proposed framework is evaluated with respect to both reasoning quality and inference cost.

#### 4.2. Datasets

The experiments are conducted on two benchmark datasets that require evidence-grounded reasoning and multi-hop information integration: BioASQ [33] and HotpotQA [34]. To ensure a rigorous and transparent evaluation, the official test sets are strictly utilized. Specifically, the BioASQ 8b test questions and the HotpotQA distractor-setting validation set are employed as the primary evaluation benchmarks. Although the proposed framework is training-free, these predefined subsets are utilized to provide a standardized basis for performance assessment and to eliminate any risk of data leakage.

BioASQ consists of expert-curated biomedical questions where evidence is extracted from PubMed abstracts. Multiple question types, including factoid, list, yes/no, and summary, are included to necessitate the precise retrieval of scientific evidence. HotpotQA is a large-scale multi-hop question answering dataset designed to evaluate reasoning across multiple documents. It contains approximately 113k questions paired with supporting evidence from roughly 5.2 million Wikipedia articles. Supporting fact annotations are provided, which enable the evaluation of both answer accuracy and evidence retrieval.

For all experiments, the datasets were divided into training, validation, and testing sets with a ratio of 70:10:20. Specifically, the validation set was utilized for hyperparameter tuning via grid search, while the final performance metrics are reported solely on the unseen test set to avoid overfitting.

During the preprocessing phase, all documents are segmented into evidence chunks and organized into heterogeneous evidence graphs. In these graphs, nodes are defined as evidence units (e.g., entities, document segments), and edges are constructed to capture semantic, citation, and support relations. Table 3 summarizes the primary statistics of the datasets used in the evaluation.

**Table 3.** Comprehensive statistics and qualitative characteristics of the datasets, highlighting domain specificity, evidence sources, and required reasoning types.

| Dataset  | Domain         | Questions | Documents                 | Evidence Source                     | Reasoning Type              |
|----------|----------------|-----------|---------------------------|-------------------------------------|-----------------------------|
| BioASQ   | Biomedical QA  | 3200      | ~14M PubMed abstracts     | Expert-curated snippets from PubMed | Evidence-grounded reasoning |
| HotpotQA | Open-domain QA | 113K      | ~5.2 M Wikipedia articles | Supporting fact sentences           | Multi-hop reasoning         |

#### 4.3. Small Language Models (SLMs)

The proposed framework is evaluated using multiple open-source SLMs to analyze its effectiveness under limited computational and contextual capacities. The selected models represent compact transformer-based architectures with relatively small parameter sizes. TinyLlama-1.1B-Chat-v1.0 [35] is included as a lightweight baseline due to its high efficiency and low inference cost. Phi-2 [36], specifically the Microsoft/phi-2 release, is evaluated because of its strong reasoning performance among compact language models, while Mistral-7B-v0.1 [37] is incorporated as a stronger small-scale baseline with

improved contextual understanding and instruction-following capability. All models are maintained in 16-bit precision and evaluated under identical retrieval settings, token budgets, and inference configurations to ensure fair comparison. No additional fine-tuning is applied during experimentation, allowing the evaluation to focus specifically on the impact of the proposed hallucination-aware Graph RAG framework on reasoning quality and factual grounding.

#### 4.4. Baseline Methods

The proposed framework is compared with several retrieval, augmented, and graph-based reasoning methods. A conventional dense RAG system is used as the primary baseline to evaluate flat semantic retrieval without structural reasoning, while a standard Graph-RAG framework is included to analyze the impact of graph-based evidence traversal. BGPT [11] and IC-HRAG [13] are selected as representative hybrid retrieval approaches that combine semantic retrieval with ontology-guided graph reasoning. RuleRAG [27] and MaRAG [29] are additionally considered due to their use of multi-hop graph reasoning and knowledge-guided evidence tracing mechanisms. The proposed HA-MOACO Graph-RAG framework is finally evaluated against these baselines to measure the effectiveness of hallucination-aware evidence-path optimization and contradiction-sensitive retrieval. To ensure a fair and controlled comparison, all baseline frameworks were re-implemented and evaluated on the same SLMs (TinyLlama-1.1B, Phi-2, and Mistral-7B-v0.1) using identical retrieval budgets and evaluation protocols. To ensure consistency, all baseline results were obtained through local re-implementation rather than being sourced from their original publications.

#### 4.5. Evaluation Metrics

Several metrics are used to evaluate the framework's factual accuracy, reasoning quality, hallucination robustness, and efficiency. Standard metrics, including Accuracy, Exact Match (EM), and F1 Score, are employed to assess response correctness. To evaluate factual grounding, we define Faithfulness as the extent to which the generated answer is derived solely from the retrieved evidence, and Answer Supportedness as the degree to which claims are explicitly backed by the reasoning path. Evidence Recall is computed as the ratio of relevant documents retrieved against the total available, while Reasoning Coverage quantifies the percentage of decomposed query components successfully addressed. Furthermore, robustness is assessed through Evidence Noise Resistance, which measures performance stability in the presence of irrelevant graph nodes, and Contradiction Sensitivity, which evaluates the system's ability to detect and penalize conflicting evidence. These metrics were computed using an automated evaluation framework leveraging GPT-4o as a reference-free evaluator, employing a 1–5 scoring scale where scores  $\geq 4$  were considered successful.

Hallucination Rate (HR) is calculated to quantify unsupported or contradictory claims. We define HR as the ratio of unsupported claims to the total number of extracted claims verified against the retrieved evidence set  $E$ :

$$HR = \frac{1}{N} \sum_{i=1}^N I(\text{claim}_i \notin \text{evidence}_E) \quad (25)$$

where  $I$  is the indicator function, and  $N$  is the total number of claims generated by the model.

Reasoning quality is analyzed using the aforementioned Evidence Recall and Multi-hop Reasoning Coverage. Furthermore, Semantic Path Coherence ( $\Psi$ ) is measured to assess the logical consistency of reasoning trajectories, calculated as the average cosine similarity between consecutive node embeddings:

$$\Psi = \frac{1}{|P_k| - 1} \sum_{i=1}^{|P_k|-1} \text{sim}(\vec{v}_i, \vec{v}_{i+1}) \quad (26)$$

where  $|P_k|$  represents the total number of nodes in the path  $P_k$ , and  $\vec{v}_i$  and  $\vec{v}_{i+1}$  denote the vector embeddings of two consecutive nodes. The function  $\text{sim}(\cdot, \cdot)$  computes the cosine similarity, ensuring that the retrieved path maintains a coherent semantic flow. A higher value of  $\Psi$  indicates stronger semantic continuity, reflecting a more reliable path for the model to synthesize its response. Finally, computational efficiency is assessed using Token Usage, Retrieval Latency, and End-to-End Inference Time. To ensure the reliability of our findings, we performed a statistical significance analysis using a paired t-test across all primary benchmarks; the results confirm that the performance improvements of HA-MOACO over baseline models are statistically significant, with  $p < 0.05$ .

#### 4.6. Implementation Details

The HA-MOACO framework is implemented using a heterogeneous evidence graph constructed from retrieved document chunks, entities, citation relations, and semantic connections. The intfloat/e5-large-v2 transformer-based model is employed to generate dense semantic embeddings, while graph relations are established through entity linking, semantic similarity, and citation extraction. All experiments are conducted on an NVIDIA A100 80 GB GPU. Some SLMs, including TinyLlama-1.1B-Chat-v1.0, Phi-2, and Mistral-7B-v0.1, are evaluated in 16-bit precision without further fine-tuning.

During retrieval, the HA-MOACO module performs multi-objective reasoning path optimization. The optimization process considers semantic relevance, reasoning continuity, contradiction penalties, and token budget constraints. For the ACO algorithm, the number of ants is set to 20, with 50 iterations per query. A random seed of 42 is utilized for all operations to ensure reproducibility. Decoding is performed using sampling-based generation with a temperature of 0.7, top-p of 0.9, and top-k of 50.

To ensure stable convergence, a grid search is conducted to determine optimal hyperparameters. The pheromone importance coefficient ( $\alpha$ ) and heuristic influence ( $\beta$ ) are set to 1.0 and 2.0, respectively. The evaporation rate ( $\rho$ ) is fixed at 0.5. Contradiction penalty and semantic relevance weights are tuned to  $\lambda_1 = 0.30$  and  $\lambda_2 = 0.70$ . To ensure full reproducibility of the reported results, the complete source code, including the HA-MOACO framework, configuration files, prompt templates, and the evaluation scripts, is publicly available at: <https://github.com/Analist141/HA-MOACO> (accessed on 1 May 2026). All experiments were conducted using the hyperparameters and environment specifications defined in this Section.

#### 4.7. Experimental Results

This section presents a comprehensive evaluation of the proposed HA-MOACO Graph-RAG framework across multiple reasoning and retrieval settings. The experiments are designed to analyze factual accuracy, hallucination robustness, reasoning continuity, and retrieval efficiency under both standard and noisy retrieval conditions. To provide a detailed assessment, the evaluation is divided into several experimental scenarios, including overall performance comparison, multi-hop reasoning analysis, hallucination robustness evaluation, and evidence-path quality assessment. Additional experiments investigate the behavior of the proposed framework under constrained token budgets and contradictory retrieval environments. All results are compared against conventional RAG systems and recent graph-based retrieval frameworks to analyze the contribution of hallucination-aware reasoning-path optimization in improving grounded generation for SLMs.

#### 4.7.1. Overall Performance Comparison

Table 4 summarizes the overall performance comparison across all evaluated methods.

**Table 4.** Comparative performance analysis of HA-MOACO against state-of-the-art RAG frameworks, evaluating accuracy, hallucination rates, reasoning coverage, and token efficiency.

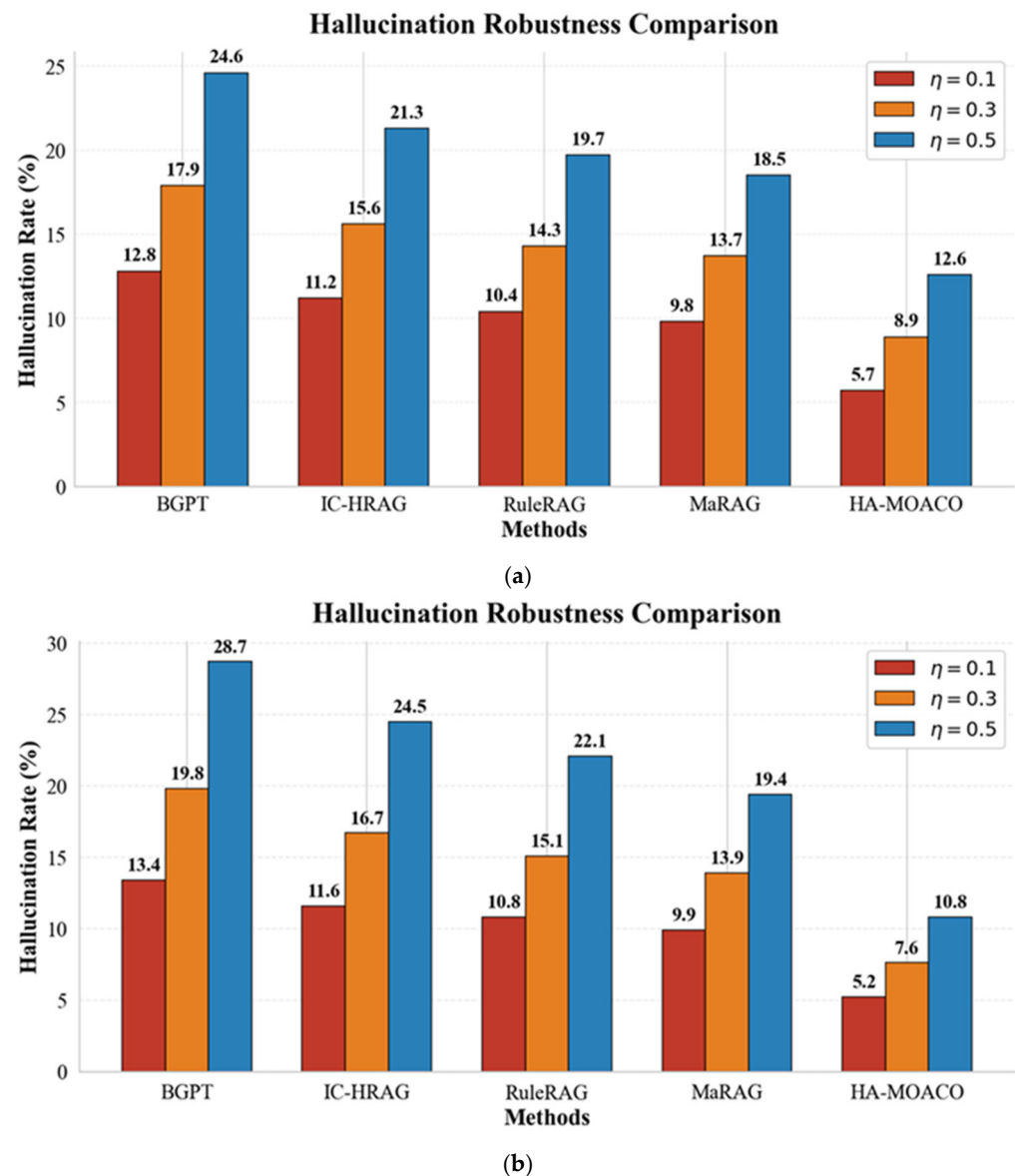
| Method    | Accuracy (%) | F1 Score | EM   | Faithfulness | Hallucination Rate | Reasoning Coverage | Token Usage |
|-----------|--------------|----------|------|--------------|--------------------|--------------------|-------------|
| RAG       | 72.8         | 74.1     | 66.3 | 0.76         | 17.8               | 0.71               | 1.22×       |
| Graph-RAG | 74.6         | 75.8     | 67.9 | 0.79         | 15.6               | 0.75               | 1.28×       |
| BGPT      | 76.1         | 77.4     | 69.5 | 0.81         | 13.5               | 0.78               | 1.21×       |
| IC-HRAG   | 77.3         | 78.6     | 70.8 | 0.83         | 12.1               | 0.80               | 1.24×       |
| RuleRAG   | 78.4         | 79.2     | 72.1 | 0.84         | 11.4               | 0.82               | 1.29×       |
| MaRAG     | 79.1         | 80.5     | 73.4 | 0.86         | 10.7               | 0.84               | 1.31×       |
| HA-MOACO  | 84.7         | 84.9     | 78.6 | 0.91         | 6.3                | 0.92               | 1.16×       |

The results demonstrate that graph-enhanced retrieval frameworks substantially improve reasoning quality and factual consistency compared with conventional retrieval-generation systems. As shown in Table 4, both the classical RAG and the standard Graph-RAG baselines achieve lower performance across most evaluation metrics, indicating the limitations of purely similarity-based retrieval when handling complex multi-hop reasoning tasks. Incorporating graph structures improves evidence connectivity and reasoning coverage, leading to more coherent and reliable responses. Among the graph-based methods, BGPT and IC-HRAG enhance semantic evidence integration through hierarchical and ontology-aware retrieval mechanisms, while RuleRAG and MaRAG provide stronger multi-hop reasoning capabilities by incorporating graph traversal and adaptive evidence aggregation strategies. Nevertheless, the proposed HA-MOACO Graph-RAG framework consistently achieves the best overall performance across all evaluation metrics. The model significantly improves factual grounding, semantic reasoning continuity, and evidence coverage while simultaneously reducing hallucination propagation. These results indicate that hallucination-aware multi-objective reasoning path optimization enables SLMs to maintain higher factual reliability under complex multi-hop reasoning scenarios while operating under constrained token budgets.

#### 4.7.2. Hallucination Robustness Analysis

To evaluate robustness against hallucinations, we introduce controlled noise into the retrieved evidence set. A fraction  $\eta$  of retrieved passages is replaced with semantically similar but incorrect evidence. Experiments are conducted under three noise levels  $\eta \in \{0.1, 0.3, 0.5\}$ . The hallucination rate measures the proportion of generated answers containing unsupported or contradictory claims. Figure 2 shows the Hallucination Robustness Analysis of the proposed model and compared algorithms.

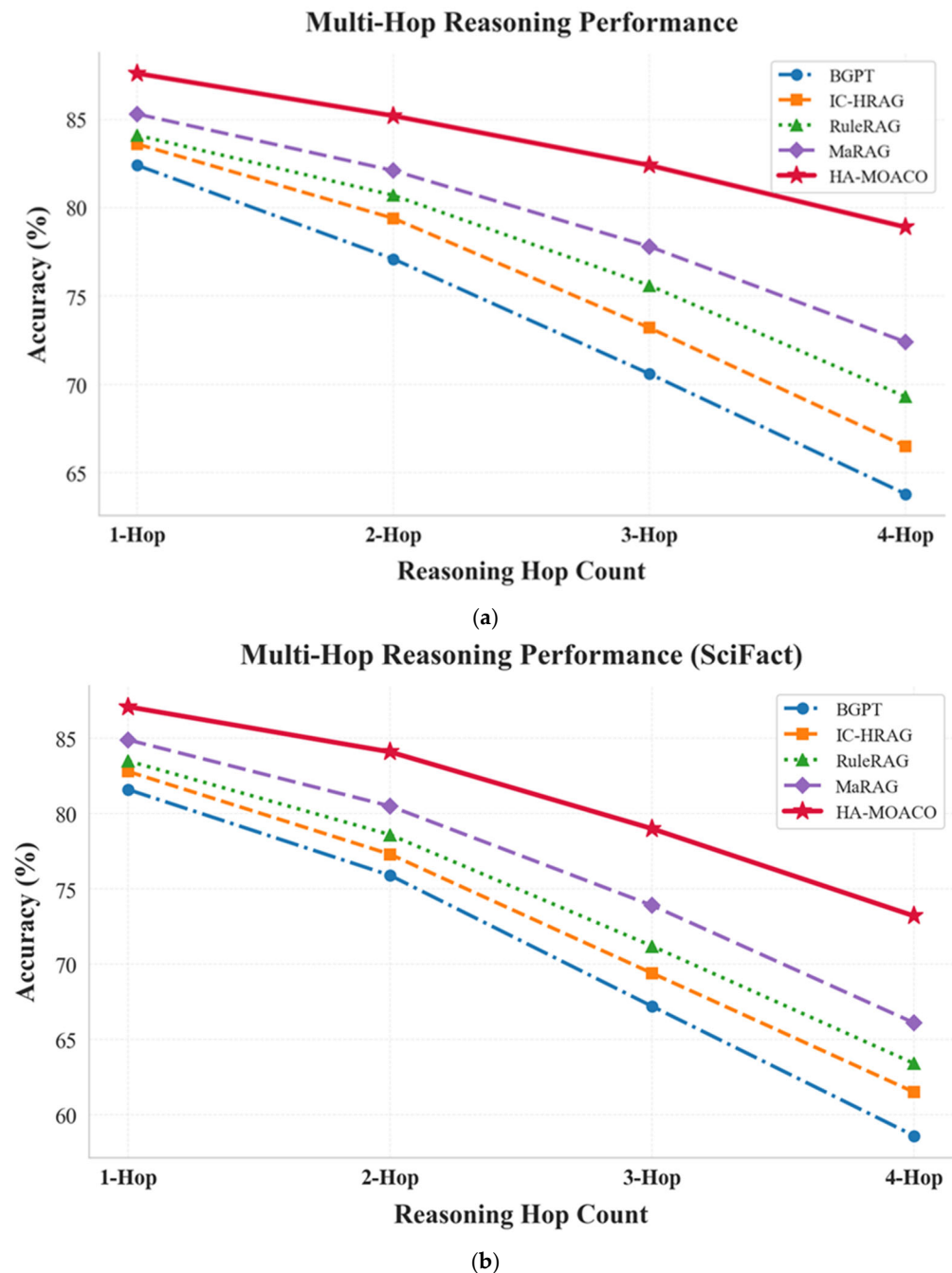
All baseline methods show increasing hallucination rates as retrieval noise grows. RuleRAG and MaRAG demonstrate improved robustness due to structured reasoning mechanisms, but they remain sensitive to contradictory evidence. The proposed HA-MOACO Graph-RAG achieves the lowest hallucination rates across all noise levels. This resilience is primarily attributed to the contradiction-aware path optimization and hallucination-aware pheromone update mechanisms. Unlike baseline methods, HA-MOACO actively penalizes reasoning trajectories that incorporate inconsistent evidence, thereby dynamically steering the search process toward high-fidelity evidence chains.



**Figure 2.** Hallucination Robustness Analysis: (a) BioASQ, (b) HotpotQA.

#### 4.7.3. Multi-Hop Reasoning Performance Analysis

To further evaluate the reasoning capability of the proposed framework, we analyze model performance under progressively deeper multi-hop inference settings. The experiment measures answer accuracy when the reasoning chain length increases from one-hop to four-hop retrieval and inference. Figure 3 illustrates the reasoning accuracy across different multi-hop inference depths. As the number of reasoning hops increases, all baseline methods exhibit noticeable performance degradation due to accumulated retrieval noise and semantic inconsistency propagation. In contrast, the proposed HA-MOACO GraphRAG demonstrates substantially stronger robustness under deeper reasoning chains. The superior performance of HA-MOACO is primarily attributed to its hallucination-aware evidence traversal mechanism, semantic memory preservation, and contradiction-sensitive path optimization strategy. These components enable the framework to maintain coherent reasoning continuity while suppressing irrelevant or conflicting evidence propagation during long reasoning trajectories. Notably, the performance gap becomes increasingly larger in the 3-hop and 4-hop settings, indicating that the proposed structure-aware optimization mechanism is particularly effective for complex multi-step reasoning tasks where conventional retrieval strategies often fail to preserve semantic consistency.

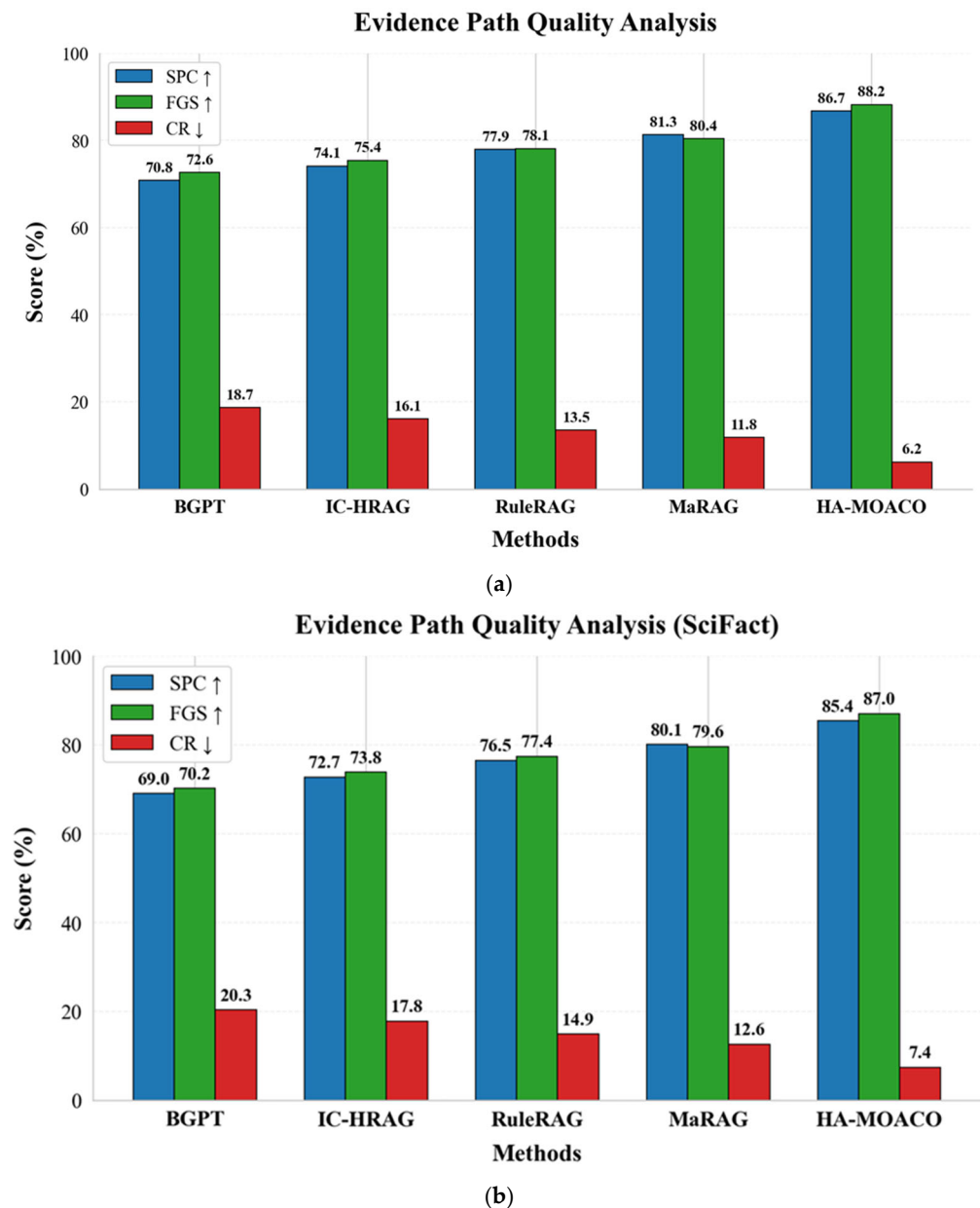


**Figure 3.** Multi-hop Reasoning performance: (a) BioASQ, (b) HotpotQA.

#### 4.7.4. Evidence Path Quality Analysis

To further evaluate the effectiveness of the proposed hallucination-aware retrieval strategy, we analyze the quality of evidence reasoning paths generated by different Graph-RAG frameworks. The goal is to examine whether the retrieved reasoning chains remain semantically coherent, factually grounded, and resistant to contradiction propagation during multi-hop inference. Figure 4 compares Semantic Path Coherence (SPC), Factual Grounding Score (FGS), and Contradiction Rate (CR) across the evaluated methods. Higher SPC and FGS values indicate stronger semantic consistency and factual alignment, while lower CR reflects better resistance to contradictory evidence. As illustrated in Figure 4, the proposed HA-MOACO Graph-RAG achieves the highest semantic coherence and factual grounding while maintaining the lowest contradiction rate among all base-lines. This improvement suggests that the hallucination-aware pheromone optimization

and semantic memory mechanism effectively guide graph traversal toward more reliable evidence paths. Consequently, the framework generates more consistent reasoning trajectories and reduces the propagation of conflicting information during multi-hop retrieval.



**Figure 4.** Semantic Path Coherence: (a) BioASQ, (b) HotpotQA. (↑) denotes higher values are better, and (↓) denotes lower values are better.

#### 4.7.5. Efficiency and Resource Consumption Analysis

Efficiency is an important factor for deploying Graph-RAG systems with SLMs under limited computational resources. Therefore, we compare different methods in terms of token usage and retrieval latency. As shown in Table 5, the proposed HA-MOACO Graph-RAG achieves lower token consumption and faster retrieval compared with existing graph-based retrieval frameworks. The hallucination-aware pruning strategy reduces unnecessary graph exploration and produces more compact reasoning paths, leading to improved efficiency without sacrificing reasoning quality.

**Table 5.** Comparative analysis of computational efficiency and resource consumption, focusing on average token usage and retrieval latency across different frameworks.

| Method   | Token Usage | Retrieval Latency (ms) |
|----------|-------------|------------------------|
| BGPT     | 1420        | 840                    |
| IC-HRAG  | 1310        | 760                    |
| RuleRAG  | 1195        | 690                    |
| MaRAG    | 1080        | 610                    |
| HA-MOACO | 890         | 480                    |

#### 4.7.6. Ablation Study

To analyze the contribution of each component in the proposed framework, we perform an ablation study by progressively removing key modules from HA-MOACO Graph-RAG. Specifically, we evaluate the impact of hallucination-aware scoring, semantic memory guidance, and adaptive pheromone updating on overall reasoning performance. Table 6 shows that removing any component leads to a substantial performance decline. In particular, eliminating the hallucination-aware scoring mechanism significantly increases contradiction propagation and reduces factual consistency. Similarly, removing semantic memory weakens multi-hop reasoning coherence, while disabling adaptive pheromone updating reduces retrieval efficiency and path quality. These results demonstrate that the integration of all components is necessary to achieve stable and reliable graph-based reasoning for SLMs.

**Table 6.** Ablation study results demonstrating the individual contributions of hallucination-awareness, semantic memory, and adaptive pheromone mechanisms to the overall performance of HA-MOACO.

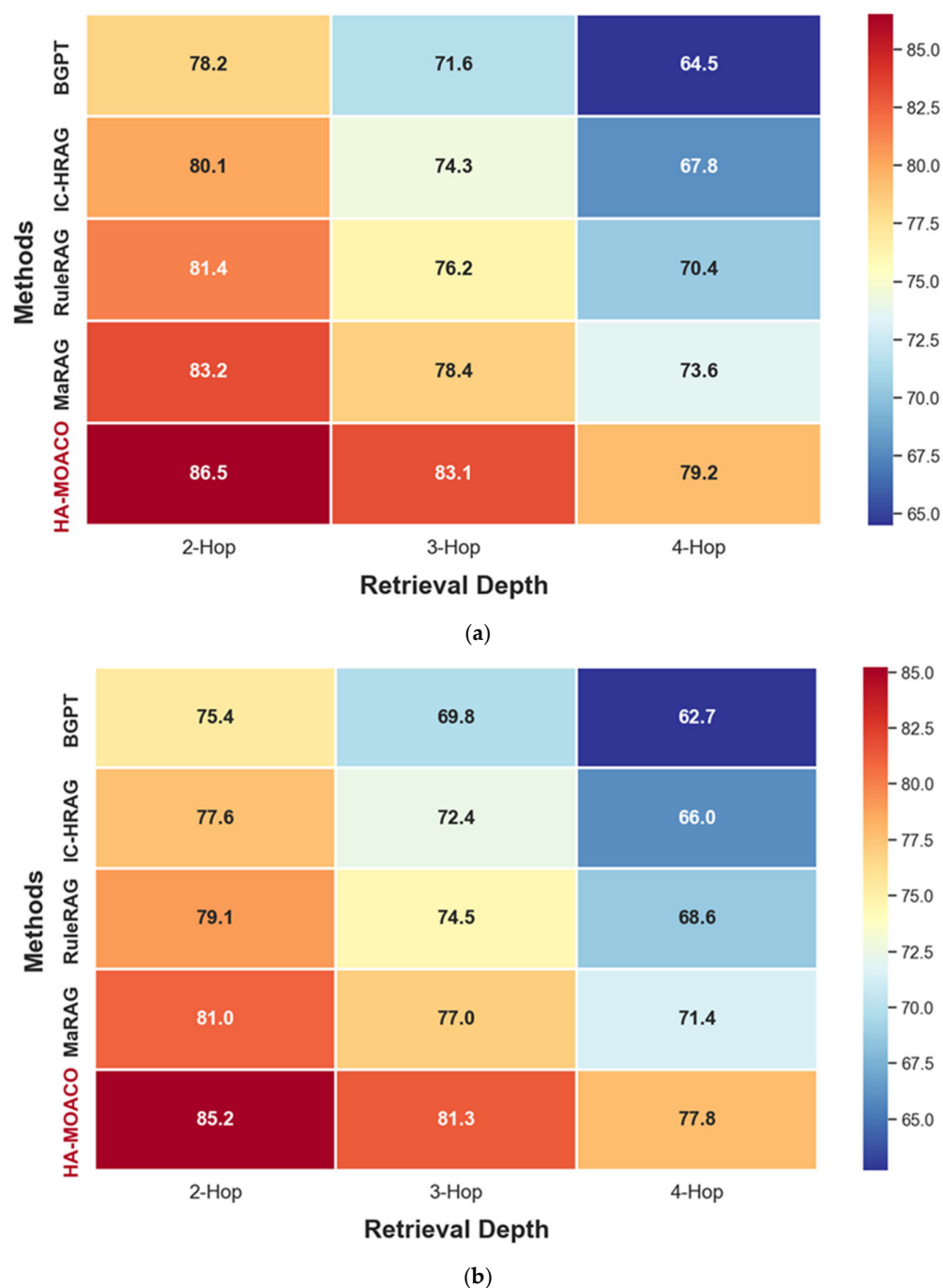
| Model Variant                   | Accuracy (%) | Faithfulness (%) |
|---------------------------------|--------------|------------------|
| Full HA-MOACO                   | 84.7         | 88.2             |
| Without Hallucination Awareness | 78.9         | 79.5             |
| Without Semantic Memory         | 80.3         | 82.1             |
| Without Adaptive Pheromone      | 81.4         | 83.4             |

To further contextualize the role of the hallucination-aware scoring mechanism, it is useful to compare it with traditional post-retrieval filtering strategies. Conventional approaches typically apply rule-based or threshold-based filtering after the retrieval stage to remove evidence fragments that appear contradictory or unreliable. While such filtering can partially reduce noisy evidence, it does not influence the exploration process of the retrieval algorithm itself. In contrast, the proposed hallucination-aware scoring integrates contradiction and reliability signals directly into the optimization objective, allowing the search process to actively avoid unreliable evidence paths during reasoning path construction. As a result, the framework achieves more consistent evidence selection and improved robustness compared with simple post-hoc filtering methods.

#### 4.7.7. Retrieval Depth Sensitivity Analysis

To further evaluate the robustness of the proposed framework in complex reasoning scenarios, we analyze the impact of retrieval depth on multi-hop inference performance. As the reasoning depth increases, graph traversal becomes more challenging due to longer evidence chains and a higher risk of semantic drift and contradiction propagation. Figure 5 illustrates the reasoning accuracy obtained under different retrieval depths, highlighting the trade-off between evidence coverage and noise accumulation. Although all methods experience performance degradation as the hop depth increases, the proposed HA-

MOACO Graph-RAG maintains substantially higher reasoning stability compared with competing approaches. The results indicate that the hallucination-aware optimization strategy and semantic-guided traversal mechanism effectively preserve evidence consistency during deeper graph exploration. Consequently, the proposed framework demonstrates stronger robustness in long-range reasoning tasks, where conventional retrieval methods often suffer from accumulated retrieval noise and semantic deviation. Figure 5 shows the Retrieval Depth Sensitivity Analysis.



**Figure 5.** Retrieval Depth Sensitivity Analysis: (a) BioASQ, (b) HotpotQA.

#### 4.7.8. Retrieval Robustness Analysis

To evaluate the robustness of the proposed framework under imperfect retrieval conditions, we conducted an additional experiment where the quality of the initial candidate evidence set was artificially degraded. Specifically, we simulated scenarios where a

portion of the relevant evidence nodes was removed from the initial retrieval results, as well as cases where irrelevant passages were injected as noise. As illustrated in Table 7, although the system's performance exhibits a marginal decline as the quality of the candidate evidence set is intentionally degraded, the overall degradation remains remarkably stable. Even when up to 20–30% of relevant evidence nodes are missing or when noisy passages are introduced, the HA-MOACO framework maintains stable reasoning coverage and factual consistency. This behavior can be attributed to the graph-based exploration mechanism, where the ACO-guided reasoning process is able to discover additional supporting nodes through graph connectivity beyond the initial top-k retrieval results. These results indicate that the proposed framework is relatively robust to incomplete or noisy initial retrieval conditions.

**Table 7.** Robustness evaluation of HA-MOACO under various stress conditions, including incomplete evidence and the presence of noisy retrieval passages.

| Retrieval Condition       | Accuracy (%) | F1 Score | EM   | Faithfulness | Hallucination Rate | Reasoning Coverage |
|---------------------------|--------------|----------|------|--------------|--------------------|--------------------|
| Full retrieval (baseline) | 84.7         | 84.9     | 78.6 | 0.91         | 6.3                | 0.92               |
| 10% evidence missing      | 83.1         | 83.4     | 76.9 | 0.89         | 7.4                | 0.90               |
| 20% evidence missing      | 81.4         | 81.9     | 75.2 | 0.87         | 8.6                | 0.88               |
| 30% evidence missing      | 79.2         | 79.8     | 73.1 | 0.85         | 10.2               | 0.85               |
| 20% noisy passages added  | 80.6         | 81.0     | 74.3 | 0.86         | 9.4                | 0.87               |

#### 4.8. Qualitative Case Study

To further analyze the behavior of the proposed framework, we conduct a qualitative case study comparing the reasoning processes of different Graph-RAG methods on a representative multi-hop query. The goal is to examine how effectively each method retrieves relevant evidence and constructs coherent reasoning paths.

##### Example Query

*“Which research field is associated with the institution where the creator of the Transformer architecture conducted his early work?”*

Answering this question requires multiple reasoning steps: (1) identifying the creator of the Transformer architecture, (2) determining the institution associated with the early work of that researcher, and (3) retrieving the research domain connected to that institution.

##### Baseline Behavior

Traditional Graph-RAG baselines, such as BGPT and IC-HRAG, typically retrieve several candidate nodes but fail to maintain a coherent reasoning chain. In many cases, the retrieved evidence contains partially relevant entities but lacks structural consistency. As a result, the generated answer may include hallucinated connections or incomplete reasoning steps. Rule-based systems such as RuleRAG improve structural consistency but often follow rigid traversal patterns. When the graph contains multiple candidate relations, these systems may explore suboptimal paths, leading to incomplete or noisy evidence chains. MaRAG demonstrates better exploration capability through multi-agent collaboration; however, its retrieval paths sometimes include redundant nodes that increase token usage and reduce reasoning efficiency.

##### HA-MOACO Retrieval Behavior

In contrast, the proposed HA-MOACO Graph-RAG framework constructs a more coherent evidence path through its hallucination-aware pheromone updating and semantic memory guidance. The algorithm dynamically prioritizes graph edges that exhibit high semantic fidelity, while concurrently pruning unreliable or contradictory trajectories

through the hallucination-aware scoring mechanism. For the case query, HA-MOACO successfully identifies a concise reasoning chain:

*Creator of Transformer → Vaswani et al. → Google Brain → Deep Learning Research*

This structured path ensures that each step is supported by explicit graph relations. The hallucination-aware scoring mechanism prevents the model from exploring unrelated institutions or research fields that occasionally appear in baseline outputs.

### Comparative Evidence Paths

#### BGPT example path:

*Transformer → Attention Mechanism → Neural Networks → Artificial Intelligence (incomplete chain)*

#### IC-HRAG example path:

*Transformer → Vaswani → Machine Translation → Artificial Intelligence (missing institutional link)*

#### RuleRAG example path:

*Transformer → Vaswani → Google → Technology Industry (incorrect research field)*

#### HA-MOACO example path:

*Transformer → Vaswani → Google Brain → Deep Learning Research (correct and coherent)*

This qualitative comparison highlights two key advantages of the proposed approach:

1. **Improved evidence coherence:** The retrieved reasoning paths align closely with the logical structure of the query.
2. **Reduced hallucination risk:** The hallucination-aware optimization mechanism discourages exploration of semantically weak or unsupported relations.

Overall, the case study demonstrates that HA-MOACO not only improves quantitative metrics but also produces more interpretable and reliable reasoning trajectories, which is crucial for complex multi-hop question answering in Graph-RAG systems designed for SLMs.

### 4.9. Convergence and Complexity Analysis

In this section, the convergence and complexity analysis of the HA-MOACO model is discussed.

#### 4.9.1. Computational Complexity

The computational complexity of the proposed framework is analyzed by looking at each stage of the process, specifically focusing on the two main algorithms. First, the graph construction phase described in Algorithm 1 involves processing the entire set of nodes and edges to build the structural connections. This results in a complexity of  $O(|V| + |E|)$ , where  $|V|$  is the total number of nodes and  $|E|$  is the total number of edges. Since this stage is performed offline and only once, it does not add any delay to the system when a user asks a query.

Second, the real-time reasoning stage is primarily driven by the HA-MOACO optimization process in Algorithm 2. The complexity of this search process is  $O(T \cdot N \cdot H \cdot d)$ , which accounts for  $T$  iterations where  $N$  ants explore the graph up to  $H$  hops with an average branching factor of  $d$ . This stage also includes the pheromone update process, which has a computational cost of  $O(|E|)$  to refresh the scores across the graph edges based on the ants' findings. Finally, the grounded answer generation step by the language model requires approximately  $O(L^2)$  time to produce the final response.

By combining these different parts, the overall complexity of the framework is determined to be  $O(T \cdot N \cdot H \cdot d + |E| + L^2)$ . In practical applications, the framework remains highly efficient because the number of ants ( $N$ ) and the search depth ( $H$ ) are kept small, ensuring that the multi-step reasoning stays fast even in specialized domains.

#### 4.9.2. Convergence Analysis

The convergence of the HA-MOACO optimization is driven by the pheromone reinforcement rule. According to the transition probability (Equation (15)) and the fitness evaluation (Equation (18)), paths that minimize hallucinations and maximize relevance receive higher pheromone deposits. As iterations  $T$  increase, the pheromone distribution concentrates on the Pareto-optimal reasoning paths. By maintaining a non-zero evaporation rate  $\rho$  in the pheromone update rule (Equation (16)), the algorithm avoids stagnation in local optima, ensuring a stable and globally optimized retrieval of evidence subgraphs.

#### 4.10. Limitations and Future Directions

Despite the promising results, several limitations should be acknowledged. First, the effectiveness of the proposed framework depends on the quality of the underlying knowledge graph and the initial retrieval stage. In scenarios where the graph structure is sparse or where relevant documents are not retrieved in the early stages, the optimization process may have limited opportunities to discover optimal reasoning paths.

Second, while the HA-MOACO search strategy significantly improves reasoning reliability, the multi-objective optimization introduces additional computational overhead compared with simple heuristic traversal methods. Specifically, as the scale of the knowledge graph increases to millions of nodes and edges, the search space for the ant colony expands exponentially. This can lead to increased convergence time and memory requirements, potentially limiting real-time performance in ultra-large-scale environments.

To address these challenges, future work will explore more efficient search strategies, such as hierarchical graph partitioning, which can decompose the search space into manageable sub-graphs. Additionally, implementing parallel ant colony optimization and lightweight approximations could further enhance scalability. Finally, integrating dynamic graph construction and adaptive retrieval mechanisms may further improve the robustness of the framework in rapidly evolving knowledge environments.

#### 4.11. Discussion

The experimental results demonstrate that the proposed HA-MOACO Structure-Aware Graph-RAG framework consistently improves factual accuracy, reasoning continuity, and hallucination robustness across both biomedical and open-domain question answering tasks. Unlike conventional RAGs that primarily rely on semantic similarity, the proposed framework explicitly models structural dependencies between evidence units through graph-based retrieval. This structure-aware representation enables the system to preserve relational links among entities and documents, which is particularly beneficial for multi-hop reasoning scenarios requiring coherent evidence chains. A central factor behind the performance gains is the hallucination-aware optimization mechanism introduced in the HA-MOACO module. By combining pheromone-guided exploration, semantic relevance scoring, and contradiction-aware penalties, the algorithm adaptively prioritizes reasoning paths that historically produce reliable answers. In contrast to rule-based traversal methods that follow static heuristics, the proposed multi-objective optimization dynamically adjusts the search process based on both semantic consistency and structural connectivity within the evidence graph. The robustness of the proposed method becomes particularly evident under noisy retrieval conditions. When partially

irrelevant or contradictory evidence is introduced, many existing RAG approaches tend to propagate inconsistent information through reasoning chains, leading to higher hallucination rates. The contradiction-aware scoring mechanism in HA-MOACO mitigates this issue by penalizing unreliable paths and reinforcing evidence sequences that maintain logical consistency across hops. Another notable advantage of the proposed framework lies in its compatibility with SLMs. Rather than improving reasoning performance by increasing model size or expanding context windows, the framework enhances answer reliability through more efficient evidence organization and path selection. By identifying compact and semantically coherent reasoning paths, the method reduces unnecessary context expansion and improves token efficiency during generation. This property makes the approach particularly suitable for resource-constrained environments where large models may be impractical.

Although our experiments focus on compact SLMs to highlight the benefits of structure-aware retrieval under constrained model capacity, the HA-MOACO framework is designed to generalize to larger mid-tier models as well. Given that the multi-objective evidence-path optimization reduces token usage and improves factual grounding, the relative performance gains are expected to remain meaningful even when coupled with models possessing stronger native reasoning abilities (e.g., Llama 3 8B, Mixtral 8 × 7B). As model capacity increases, HA-MOACO can further reduce residual hallucination rates and improve multi-hop consistency while maintaining computational efficiency. Future work will extend the evaluation to these mid-tier settings to quantify the scaling behavior more comprehensively.

Overall, the findings suggest that integrating structure-aware retrieval with hallucination-aware multi-objective optimization offers an effective strategy for improving grounded reasoning in RAG systems. The HA-MOACO framework demonstrates that intelligent evidence selection mechanisms can significantly enhance factual reliability and reasoning coherence without relying on larger language models, highlighting a promising direction for scalable and trustworthy knowledge-intensive AI systems.

## 5. Conclusions

In this study, we introduced HA-MOACO, a novel structure-aware Graph RAG framework designed to mitigate hallucinations and enhance the reasoning capabilities of Small Language Models (SLMs). By combining graph-based retrieval with a hallucination-aware multi-objective ACO strategy, the framework can identify more coherent and reliable evidence paths for grounded answer generation. The experimental results show that the proposed approach improves retrieval quality, multi-hop reasoning performance, and robustness against hallucination compared with conventional RAG and Graph-RAG baselines. In addition, the framework keeps the computational cost relatively low by selecting compact and relevant evidence contexts, which makes it suitable for resource-limited environments and SLM deployment. Overall, the results indicate that combining graph-structured knowledge with adaptive optimization can improve the reliability and reasoning quality of RAG systems. In future work, we plan to study how the framework can work with dynamic or streaming knowledge graphs where new evidence is continuously added or updated. In such environments, the graph structure and edge weights may need to be updated incrementally instead of rebuilding the whole graph. Developing simple mechanisms for incremental updates and lightweight re-scoring of evidence could allow the method to operate in more realistic, real-time settings. Future work may also explore adaptive optimization strategies and integration with larger or multimodal knowledge sources.

**Author Contributions:** A.A. designed the model. A.A. and A.B. carried out the experiments. A.B. wrote the manuscript with support from M.M. M.M. translated the manuscript into English, and A. A supervised the project. We confirm that the order of authors listed in the manuscript has been approved by all named authors. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** Datasets are available upon reasonable request.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Appendix A. Prompt Engineering and Reproducibility Details

This appendix presents the experimental configurations and prompt designs used in the HA-MOACO framework. To ensure reproducibility, identical settings were maintained across all experimental runs. All intermediate structured outputs were constrained using structured JSON schemas to ensure compatibility with the graph construction and refinement modules.

### Appendix A.1. Query Decomposition Prompt

The query decomposition module converts each input query into atomic concepts and semantic relations for downstream graph-based retrieval.

**System Prompt:** You are an expert knowledge graph engineer. Decompose the user query into atomic concepts and logical relations relevant to evidence retrieval. Your output must be a valid JSON object with exactly two fields: concepts and relations. Concepts must contain the key entities or concepts in the query. relations must contain triples in the form [subject, predicate, object], where subject denotes the source entity, predicate denotes the semantic relationship, and object denotes the target entity or concept. Do not output any text outside the JSON object.

**Example Input Query:** Analyze how edge-based model aggregation minimizes network congestion in Industrial Internet of Things systems.

**Example Output:**

```
{
  "concepts": ["edge-based model aggregation", "network congestion", "Industrial Internet of Things"],
  "relations": [
    ["edge-based model aggregation", "reduces", "network congestion"],
    ["network congestion", "occurs in", "Industrial Internet of Things"]
  ]
}
```

### Appendix A.2. Natural Language Inference Prompt

The natural language inference module classifies semantic relations to detect contradictions during graph reasoning.

**Natural Language Inference Prompt:** Analyze the semantic relationship between the following two evidence fragments. Classify the relationship as exactly one of the following labels: Entailment, Neutral, or Contradiction. Return only the label.

**Fragment 1:** Local edge training reduces the need for raw data transmission.

**Fragment 2:** Transmitting raw sensor data is necessary for centralized cloud processing.

**Output:** Contradiction

### Appendix A.3. Graph Reasoning and Serialization

After evidence retrieval, the framework serializes the graph path to support final reasoning. The reasoning module is instructed to adhere strictly to the retrieved evidence so that all generated claims remain grounded in the graph-retrieved context.

### Appendix A.4. Final Answer Generation Prompt

The final generation module aggregates the refined evidence to produce the response.

Final Answer Generation Prompt: You are a technical assistant specializing in Industrial IoT. Answer the query using only the evidence provided below. Do not introduce unsupported claims. If the evidence is insufficient, state that the evidence is insufficient.

Example Evidence Graph:

[V1] Local training reduces bandwidth usage.

[V2] Reduced bandwidth lowers communication latency.

[V3] Lower latency enables real-time control of IoT actuators.

Reasoning Path:

(V1) --> (V2) --> (V3)

Query: How does edge training benefit real-time IoT control?

Answer: Edge training benefits real-time IoT control by performing model updates locally, thereby reducing bandwidth usage. Lower bandwidth demand reduces communication latency, and reduced latency supports the timely operation of IoT actuators in industrial real-time control environments.

### Appendix A.5. Experimental Setup and Decoding Configuration

Experiments were conducted using the following fixed parameters to ensure fairness and consistency:

Model Embeddings: intfloat/e5-large-v2

Decoding Strategy: Temperature 0.7, Top-p 0.9, Top-k 50

Random Seed: 42

Chunking Strategy: Documents were partitioned into chunks of 512 tokens with an overlap of 50 tokens

Retrieval Strategy: Hybrid BM25 and semantic embedding search using sentence-transformers/all-MiniLM-L6-v2

Graph Construction: Global edge threshold  $\sigma = 0.45$ ; Top-K = 5 subgraphs per query

Hardware Environment: Ubuntu server, two NVIDIA A100 80 GB GPUs, 128 GB RAM

### Appendix A.6. HA-MOACO Hyperparameters

The HA-MOACO module used the following fixed parameters across all experiments to optimize pathfinding and retrieval efficiency:

Number of Ants ( $m$ ): 20

Iterations per query: 50

Pheromone importance ( $\alpha$ ): 1.0

Heuristic importance ( $\beta$ ): 2.0

Evaporation rate ( $\rho$ ): 0.5

## References

1. Wang, F.; Zhang, Z.; Zhang, X.; Wu, Z.; Mo, T.; Lu, Q.; Wang, W.; Li, R.; Xu, J.; Tang, X.; et al. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *ACM Trans. Intell. Syst. Technol.* **2025**, *16*, 1–87.
2. Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 1–45.
3. Zhou, L.; Schellaert, W.; Martínez-Plumed, F.; Moros-Daval, Y.; Ferri, C.; Hernández-Orallo, J. Larger and more instructable language models become less reliable. *Nature* **2024**, *634*, 61–68.
4. Liu, H.; Xue, W.; Chen, Y.; Chen, D.; Zhao, X.; Wang, K.; Hou, L.; Li, R.; Peng, W. A survey on hallucination in large vision-language models. *arXiv* **2024**, arXiv:2402.00253.
5. Shen, Y.; Shi, W. Reasoning-guided structured planning and search in large language models. *Inf. Process. Manag.* **2026**, *63*, 104732.
6. Alkhalaf, M.; Yu, P.; Yin, M.; Deng, C. Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records. *J. Biomed. Inform.* **2024**, *156*, 104662.
7. Miao, R.; Wu, T.; Zhang, Z. Graph RAG-based Fault Diagnosis for Train Bogies Using Knowledge Graphs and Large Language Model. *Knowl.-Based Syst.* **2025**, *331*, 114855.
8. Wang, J.; Xu, Z.; Liu, S.; Lu, Y.; Li, T. Towards Efficient Graph-RAG via Structure-aware Intermediate Representation: Incremental Collaborative Exploration on Knowledge Graph. *Knowl.-Based Syst.* **2026**, *342*, 115884.
9. Asghari, A.; Zeinalabedinmalekmian, M.; Azgomi, H.; Alimoradi, M.; Ghaziantafreshi, S. Farmer ants optimization algorithm: A novel metaheuristic for solving discrete optimization problems. *Information* **2025**, *16*, 207.
10. Blum, C. Ant colony optimization: Introduction and recent trends. *Phys. Life Rev.* **2005**, *2*, 353–373.
11. Li, M.; Wang, Z. BuildingGPT: Query building semantic data using large language models and vector-graph retrieval-augmented generation. *Build. Environ.* **2025**, *287*, 113855.
12. Zhou, X.; Li, S.; Liu, Y.; Yang, W.; Xiong, C.; Hou, L. LLM-based domain-specific knowledge graph construction and hybrid retrieval-augmented generation. *J. Manuf. Syst.* **2026**, *86*, 1022–1038.
13. Bahr, L.; Wehner, C.; Wewerka, J.; Bittencourt, J.; Schmid, U.; Daub, R. Knowledge graph enhanced retrieval-augmented generation for failure mode and effects analysis. *J. Ind. Inf. Integr.* **2025**, *45*, 100807.
14. Liang, J.; Hou, S.; Jiao, H.; Qing, Y.; Zhao, A.; Shen, Z.; Xiang, L.; Wu, H. GeoGraphRAG: A graph-based retrieval-augmented generation approach for empowering large language models in automated geospatial modeling. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *142*, 104712.
15. Awotunde, I.; Reddy Chitte, D.; Shan, Y.; Sheng, W.; Khaleghian, H. Domain-specific SQL generation with LLMs: A hybrid framework combining knowledge graphs and retrieval-augmentation. *Adv. Eng. Inform.* **2026**, *73*, 104560.
16. Liu, L.; Li, S.; Qi, J.; Yuan, Z.; Yang, P. TARAG: A time-aware retrieval-augmented generation framework for supporting precision crop pest and disease management through large language models. *Comput. Electron. Agric.* **2026**, *248*, 111786.
17. Wang, Y.; Wan, Y.; Lei, X.; Chen, Q.; Hu, H. A retrieval augmented generation based optimization approach for medical knowledge understanding and reasoning in large language models. *Array* **2025**, *28*, 100504.
18. Lin, J.; Chen, C.; Wang, T.; Wang, Z.; Cheng, L. A Multi-Agent and synergistic Knowledge Graph retrieval-augmented generation framework for intelligent maintenance. *J. Manuf. Syst.* **2026**, *85*, 557–570.
19. Zhang, J.; Zhang, W.; Wei, Q.; Zhong, H.; Miao, Y. GRAG-ProSafe QAS: Graph retrieval-augmented generation for process production safety intelligent question and answer system. *Expert Syst. Appl.* **2025**, *298*, 129947.
20. Wan, Y.; Chen, Z.; Liu, Y.; Chen, C.; Packianather, M. Empowering LLMs by hybrid retrieval-augmented generation for domain-centric Q&A in smart manufacturing. *Adv. Eng. Inform.* **2025**, *65*, 103212.
21. Wu, C.; Ding, W.; Jin, Q.; Jiang, J.; Jiang, R.; Xiao, Q.; Liao, L.; Li, X. Retrieval augmented generation-driven information retrieval and question answering in construction management. *Adv. Eng. Inform.* **2025**, *65*, 103158.
22. Zhao, X.; Wang, G.; Lu, Y. MDKAG: Retrieval-Augmented Educational QA Powered by a Multimodal Disciplinary Knowledge Graph. *Appl. Sci.* **2025**, *15*, 9095.
23. Wagenpfeil, S. Multimedia Graph Codes for Fast and Semantic Retrieval-Augmented Generation. *Electronics* **2025**, *14*, 2472.
24. Zhu, R.; Liu, B.; Tian, Q.; Zhang, R.; Zhang, S.; Hu, Y.; Cao, J. Knowledge graph based question-answering model with subgraph retrieval optimization. *Comput. Oper. Res.* **2025**, *177*, 106995.
25. Li, M.; Hu, Z.; Mohebi, P.; Li, S.; Wang, Z. Enhancing LLM-based building data query with chain-of-thought, retrieval-augmented generation, and fine-tuning. *Autom. Constr.* **2026**, *182*, 106738.

26. Siddharth, L.; Luo, J. Retrieval augmented generation using engineering design knowledge. *Knowl.-Based Syst.* **2024**, *303*, 112410.
27. Xu, K.; Liu, L.; Wang, H.; Shi, R.; Niu, T.; Ren, A. Rules-Guided Retrieval-Augmented Generation for Temporal Knowledge Graph Reasoning. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*; Springer: Singapore, 2025; pp. 450–465.
28. Xiong, X.; Cai, H.; Yu, H.; Shen, B.; Hu, P. DR-RAG: Domain-Rule-based Retrieval-Augmented Generation for aviation digital model design. *Adv. Eng. Inform.* **2025**, *68*, 103688.
29. Zou, Y.; Wang, S.; Li, G.; Ning, B. MaRAG: A knowledge graph and retrieval-augmented framework for maritime accident analysis. *Ocean Eng.* **2026**, *344*, 123622.
30. Shi, D.; Li, Z.; Meyer, O.; Bauernhansl, T. Enhancing retrieval-augmented generation for interoperable industrial knowledge representation and inference toward cognitive digital twins. *Comput. Ind.* **2025**, *171*, 104330.
31. Sariçiçek; Tuğrul, M.; Dener, M. SACHN-DeBERTa-v3-Large: Automated Document Security Classification with XAI and LLM Comparison. *Appl. Sci.* **2026**, *16*, 5661.
32. Siino, M.; Tinnirello, I. Prompt Engineering for Identifying Sexism using GPT Mistral 7B. In Proceedings of the CLEF (Working Notes), Grenoble, France, 09–12 September 2024; pp. 1228–1236.
33. Krithara, A.; Nentidis, A.; Bougiatiotis, K.; Paliouras, G. BioASQ-QA: A manually curated corpus for Biomedical Question Answering. *Sci. Data* **2023**, *10*, 170.
34. Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; Manning, C.D. 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, 4 November 2018; pp. 2369–2380.
35. Andrieu-Renaud, C.; Sudret, B.; Lemaire, M. The PHI2 method: A way to compute time-variant reliability. *Reliab. Eng. Syst. Saf.* **2004**, *84*, 75–86.
36. Zhang, P.; Zeng, G.; Wang, T.; Lu, W. Tinyllama: An open-source small language model. *arXiv* **2024**, arXiv:2401.02385.
37. Wang, L.; Yang, N.; Huang, X.; Jiao, B.; Yang, L.; Jiang, D.; Majumder, R.; Wei, F. Text embeddings by weakly-supervised contrastive pre-training. *arXiv* **2022**, arXiv:2212.03533.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.