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

Longkun Li, Yuanben Zou, Jinghan Wu, Yuqing Wen, Jing Li, Hangwei Qian, Ivor Tsang

Research Problem

- **Background.** Existing Graph-RAG systems assume centralized knowledge graphs, which is unrealistic in practice.
- **Reality.** Knowledge is distributed, private, and cost-gated (e.g., hospitals, countries, organizations). Exhaustive cross-domain retrieval leads to high latency and token cost.
- **Core Challenge.** Cross-domain reasoning under partial observability and strict budgets.

Method Overview

Core Idea. We model cross-domain Graph-RAG as a sequential, agentic decision process instead of one-shot retrieval.

SCOUT-RAG progressively decides:
- *which domains to query*,
- *how deep or broad to traverse*,
- *when to stop*, under explicit time and cost budgets.

SCOUT-RAG Framework

SCOUT-RAG consists of three stages coordinated by four agents (Figure. 1).

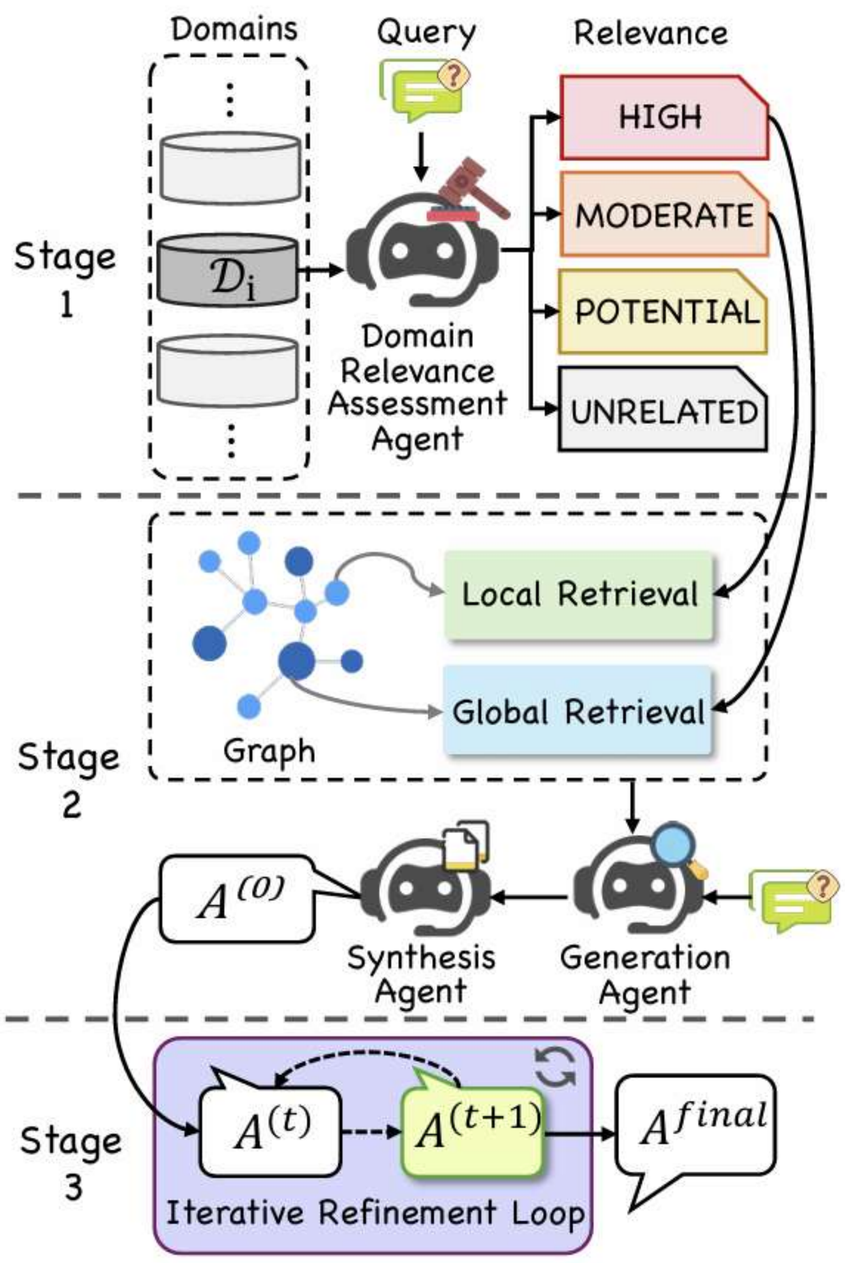

Figure 1: Overview of the proposed SCOUT-RAG framework

- **Stage I: Domain Relevance Assessment.** Classifies domains into HIGH / MODERATE / POTENTIAL / IRRELEVANT.
- **Stage II: Domain-Scoped Seeding.** HIGH → global retrieval, MODERATE → local retrieval; partial answers are synthesized into an initial response.
- **Stage III: Iterative Refinement.** Answer quality is evaluated and refined via selective depth/breadth expansion (Figure. 2).

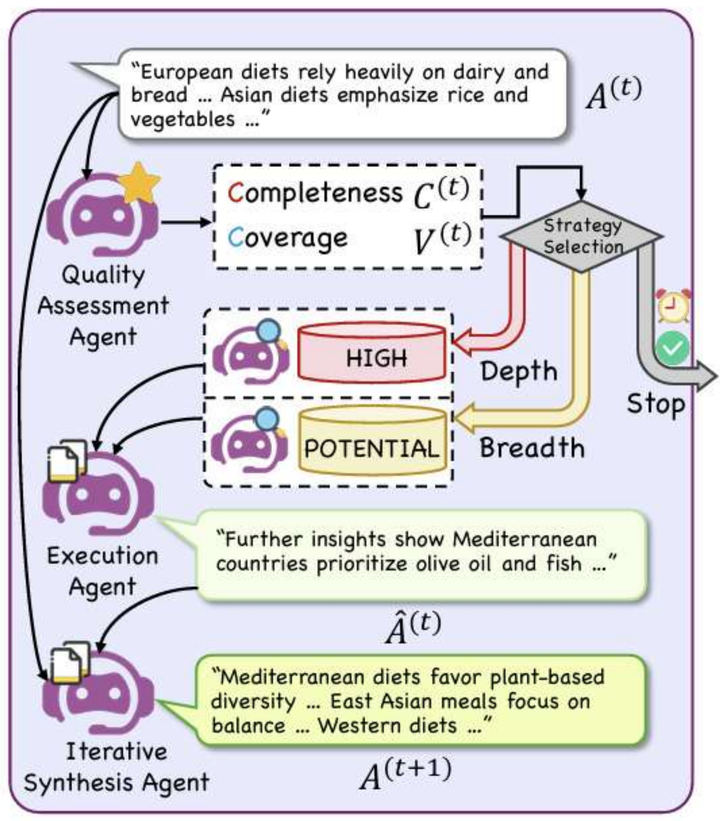

Figure 2: Illustration of Stage III

Retrieval is governed by explicit, interpretable agent decisions rather than fixed pipelines.

- **Domain Control.** Relevance is estimated from semantic similarity, knowledge richness, and historical utility, enabling cold-start deployment.
- **Depth–Breadth Control.**

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

- **Best-Answer Track.** A best-answer tracking mechanism prevents late-stage degradation.

Experiments

Setup. We evaluate on 89 question-answering tasks across 45 distributed domains (*each question involves a different subset of domains*), comparing against centralized and decentralized GraphRAG baselines with DRIFT.

Results. SCOUT-RAG achieves near centralized DRIFT quality with:
- >80% token reduction,
- >80% latency reduction, compared to exhaustive decentralized traversal,
- Near-centralized quality at a fraction of the cost.

A Closer Look at the Time Budget.

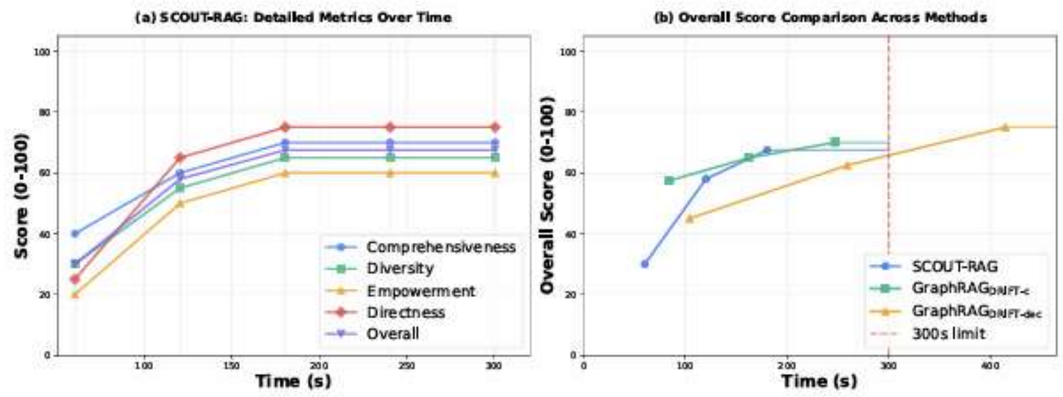

Figure 3: Time-performance monitoring and comparison.

Agency for Science, Technology and Research SINGAPORE
NUS National University of Singapore
XI'AN JIAOTONG UNIVERSITY 1896