SCOUT-RAG: Scalable and Cost-Efficient Unifying Traversal for Agentic Graph-RAG over Distributed Domains

Longkun Li[1*], **Yuanben Zou**[2,*], **Jinghan Wu**[3,*], **Yuqing Wen**[4,*]
Jing Li[5, 6†] **Hangwei Qian**[5,6], **Ivor Tsang**[5, 6]

[1]Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[2]College of Design and Engineering, National University of Singapore, Singapore
[3]Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, China
[4]Faculty of Science, National University of Singapore, Singapore
[5]Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore
[6]Centre for Frontier AI Research, Agency for Science, Technology and Research, Singapore
{e1324260, e1142979, e1351400}@u.nus.edu, wujing4022@stu.xjtu.edu.cn,
{li_jing, qian_hangwei, ivor_tsang}@a-star.edu.sg

Abstract

Graph-RAG improves LLM reasoning using structured knowledge, yet conventional designs rely on a centralized knowledge graph. In distributed and access-restricted settings (e.g., hospitals or multinational organizations), retrieval must select relevant domains and appropriate traversal depth without global graph visibility or exhaustive querying. To address this challenge, we introduce **SCOUT-RAG** (*Scalable and COst-efficient Unifying Traversal*), a distributed agentic Graph-RAG framework that performs progressive cross-domain retrieval guided by incremental utility goals. SCOUT-RAG employs four cooperative agents that: (i) estimate domain relevance, (ii) decide when to expand retrieval to additional domains, (iii) adapt traversal depth to avoid unnecessary graph exploration, and (iv) synthesize the high-quality answers. The framework is designed to minimize retrieval regret, defined as missing useful domain information, while controlling latency and API cost. Across multi-domain knowledge settings, SCOUT-RAG achieves performance comparable to centralized baselines, including DRIFT and exhaustive domain traversal, while substantially reducing cross-domain calls, total tokens processed, and latency.

Introduction

RAG (Retrieval-Augmented Generation) (Lewis et al. 2020) enhances large language models by retrieving and integrating external knowledge for grounded response generation. Traditional deployments primarily rely on vector-based retrieval, where individually relevant text segments are retrieved based on similarity signals. Yet many real-world queries demand global sense-making, synthesizing relationships, dependencies, and abstractions across entities rather than isolating local matches. To address this gap, Graph-RAG systems (Edge et al. 2024) have emerged, enabling multi-hop traversal, hierarchical abstraction, and structured knowledge integration. These capabilities have led major industry platforms, including Microsoft, NebulaGraph, AntGroup, and Neo4j (Peng et al. 2024), to adopt graph-centric retrieval pipelines, underscoring the emerging role of structured knowledge access in complex information environments.

However, most RAG and Graph-RAG frameworks assume centralized access to a unified knowledge store. In practice, data often resides across independent organizations that cannot share raw content due to privacy, ownership, or regulatory constraints, motivating retrieval across distributed and siloed domains rather than a single graph. This setting has been formalized as multi-domain or federated RAG (Shojaee et al. 2025), where each domain retains data control and responses must respect access boundaries. Further, each cross-domain request may incur latency or monetary costs, requiring systems that minimize unnecessary retrieval while preserving answer fidelity. Recent work uses supervised domain routers (Shojaee et al. 2025; Guerraoui et al. 2025) to filter domains, but these approaches depend on labeled query–domain pairs and do not adapt to evolving utility signals or cold-start deployments.

Distributed Graph-RAG features partial observability, heterogeneous access constraints, and incremental evidence accumulation, which collectively motivate an **agentic** formulation. Rather than predetermining a fixed retrieval plan, the system must treat cross-domain traversal as a sequential decision process: evaluating intermediate signals, updating beliefs about domain relevance, and allocating retrieval budget adaptively across local exploration and global expansion. This perspective naturally accommodates privacy constraints, variable query structure, and shifting domain utility without requiring labeled access patterns or centralized visibility. By framing retrieval as iterative utility estimation and constrained action selection, an agentic approach supports controlled cost–quality trade-offs and scalable reasoning across distributed knowledge holders.

Building on this perspective, we present **SCOUT-RAG**,

*Work done during an internship at A*STAR.
†Corresponding Author

Copyright © 2026, Trustworthy Agentic AI Workshop@ Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

a **layered agentic architecture for distributed Graph-RAG**, explicitly designed to operate under cost, privacy, and scalability constraints. Our design is grounded in three core insights. **First**, we introduce training-free domain relevance estimation, avoiding supervised routing signals and enabling deployment in cold-start and privacy-restricted settings. **Second**, we generalize Graph-RAG to decentralized knowledge holders, treating each domain as a partially observable subgraph and enabling adaptive switching between local traversal and cross-domain expansion as context evolves. **Third**, motivated by agent-based RAG systems (Singh et al. 2025) and recent agentic Graph-RAG efforts (Lee et al. 2025; Luo et al. 2025), we formalize retrieval as a sequential decision process that incrementally assesses utility, allocates exploration budget, and adjusts traversal depth and breadth under explicit cost constraints. Together, these principles enable scalable, privacy-aware, and cost-efficient cross-domain knowledge aggregation without assuming centralized visibility or requiring supervised domain routing. To the best of our knowledge, **SCOUT-RAG is the first to operationalize scalable agentic Graph-RAG across independent knowledge domains**, supporting adaptive multi-domain traversal under explicit cost and privacy objectives.

Overall, the main contributions of this work are as follows:

- We introduce the first cross-domain Graph-RAG setting, where knowledge is distributed across independent data holders and cross-domain access is gated by privacy and cost constraints. This setting highlights the scalability and utility challenges of decentralized sense-making queries.
- We propose SCOUT-RAG, a layered agentic system that conducts training-free domain relevance estimation, local vs. global traversal decisions, and adaptive depth–breadth exploration across multiple domains.
- We design a closed-loop evaluation and improvement mechanism in which assessment agents monitor answer sufficiency and factual coverage, guiding strategy agents to selectively expand domains or deepen graph traversal when needed.
- SCOUT-RAG implements multi-level safeguards, including strict time-budget enforcement, best-answer tracking, and coordinated parallelism, to ensure reliable execution and prevent cascading failures in decentralized access settings.
- Evaluations on 89 multi-domain queries across 1–40 countries show that SCOUT-RAG outperforms centralized local and global baselines, while nearly matching centralized DRIFT (56 vs. 63) with over 4× fewer tokens and lower latency, demonstrating efficient and scalable decentralized retrieval.

Related Work

Gragh Retrieval-Augmented Generation. RAG improves language model grounding by retrieving external evidence (Lewis et al. 2020). GraphRAG further enhances contextual reasoning by leveraging graph structure for hierarchical retrieval and local exploration (Edge et al. 2024). Typically, users are required to specify the retrieval mode—global or local—when using Graph-RAG systems, in order to indicate which level of granularity is preferred for the current query. Recent advances have explored organizing retrieved information through hierarchical summary trees (Sarthi et al. 2024), constructing entity graphs from document-level chunks (Zhao et al. 2025), or extracting entities and relations with industrial-grade NLP libraries (Min et al. 2025). However, these methods work under a unified corpus or global knowledge graph, which can be complementary for the problem setup where information is siloed, access is constrained, and retrieval must remain selective and budget-aware. The DRIFT framework [1] from Microsoft Research introduced the idea of combining global reasoning with targeted local inspection to improve retrieval efficiency. Our proposed framework draws inspiration from this design: SCOUT-RAG adopts DRIFT-style global–local balancing but extends it to a multi-domain setting, where retrieval policies must reinterpret the notions of ”global” and ”local” in order to determine the most appropriate level of granularity for each specific query.

Cross-Domain Traversal. Beyond domain-specific reasoning in Graph-RAG (Xiao et al. 2025), recent practical applications have highlighted the need to retrieve information across multiple domains (Liu et al. 2024; Shojaee et al. 2025), which calls for more principled search strategies. Ideally, domains can be accessed independently and concurrently, allowing overall latency to be reduced as multiple domain searches proceed in parallel. However, in practice, latency may be amplified due to bandwidth and network overhead, the dominance of the slowest domain, or additional computation and synchronization delays across domains. A recent automated platform for multi-domain evaluation of RAG systems, OmniBench-RAG (Liang, Zhou, and Wang 2025), has incorporated efficiency as a key dimension in performance quantification. In addition, many systems, such as KG-GPT (Kim et al. 2023) and StructGPT (Jiang et al. 2023), have integrated LLM-based retrievers, implying that exhaustive retrieval across all domains would be costly. Regarding the efficiency problems, existing approaches rely on labeled data to train models that identify relevant domains, commonly referred to as domain routers (Shojaee et al. 2025; Guerraoui et al. 2025). Our proposed SCOUT-RAG leverages an agent to rank domain relevance and dynamically explore appropriate domains based on multiple key signals. The agent can operate under a cold-start condition and becomes increasingly reliable as more queries are accumulated. To enhance usability, we also introduce explicit termination conditions, such as time tolerance and answer-quality constraints.

Agentic AI Systems. Recent advances in multi-agent systems explore coordinated LLM roles for complex reasoning and tool-use workflows. AutoGen (Wu et al. 2024) demonstrates conversational collaboration among agents, while MetaGPT (Hong et al. 2023) formalizes structured

[1] https://microsoft.github.io/graphrag/query/drift_search/

role assignments inspired by software engineering teams. ReAct (Yao et al. 2022) couples reasoning traces with action execution for controllability. Several pioneering works have explored the integration of agentic AI into Graph-RAG systems. KG-R1 (Song et al. 2025) employs a single agent that interacts with the knowledge graph as its environment, learning to retrieve relevant information step by step and integrating it into its reasoning and text generation process. Agent-G (Lee et al. 2025) introduces feedback loops that integrate an agent to improve retrieval quality, where an auxiliary critic module rejects unsatisfactory answers. We build on these insights but shift focus to governing retrieval decisions for multi-sources scenarios: SCOUT-RAG operationalizes four specialized agents for domain assessment, evidence extraction, synthesis, and iterative refinement, each explicitly grounded in cross-domain information access needs.

Methodology

Problem Formulation

We consider retrieval-augmented generation over M distributed knowledge domains. Each domain $i \in \{1, \ldots, M\}$ maintains a private knowledge graph $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i, \mathbf{X}_i)$ consisting of entities, relations, and associated text or structured attributes. Raw data cannot be shared across domains due to privacy and ownership constraints. Given a natural-language query q, the system aims to synthesize an answer A grounded in evidence retrieved from multiple domains, without centralizing the underlying graphs and adhering to a prescribed cost budget.

Let π_i denote the retrieval policy for domain i, specifying whether and how to perform retrieval, with associated cost c_i (e.g., time or computational resources). The objective is to produce an answer that maximizes the answer quality:

$$\begin{aligned} &\max_{A} \quad \mathrm{Qual}(A \mid q,\ \{\pi_i\}_{i=1}^{M}) \\ &\text{s.t.} \quad A = \Phi\big(\{f_{\pi_i}(q, \mathcal{G}_i)\}_{i=1}^{M}\big), \quad \sum_{i=1}^{M} c_i \le \mathcal{C}_{\max}, \end{aligned} \tag{1}$$

where $f_{\pi_i}(\cdot)$ is a generator producing a local answer A_i under policy π_i, $\Phi(\cdot)$ denotes the aggregation function synthesizing the final answer across domains, and $\mathrm{Qual}(\cdot)$ quantifies the grounding fidelity and relevance of the generated answer with respect to retrieved evidence.

In practice, the retrieval policy $\{\pi_i\}$ is not executed in a single pass. Instead, SCOUT-RAG operates through *progressive policy refinement*: (i) initial estimation of domain relevance, (ii) domain-scoped partial retrieval and answer seeding, and (iii) iterative improvement via local depth or breadth expansions guided by answer-quality feedback and remaining budget. This sequential decision process incrementally allocates retrieval resources while honoring $\mathcal{C}_{\max}$, terminating when no further utility is gained or the budget is exhausted.

This formulation captures the fundamental characteristics of distributed Graph-RAG: (i) *partial observability*, since the global graph $\bigcup_{i=1}^{M} \mathcal{G}_i$ is never exposed; (ii) *adaptive cross-domain reasoning*, requiring decisions on when to remain local vs. expand to new domains; and (iii) *cost-aware retrieval refinement*, where utility estimates guide incremental evidence gathering rather than exhaustive querying. Building upon this formulation, we design a decentralized solution that operates the above objectives through coordinated retrieval and synthesis. Specifically, we present the **SCOUT-RAG** framework, which operates in three coordinated stages and leverages four specialized agents. We first identify relevant knowledge domains, then perform adaptive local-to-global graph retrieval, and finally refine answers through iterative quality-guided evidence expansion. An illustrative overview of the framework is provided in Figure 1.

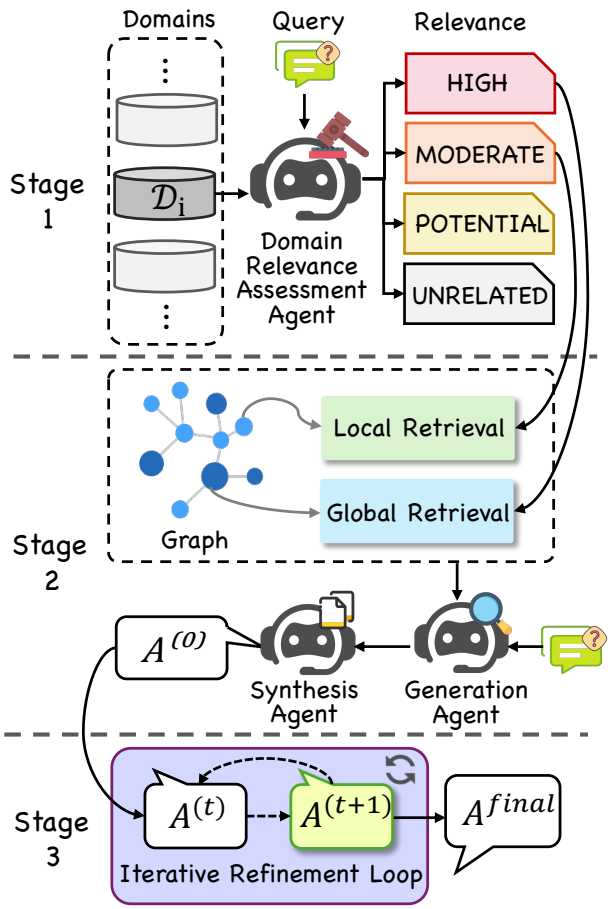

Figure 1: Overview of the proposed **SCOUT-RAG** framework. The framework operates in three stages: (i) domain relevance estimation, (ii) domain-aware retrieval (local vs. global), and (iii) iterative answer refinement via synthesis and generation agents. These agentic components collectively support adaptive, cross-domain reasoning without centralized graph access.

Stage I. Domain Relevance Assessment

The first stage determines which distributed domains are likely to contain knowledge relevant to the query. Accurate relevance estimation is crucial: it suppresses unnecessary retrieval and prioritizes domains that meaningfully contribute

to grounding, thereby improving cost-efficiency and cross-domain reasoning quality.

Domain Relevance Assessment Agent (DRAA). Given distributed knowledge domains $\{\mathcal{D}_i\}_{i=1}^{M}$, we employ a **DRAA** to classify each domain into one of four relevance tiers: `HIGH`, `MODERATE`, `POTENTIAL`, and `IRRELEVANT`.

For each domain $\mathcal{D}_i$ and query q, three complementary information signals are computed:

- **Semantic similarity:** $s_i^{\text{sim}} = \text{Sim}(q, \mathcal{D}_i)$, representing embedding-based cosine similarity between the query and domain-level representations.
- **Knowledge richness:** $s_i^{\text{rich}} = |\mathcal{R}_i| / \max_j |\mathcal{R}_j|$, quantifying relative data abundance by the normalized report count.
- **Historical performance:** $s_i^{\text{hist}} = \frac{1}{|\mathcal{H}_i|} \sum_{h \in \mathcal{H}_i} Q(h)$, capturing average answer quality over past queries associated with the domain.

These features jointly characterize domain–query compatibility from semantic, evidential, and behavioral perspectives. The DRAA combines them with query context to produce a relevance prediction:

$$(\tau_i, r_i) = f_{\text{DRAA}}\left(q, \mathcal{D}_i, s_i^{\text{sim}}, s_i^{\text{rich}}, s_i^{\text{hist}}\right), \tag{2}$$

where τ_i denotes the assigned relevance tier and r_i is a natural-language rationale to support transparency and inspection.

Stage II. Domain-Scoped Seeding

After Stage I identifies candidate domains, Stage II initializes retrieval and constructs the first round answer. This stage balances *coverage* and *precision*: high-confidence domains provide broad contextual grounding, while moderate-confidence domains contribute focused evidence. Two agents collaborate in this phase: the **Partial Answer Generation Agent (PAGA)** retrieves domain-specific evidence, and the **Overall Answer Synthesis Agent (OASA)** synthesizes these signals into an initial cross-domain response.

Partial Answer Generation Agent (PAGA). The PAGA executes domain-adaptive retrieval, guided by the relevance tiers τ_i from Stage I. Retrieval is performed in parallel, but with differentiated granularity:

- `HIGH`-relevance domains: **Global retrieval**, capturing broad semantic context via community-level summaries.
- `MODERATE`-relevance domains: **Local retrieval**, retrieving fine-grained entity-centric or relation-level evidence.
- `POTENTIAL` domains: Deferred, held in reserve for possible activation in Stage III.
- `IRRELEVANT` domains: Excluded to avoid unnecessary cost and noise.

Thus, for each domain $\mathcal{D}_i$ with graph $\mathcal{G}_i$, the partial answer is:

$$\mathcal{A}_i = \begin{cases} f_{\text{PAGA}}^{g}(q, \mathcal{G}_i), & \text{if } \tau_i = \texttt{HIGH}, \\ f_{\text{PAGA}}^{l}(q, \mathcal{G}_i), & \text{if } \tau_i = \texttt{MODERATE}, \\ \emptyset, & \text{otherwise.} \end{cases} \tag{3}$$

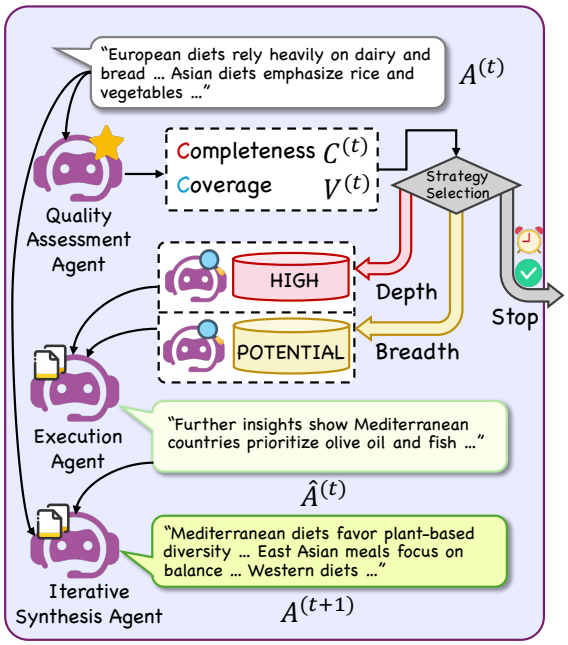

Figure 2: Illustration of Stage III, where the system evaluates and improves answers through selective retrieval and synthesis. *Example query:* "What are the main differences in dietary habits among countries such as Japan, France, and the United States?".

Here, f_{PAGA}^{g} and f_{PAGA}^{l} denote domain-specific global and local graph retrieval operators, respectively. This tiered policy ensures that retrieval cost concentrates only on domains with the highest expected payoff.

Overall Answer Synthesis Agent (OASA). The OASA merges the partial answers into a coherent initial response:

$$A^{(0)} = \Phi_{\text{OASA}}\left(\bigcup_{i=1}^{M} \mathcal{A}_i\right), \tag{4}$$

where $\Phi_{\text{OASA}}(\cdot)$ performs cross-domain answer fusion with explicit source attribution and light consistency checking. The resulting $A^{(0)}$ serves as a grounded *seed* answer: it establishes an initial scaffold, which Stage III subsequently expands and refines as additional evidence is acquired.

Stage III. Iterative Cross-Domain Refinement

Stage III incrementally improves the initial seed answer from Stage II through targeted, quality-guided refinement. The system evaluates the current answer, identifies missing content or uncertainty, and selectively activates retrieval to close knowledge gaps. This iterative loop continues until quality converges or the time budget is exhausted (Figure 2).

Answer Quality Assessment Agent (AQAA). The **AQAA** assesses the current answer $A^{(t)}$ using structured LLM-based evaluation prompts, producing a tuple of metrics and follow-up queries:

$$(C^{(t)}, V^{(t)}, G^{(t)}, \mathcal{F}^{(t)}) = f_{\text{AQAA}}(A^{(t)}), \tag{5}$$

where $C^{(t)}$ measures **completeness** (coverage of core conceptual elements), $V^{(t)}$ measures **breadth** (coverage across relevant subtopics), $G^{(t)}$ denotes unresolved knowledge gaps, and $\mathcal{F}^{(t)} = q_k^f$ is a set of targeted follow-up queries. This structured evaluation provides actionable signals on what to refine next and where.

Strategy Selection. A decision policy selects refinement mode $\alpha^{(t)}$ based on the current quality state and remaining time budget T_r according to our empirical evaluations:

$$\alpha^{(t)} = \begin{cases} \text{Depth}, & \text{if } C^{(t)} < 0.75 \wedge T_\text{r} > 15\text{s} \\ \text{Breadth}, & \text{if } V^{(t)} < 0.70 \wedge T_\text{r} > 10\text{s} \\ \text{Hybrid}, & \text{if } C^{(t)} < 0.75 \wedge V^{(t)} < 0.70 \wedge T_\text{r} > 20\text{s} \\ \text{Stop}, & \text{otherwise} \end{cases} \tag{6}$$

Refinement terminates when:

- Quality threshold met: $Q^{(t)} \geq 0.85$
- Time budget depleted: $T_\text{r} < 5\text{s}$
- Improvement stagnation: $\Delta Q^{(t)} = Q^{(t)} - Q^{(t-1)} < \epsilon$

where $Q^{(t)} = \frac{1}{2}(C^{(t)} + V^{(t)})$.

Once a strategy is selected, refinement proceeds along one of three modes: a depth-focused pass that queries only `HIGH`-relevance domains using the targeted follow-up questions $\mathcal{F}^{(t)}$ to close core completeness gaps (similar in spirit to focused re-retrieval approaches such as DRIFT (Edge et al. 2024)); a breadth-expansion pass that activates previously dormant `POTENTIAL` domains to capture peripheral or emerging subtopics; or a hybrid mode that runs both paths in parallel when the answer is simultaneously incomplete and narrow. If none of the strategies are triggered, e.g., quality saturation or insufficient time, the system halts refinement.

Partial Answer Generation Agent (PAGA). The **PAGA** executes the selected refinement mode by conducting targeted local retrieval:

$$\hat{\mathcal{A}}^{(t)} = \begin{cases} \bigcup_{i \in \mathcal{D}_\text{high}} f^l_\text{PAGA}(\mathcal{F}^{(t)}, \mathcal{G}_i), & \alpha^{(t)} = \text{Depth}, \\ \bigcup_{j \in \mathcal{D}_\text{pot}} f^l_\text{PAGA}(q, \mathcal{G}_j), & \alpha^{(t)} = \text{Breadth}, \\ \text{both above}, & \alpha^{(t)} = \text{Hybrid}. \end{cases} \tag{7}$$

Depth queries refine critical details in high-relevance domains, while breadth queries activate latent domains to capture overlooked context.

Overall Answer Synthesis Agent (OASA). The **OASA** integrates newly retrieved evidence with the current answer:

$$A^{(t+1)} = \Phi_\text{OASA}\left(A^{(t)}, \hat{\mathcal{A}}^{(t)}\right). \tag{8}$$

To ensure stability, the system tracks all intermediate candidates $\{A^{(0)}, A^{(1)}, \ldots, A^{(t)}\}$ and returns the highest-quality answer $A^* = \arg\max_i Qual(A^{(i)})$ upon termination, ensuring resilience against late-stage drift or noisy retrieval. For clarity, we present a summary of SCOUT-RAG in Algorithm 1.

Algorithm 1: Scalable Cost-Efficient Cross-Domain Traversal for Distributed Agentic Graph-RAG

Require: Query q, domains $\{\mathcal{D}_i\}_{i=1}^M$, time budget T_{max}
Ensure: Final answer A^*

Stage 1: Domain Relevance Assessment
for $D_i \in \mathcal{D}$ **do**
  Compute domain relevance via DRAA: `HIGH`/`MODERATE`/`POTENTIAL`/`IRRELEVANT`
end for
Stage 2: Domain-Scoped Seeding
Perform global retrieval on `HIGH` domains
Perform local retrieval on `MODERATE` domains
PAGA produces partial answers
OASA synthesizes seed answer $A^{(0)}$
Stage 3: Iterative Cross-Domain Refinement
while $T_r > 0$ **do**
  Evaluate current answer $A^{(t)}$ via AQAA to obtain
  $(C^{(t)}, V^{(t)}, G^{(t)}, \mathcal{F}^{(t)})$
  Compute quality $Q^{(t)} = \frac{1}{2}(C^{(t)} + V^{(t)})$
  Select refinement strategy via Eq. (6):
  if $\alpha^{(t)} = \text{STOP}$ **then**
    break
  end if
  Conduct retrieval via PAGA according to $\alpha^{(t)}$
  Update answer via OASA fusion to obtain $A^{(t+1)}$
  Re-evaluate via AQAA
  if quality improves **then** update best answer A^*
  end if
end while
return A^*

Experiments

Experimental Setup

Dataset To assess the scalability of our framework in handling a large number of domains, we employ Wikipedia articles from 45 countries as simulated domain datasets. Within each domain, a graph-based knowledge base is constructed, comprising 9–77 community reports. We also generate 100 queries with varying levels of domain coverage to comprehensively evaluate our framework, of which 89 are effectively answered by all methods. The queries are categorized into five levels: single-domain (20 queries, each involving one country), small multi-domain (20 queries, each involving five countries), medium multi-domain (20 queries, each involving ten countries), large multi-domain (20 queries, each involving twenty countries), and very large multi-domain (9 queries, each involving forty countries).

Baseline We compare against four GraphRAG baselines across two deployment paradigms. Under the centralized setting (all data from 45 countries consolidated into a single unified domain), we evaluate:

- **GraphRAG$_\text{Local}$**: Local search for detailed entity-level information retrieval from specific community reports.
- **GraphRAG$_\text{Global}$**: Global search across hierarchical structures for comprehensive overviews via community

Table 1: Average Performance Across Evaluation Dimensions. The best scores for centralized and decentralized methods are highlighted in bold.

Category	Method	Quality					Cost	
		Comp. (↑)	Div. (↑)	Emp. (↑)	Dir. (↑)	Overall (↑)	Time(s) (↓)	Token (↓)
Centralized	$\text{GraphRAG}_{\text{Local}}$	65	55	35	58	53	**34.40**	**11,223**
	$\text{GraphRAG}_{\text{Global}}$	60	50	30	55	49	45.89	640,574
	$\text{GraphRAG}_{\text{DRIFT-c}}$	**72**	**70**	**45**	**63**	**63**	231.85	693,731
Decentralized	$\text{GraphRAG}_{\text{DRIFT-dec}}$	**90**	**88**	**75**	**88**	**85**	414.88	879,911
	SCOUT-RAG (Ours)	65	60	40	58	56	**75.32**	**159,169**

summaries.

- **$\text{GraphRAG}_{\text{DRIFT-c}}$**: One global search followed by two refinement rounds (three follow-up local searches per round) to iteratively improve answer quality.

Under the decentralized setting (data distributed across 45 country-specific domains), we include:

- **$\text{GraphRAG}_{\text{DRIFT-dec}}$**: Applies DRIFT independently in each domain—one global search plus two refinement rounds per domain, with final synthesis across all 45 domain-level answers.

Mectric Following interactive agent evaluation methodologies (Zhou et al. 2023), we assess four dimensions adapted from GraphRAG (Edge et al. 2024): Comprehensiveness (coverage of relevant aspects), Diversity (breadth of perspectives and sources), Empowerment (utility for follow-up exploration and actionable insights), and Directness (conciseness and clarity). Each dimension is scored 0–100 by GPT-4o as an LLM-as-judge using structured prompts.

All experiments were conducted on a system equipped with an NVIDIA RTX 3070 GPU with 9 GB VRAM (driver version 575.57.08). The compute environment consisted of 16 AMD EPYC 7B12 vCPUs, 31 GB memory.

Main Results

Answer Quality Comparison Table 1 summarizes the average performance across 89 queries spanning five complexity levels, evaluated by GPT-4o under two contrasting paradigms that comprise three centralized methods and two decentralized methods.

In decentralized mode, $\text{GraphRAG}_{\text{DRIFT-dec}}$ achieves the highest overall score (85) among all methods through exhaustive per-domain retrieval. Among centralized approaches, $\text{GraphRAG}_{\text{DRIFT-c}}$ reaches 63, outperforming $\text{GraphRAG}_{\text{Local}}$ (53) and $\text{GraphRAG}_{\text{Global}}$ (49) by more than 10 points. The proposed SCOUT-RAG system explores a third approach that combines decentralized deployment with intelligent and quality-controlled orchestration. It achieves an overall score of 56, which is 7 points lower than $\text{GraphRAG}_{\text{DRIFT-c}}$ but 7 points higher than $\text{GraphRAG}_{\text{Global}}$ and 3 points higher than $\text{GraphRAG}_{\text{Local}}$. This near-parity is notable given that SCOUT-RAG operates under distributed constraints without centralized coordination. The small performance gap indicates that quality-controlled orchestration effectively preserves answer quality within decentralized architectures.

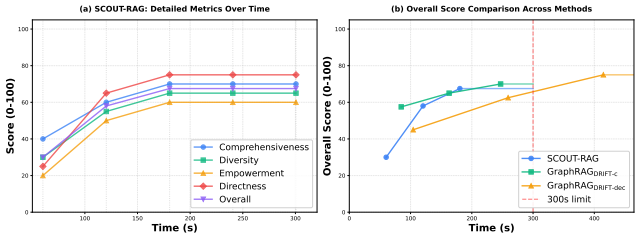

Figure 3: Time-performance monitoring and comparison.

Across individual evaluation dimensions, SCOUT-RAG demonstrates particularly strong performance in diversity (60), surpassing centralized baselines by 5–10 points. This improvement results from hierarchical domain categorization (`HIGH`/`MODERATE`/`POTENTIAL`) and quality-driven activation, which encourage systematic exploration of multiple perspectives. Although $\text{GraphRAG}_{\text{DRIFT-dec}}$ achieves the highest overall performance, it incurs an extreme computational cost of 414.88 seconds and 879,911 tokens, which is more than 5.5 times the resource consumption of SCOUT-RAG, as analyzed in the following section.

Cost-Efficiency Analysis A critical advantage of SCOUT-RAG lies in resource efficiency. Our system consumes 159,169 tokens and completes queries in 75.32 seconds on average, representing 81.9% token reduction and 81.8% speedup compared to $\text{GraphRAG}_{\text{DRIFT-dec}}$ (879,911 tokens, 414.88s). Even against centralized $\text{GraphRAG}_{\text{Global}}$ (640,574 tokens), we achieve 75.2% fewer tokens while maintaining comparable quality (56 vs. 49).

The cost and performance trade-off is particularly notable. $\text{GraphRAG}_{\text{DRIFT-dec}}$ attains the highest overall score (85) through exhaustive per-domain retrieval but requires 414.88s and 879,911 tokens, which is approximately 5.5 times SCOUT-RAG's execution time and token usage. Centralized $\text{GraphRAG}_{\text{DRIFT-c}}$ (63) also incurs significant overhead, consuming 231.85s and 693,731 tokens, which are 3.1 times and 4.4 times higher, respectively. Although $\text{GraphRAG}_{\text{Local}}$ executes faster (34.40s), it does so at the expense of diversity (55 vs. 60) and overall quality.

These efficiency gains enable practical deployment in resource-constrained scenarios. By achieving near-baseline performance with >80% cost reduction compared to decentralized alternatives, SCOUT-RAG demonstrates the viability of quality-controlled orchestration for distributed knowledge retrieval at scale.

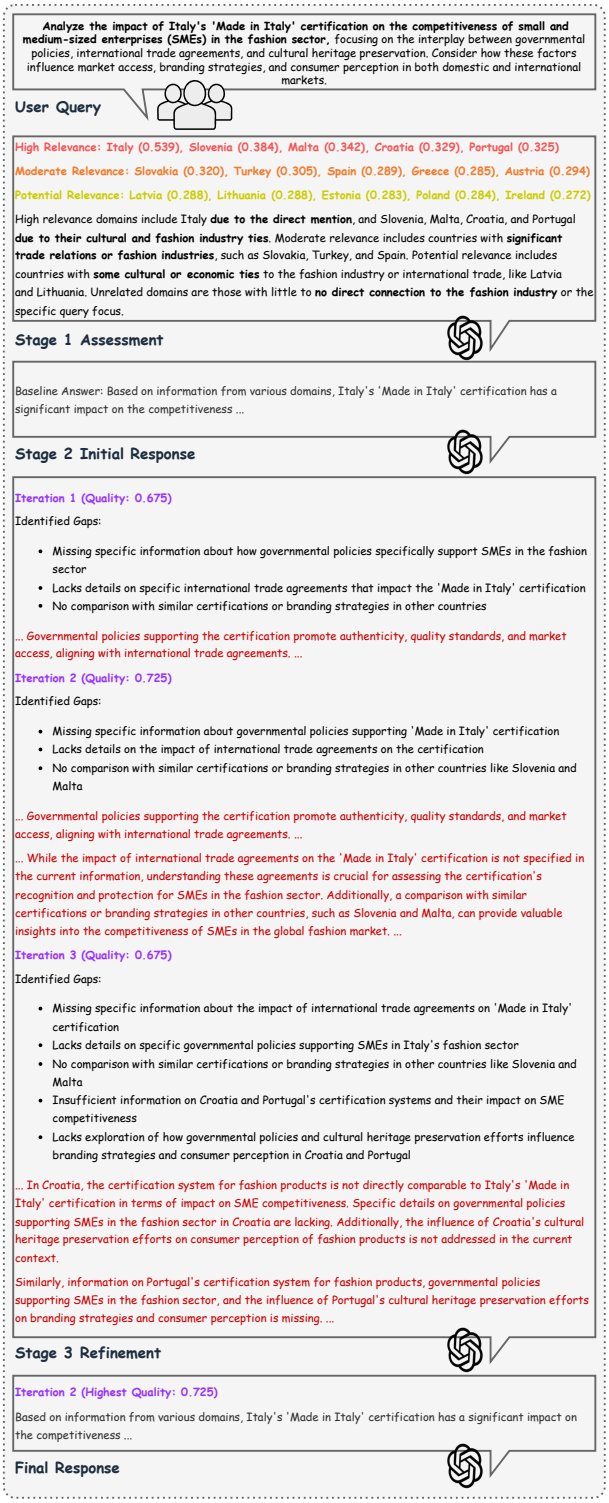

Figure 4: Case study on Italy's "*Made in Italy*" certification.

A Closer Look at the Time Budget

Figure 3 (left) demonstrates how SCOUT-RAG's answer quality evolves over time. All evaluation dimensions—Comprehensiveness, Diversity, Empowerment, and Directness—show rapid improvement in the first 120 seconds, followed by a performance plateau around 180s, indicating diminishing returns beyond this point. Figure 3 (right) compares three methods under different time constraints. SCOUT-RAG converges quickly to stable quality within the 300s budget. GraphRAG$_{\text{DRIFT-c}}$ achieves similar performance with comparable time efficiency. GraphRAG$_{\text{DRIFT-dec}}$, though requiring more time, ultimately reaches the highest overall score, demonstrating a trade-off between response speed and answer quality.

Case Study

We conducted a case study of SCOUT-RAG using a complex query. The overall workflow is presented in Figure 4.

In Stage I, the domain relevance module identified 10 relevant domains within 8.28 seconds. Italy ranked first (0.539), followed by Slovenia, Malta, Croatia, and Portugal, reflecting strong cultural and institutional similarity. This outcome demonstrates the system's ability to perform accurate geocultural alignment, ensuring that subsequent reasoning is grounded in comparable certification frameworks.

Stage II produced an initial domain-scoped synthesis in 53.26 seconds. The system captured the main analytical dimensions concerning competitiveness, policy instruments, trade agreements, and heritage branding, but lacked concrete examples and cross-country references. AQAA rated breath at 0.65 and completeness at 0.70, prompting refinement to enhance specificity and comparative depth.

Stage III executed three iterative refinement rounds over 41.12 seconds. The first round expanded retrieval to Slovenia, Malta, Croatia, and Portugal, introducing policy and branding details and improving coverage to 0.75. The second iteration focused on Italy, Malta, and Slovenia, achieving the highest overall quality (0.725) through contextual and evidential balance. The third round employed Breadth to explore `POTENTIAL` domains, but added minor noise and reduced final quality to 0.675. The best-track mechanism retained Iteration 2, preventing late-stage degradation.

Overall, this case demonstrates that SCOUT-RAG effectively integrates policy, trade, and cultural heritage information across domains. Through adaptive retrieval, iterative synthesis, and best-track optimization, the system maintains analytical precision and interpretive coherence in addressing complex, interdisciplinary socio-economic questions such as "Made in Italy."

Conclusion

This work presented SCOUT-RAG, a scalable and cost-efficient agentic framework for retrieval-augmented generation across distributed knowledge domains. Unlike centralized Graph-RAG systems, SCOUT-RAG enables progressive cross-domain retrieval under explicit cost and privacy constraints. Empirical evaluations demonstrate that SCOUT-RAG achieves performance comparable to the centralized DRIFT, while maintaining a modest gap compared to the near-exhaustive Graph-RAG with DRIFT. Cost-efficiency analysis shows that SCOUT-RAG achieves significant reductions in both token consumption and execution time, demonstrating its potential for real-world deployments.

Acknowledgments

This research is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-003). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. This research is supported by A*STAR Career Development Fund <Project No. C243512010>.

References

Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; Metropolitansky, D.; Ness, R. O.; and Larson, J. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Guerraoui, R.; Kermarrec, A.-M.; Petrescu, D.; Pires, R.; Randl, M.; and de Vos, M. 2025. Efficient federated search for retrieval-augmented generation. In *Proceedings of the 5th Workshop on Machine Learning and Systems*, 74–81.

Hong, S.; Zhuge, M.; Chen, J.; Zheng, X.; Cheng, Y.; Wang, J.; Zhang, C.; Wang, Z.; Yau, S. K. S.; Lin, Z.; et al. 2023. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Jiang, J.; Zhou, K.; Dong, Z.; Ye, K.; Zhao, W. X.; and Wen, J.-R. 2023. Structgpt: A general framework for large language model to reason over structured data. *arXiv preprint arXiv:2305.09645*.

Kim, J.; Kwon, Y.; Jo, Y.; and Choi, E. 2023. Kg-gpt: A general framework for reasoning on knowledge graphs using large language models. *arXiv preprint arXiv:2310.11220*.

Lee, M.-C.; Zhu, Q.; Mavromatis, C.; Han, Z.; Adeshina, S.; Ioannidis, V. N.; Rangwala, H.; and Faloutsos, C. 2025. Agent-G: An Agentic Framework for Graph Retrieval Augmented Generation.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474.

Liang, J.; Zhou, S.; and Wang, K. 2025. OmniBench-RAG: A Multi-Domain Evaluation Platform for Retrieval-Augmented Generation Tools. *arXiv preprint arXiv:2508.05650*.

Liu, Z.; Yang, K.; Xie, Q.; de Kock, C.; Ananiadou, S.; and Hovy, E. 2024. RAEmoLLM: Retrieval augmented LLMs for cross-domain misinformation detection using in-context learning based on emotional information. *arXiv preprint arXiv:2406.11093*.

Luo, H.; E, H.; Chen, G.; Lin, Q.; Guo, Y.; Xu, F.; Kuang, Z.; Song, M.; Wu, X.; Zhu, Y.; and Tuan, L. A. 2025. Graph-R1: Towards Agentic GraphRAG Framework via End-to-end Reinforcement Learning. *arXiv preprint arXiv:2507.21892*.

Min, C.; Mathew, R.; Pan, J.; Bansal, S.; Keshavarzi, A.; and Kannan, A. V. 2025. Efficient Knowledge Graph Construction and Retrieval from Unstructured Text for Large-Scale RAG Systems. *arXiv preprint arXiv:2507.03226*.

Peng, B.; Zhu, Y.; Liu, Y.; Bo, X.; Shi, H.; Hong, C.; Zhang, Y.; and Tang, S. 2024. Graph retrieval-augmented generation: A survey. *arXiv preprint arXiv:2408.08921*.

Sarthi, P.; Abdullah, S.; Tuli, A.; Khanna, S.; Goldie, A.; and Manning, C. D. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*.

Shojaee, P.; Harsha, S. S.; Luo, D.; Maharaj, A.; Yu, T.; and Li, Y. 2025. Federated retrieval augmented generation for multi-product question answering. *arXiv preprint arXiv:2501.14998*.

Singh, A.; Ehtesham, A.; Kumar, S.; and Khoei, T. T. 2025. Agentic retrieval-augmented generation: A survey on agentic rag. *arXiv preprint arXiv:2501.09136*.

Song, J.; Wang, S.; Shun, J.; and Zhu, Y. 2025. Efficient and Transferable Agentic Knowledge Graph RAG via Reinforcement Learning. *arXiv preprint arXiv:2509.26383*.

Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; Liu, J.; et al. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*.

Xiao, Y.; Dong, J.; Zhou, C.; Dong, S.; Zhang, Q.-w.; Yin, D.; Sun, X.; and Huang, X. 2025. GraphRAG-Bench: Challenging Domain-Specific Reasoning for Evaluating Graph Retrieval-Augmented Generation. *arXiv preprint arXiv:2506.02404*.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K. R.; and Cao, Y. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Zhao, Y.; Zhu, J.; Guo, Y.; He, K.; and Li, X. 2025. Eˆ 2GraphRAG: Streamlining Graph-based RAG for High Efficiency and Effectiveness. *arXiv preprint arXiv:2505.24226*.

Zhou, X.; Zhu, H.; Mathur, L.; Zhang, R.; Yu, H.; Qi, Z.; Morency, L.-P.; Bisk, Y.; Fried, D.; Neubig, G.; et al. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.