ECHORAG: EVOLVING COMMUNITIES AND HARMONIZING INTENT FOR RETRIEVAL-AUGMENTED GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph-based retrieval-augmented generation (RAG) organizes external corpora into reusable knowledge communities, but typically overlooks the complex query process that determines knowledge utility. This creates two coupled gaps: an *intent-blind knowledge utilization gap*, where underspecified queries yield evidence misaligned with latent needs, and a *query-blind knowledge organization gap*, where structurally coherent communities ignore downstream query distributions. To bridge them, we propose **E**volving **C**ommunities and **H**arm**o**nizing Intent for **R**etrieval-**A**ugmented **G**eneration (EchoRAG), a self-evolving framework that approximates query-specific knowledge organization without per-query reconstruction. EchoRAG begins with a structural initialization to establish a baseline graph partition and then introduces query-guided collaborative community evolution: accumulated query feedback determines where and when to update the knowledge base, while structural-semantic signals guide community evolution. EchoRAG further incorporates an intent-harmonized knowledge utilization mechanism: it infers latent intent to retrieve and transform retrieved knowledge into grounded, intent-aligned evidence anchored to source spans. Utilization traces provide delayed feedback for future evolution, closing the loop between organization and utilization. Extensive experiments on three benchmark datasets demonstrate that EchoRAG significantly outperforms existing state-of-the-art baseline approaches. Our code is available at `https://anonymous.4open.science/r/EchoRAG`.

1 INTRODUCTION

Retrieval-Augmented Generation (RAG) has emerged as a pivotal paradigm for fortifying Large Language Models (LLMs) with reliable, up-to-date external knowledge. A standard RAG framework operates by retrieving query-relevant text chunks and seamlessly weaving them into the generative process as contextual evidence (Lewis et al., 2020). Despite its empirical success, traditional RAG inherently suffers from the isolated knowledge retrieval bottleneck. Specifically, retrieved text chunks are processed as structurally independent units, leaving their underlying relationships entirely unmodeled. To circumvent this limitation, recent literature (Edge et al., 2024; Guo et al., 2024; Wang et al., 2025b) introduces graph-based RAG frameworks. These approaches systematically structure textual corpora into interconnected graph representations, thereby unlocking capacities for relational reasoning and topology-aware exploration.

Formally, graph-based RAG includes $\mathcal{O} : \mathcal{G} \rightarrow \mathcal{K}$, which denotes the knowledge organization mapping that transforms a knowledge graph $\mathcal{G}$ into a set of reusable knowledge units $\mathcal{K}$, and $\mathcal{U} : \mathcal{K} \times \mathcal{Q} \rightarrow \mathcal{Y}$ denotes the knowledge utilization function that processes these units alongside a user query $q \in \mathcal{Q}$ to yield a final response $y \in \mathcal{Y}$. GraphRAG (Edge et al., 2024) establishes a pioneering paradigm of *offline indexing followed by online retrieval*. During the offline stage, it leverages LLMs to extract entities and relations from raw corpora to construct $\mathcal{G}$, applies graph partitioning algorithms (e.g., Leiden (Traag et al., 2019)) to detect densely connected subgraphs, i.e., communities, and generates community-level summaries to instantiate $\mathcal{K}$. In the online stage, it retrieves query-aligned summaries to synthesize the response. Following this paradigm, subsequent studies expand the structural expressiveness of $\mathcal{G}$ to optimize the mapping $\mathcal{O}$. For instance,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ArchRAG (Wang et al., 2025b) formulates $\mathcal{G}$ as an attributed graph to embed fine-grained metadata into $\mathcal{K}$; Zeng et al. (Zeng et al., 2026) model $\mathcal{G}$ as a dynamic graph to ensure temporal consistency in $\mathcal{K}$ over evolving corpora; and CogRAG (Hu et al., 2026) upgrades $\mathcal{G}$ to a hypergraph topology, enabling $\mathcal{O}$ to encapsulate higher-order relationships among multiple entities. A comprehensive related work is in Appendix B.

However, existing literature predominantly concentrates on optimizing the knowledge base construction and retrieval mechanisms, leaving the inherent complexity of the query process largely unexplored. This fundamental oversight precipitates two gaps:

Intent-Blind Knowledge Utilization Gap. Current frameworks naively deploy the raw query q as the explicit search directive. Yet, the intrinsic quality of q is rarely guaranteed. Constrained by limited domain expertise or suboptimal articulation, users frequently fail to formulate their true underlying intent, denoted as $I(q)$. For example, a patient querying "sudden vision loss and jaw pain" is essentially seeking treatment information regarding "giant cell arteritis", but without explicitly articulating this precise medical entity, the semantic matching within the graph inevitably fails. Consequently, the knowledge retrieved directly from the raw query fundamentally diverges from the knowledge required to satisfy the true intent, mathematically expressed as $R(q) \neq R(I(q))$, where $R(\cdot)$ represents the retrieval function.

Query-Blind Knowledge Organization Gap. Existing graph-based RAG frameworks formulate the optimization of $\mathcal{O}$ as entirely agnostic to q. However, the practical utility of an organized knowledge structure $\mathcal{K}$ is not solely dictated by the topological density of entities within $\mathcal{G}$, but profoundly hinges on the distribution of downstream queries and how they intend to traverse these structures. Intuitively, this is analogous to archiving corporate documents strictly by departmental boundaries based on structural affiliation. While topologically coherent, this organization mapping $\mathcal{O}$ becomes inherently dysfunctional when faced with a cross-functional query q regarding a "product launch workflow", which necessitates traversing scattered nodes across multiple departments. By divorcing $\mathcal{O}$ from $\mathcal{Q}$, current methods inevitably yield sub-optimal knowledge units that are structurally isolated from actual retrieval demands.

To bridge these fundamental gaps, we propose **EchoRAG** (**E**volving **C**ommunities and **H**arm**o**nizing Intent for **R**etrieval-**A**ugmented **G**eneration). EchoRAG pioneers a unified paradigm that tightly couples query-guided knowledge organization with intent-harmonized knowledge utilization. Specifically, to resolve the query-blind knowledge organization gap, EchoRAG introduces a *query-guided collaborative community evolution mechanism*. The process begins with a structural initialization to establish an optimal baseline graph partition. During online serving, query-intent distributional shifts dictate the precise spatial and temporal triggers for knowledge maintenance. A joint structural and semantic objective governs the evolutionary trajectory of the communities, with LLMs invoked only when the two signals conflict. Consequently, downstream query feedback steers selective knowledge maintenance efficiently, bypassing the prohibitive computational overhead of reconstructing the entire graph topology. To bridge the intent-blind knowledge utilization gap, EchoRAG incorporates an *intent-harmonized knowledge utilization mechanism*. It executes intent parsing according to a raw user query q to formulate an ideal answer hypothesis, serving as a high-fidelity proxy for the latent information need $I(q)$. This decoded intent is subsequently fused with the explicit query to retrieve complementary community structures. EchoRAG further sculpts the retrieved raw content into tightly grounded, intent-aligned evidence while strictly preserving the original textual spans as immutable factual anchors. Through this lightweight adaptation process, EchoRAG successfully maps the sub-optimal query-relevant knowledge $R(q)$ onto the optimal intent-aligned knowledge distribution $R(I(q))$.

In summary, we formalize the systemic neglect of query dynamics, which inherently manifests as the intent-blind knowledge utilization gap and the query-blind knowledge organization gap. Based on a solid mathematical formulation and theoretical justification for these limitations, we propose **EchoRAG**, a pioneering self-evolving graph RAG paradigm. By seamlessly integrating query-guided collaborative community evolution with intent-harmonized knowledge utilization, EchoRAG dynamically distills workload feedback into reusable knowledge units and transforms retrieved communities into grounded evidence strictly tailored to the latent information need. Extensive empirical evaluations across diverse benchmark datasets demonstrate that EchoRAG consistently surpasses state-of-the-art baselines in both retrieval precision and generation fidelity, comprehensively validating the superiority and generalizability of the proposed paradigm.

2 PRELIMINARIES

This section provides the key design principles of our framework: adapting knowledge organization to evolving query demand and incorporating inferred intent into evidence retrieval. These results directly motivate the selective community-update strategy and intent-conditioned retrieval mechanism introduced in the subsequent section. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathbf{X}, \mathbf{E})$ be a text-attributed graph. Graph RAG can be viewed as marginalizing over a discrete space $\mathbf{K}$ of knowledge organizations:

$$p(y \mid q, \mathcal{G}) = \sum_{\mathcal{K} \in \mathbf{K}} p_\theta(y \mid q, \mathcal{K}) p_\phi(\mathcal{K} \mid \mathcal{G}), \tag{1}$$

where p_ϕ organizes the graph offline and p_θ utilizes the resulting knowledge online. This separates how evidence is grouped from how it is selected for a query. We next show that both query-blind organization and intent-blind utilization can be suboptimal.

2.1 QUERY-BLIND KNOWLEDGE ORGANIZATION

A query q induces routing weights $a(q) \in \Delta^{C-1}$ over C communities. For a query distribution μ, define the population demand and organization risk as

$$\pi_\mu(c) := \mathbb{E}_{Q \sim \mu}[a_c(Q)], \tag{2}$$

$$\mathcal{R}_\pi(\mathcal{K}) := \sum_{c=1}^{C} \pi_\mu(c) L_c(\mathcal{K}), \qquad L_c(\mathcal{K}) \in [0, L], \tag{3}$$

where $\mathcal{K} \in \mathfrak{K}_B(\mathcal{G})$ obeys storage budget B and L_c measures the deficiency associated with community c. Thus, $\mathcal{R}_\pi$ evaluates an organization under its query workload rather than graph structure alone.

Theorem 1 (Demand-induced regret and selective-update certificate)**.** *For any query-blind organization $\mathcal{K}_0$, define*

$$\Gamma_{\mathrm{org}}(\pi; \mathcal{K}_0) := \mathcal{R}_\pi(\mathcal{K}_0) - \min_{\mathcal{K} \in \mathfrak{K}_B(\mathcal{G})} \mathcal{R}_\pi(\mathcal{K}). \tag{4}$$

Then $\Gamma_{\mathrm{org}} \geq 0$, with equality for every π iff $\mathcal{K}_0$ minimizes every L_c. Moreover, for $D_\pi(\mathcal{K}, \mathcal{K}') := \mathcal{R}_\pi(\mathcal{K}) - \mathcal{R}_\pi(\mathcal{K}')$,

$$|D_{\pi_1} - D_{\pi_0}| \leq L \|\pi_1 - \pi_0\|_1. \tag{5}$$

Finally, let $d_\epsilon(c) := |\pi_1(c) - \pi_0(c)| / [\pi_1(c) + \pi_0(c) + \epsilon]$ and $\mathcal{A}_\tau := \{c : d_\epsilon(c) > \tau\}$. The drift ignored by updating only $\mathcal{A}_\tau$ is bounded by

$$\left| \sum_{c \notin \mathcal{A}_\tau} [\pi_1(c) - \pi_0(c)][L_c(\mathcal{K}) - L_c(\mathcal{K}')] \right| \leq L\tau(2 + C\epsilon). \tag{6}$$

Proof sketch. Nonnegativity follows from feasibility. Choosing $\pi = e_c$ isolates L_c and proves the equality characterization. Expanding $D_{\pi_1} - D_{\pi_0}$ and using $|L_c(\mathcal{K}) - L_c(\mathcal{K}')| \leq L$ gives Eq. equation 5. On unselected communities, the threshold definition bounds each demand change; summing these terms yields Eq. equation 6. See Appendix A.1. □

The result motivates EchoRAG's demand logging and selective evolution: changes in π determine where maintenance is needed, while τ limits the update scope with a bounded residual. Appendix A.3 further establishes when this trigger is reliable under finite query samples.

Let $\mathcal{F}_\mu(\mathcal{K})$ denote the probability that the evidence required by a query is not contained in any single community of partition $\mathcal{K}$.

Proposition 1 (Structural metrics can be maximally misaligned)**.** *There exist a graph and workload for which the unique modularity-maximizing partition has $\mathcal{F}_\mu = 1$, whereas a workload-aware partition has $\mathcal{F}_\mu = 0$.*

Proof sketch. For the disjoint edges (a, b) and (c, d), modularity selects $\{\{a, b\}, \{c, d\}\}$, which fragments every request in a workload requiring $\{a, c\}$ or $\{b, d\}$; the crossed partition has zero risk. See Appendix A.2. □

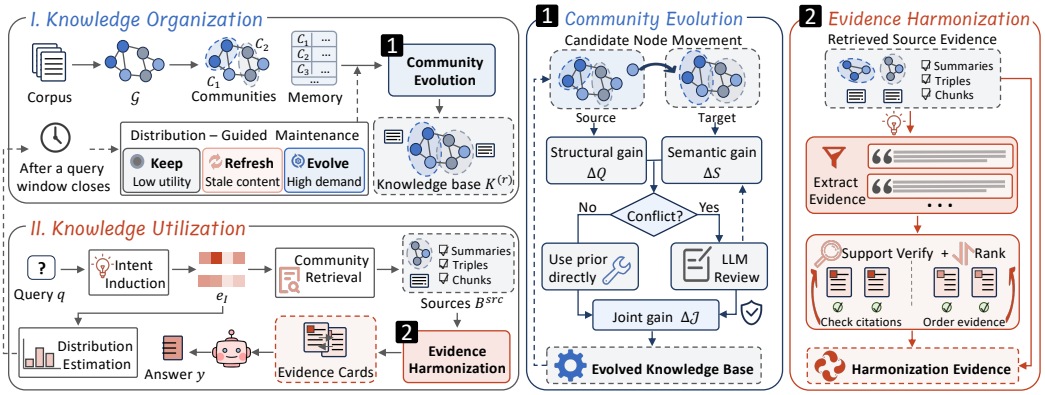


Figure 1: The framework of EchoRAG.

Thus, topology provides a useful warm start but not a sufficient retrieval objective, motivating the topology–semantic refinement engine in Section 3.1.

2.2 Intent-Blind Knowledge Utilization

Let e_q, z_q, and $\widehat{z}_q$ be the explicit-query, true-intent, and estimated-intent representations. Define

$$v_q^\star := (1-\alpha)e_q + \alpha z_q, \quad \widehat{v}_q := (1-\alpha)e_q + \alpha\widehat{z}_q, \quad \eta_q := \|\widehat{z}_q - z_q\|_2. \tag{7}$$

For $\|m_c\|_2 \le 1$, let $u_\alpha(c \mid q) := \langle v_q^\star, m_c\rangle$, $\widehat{s}_\alpha(c \mid q) := \langle \widehat{v}_q, m_c\rangle$, and $U_q(S) := \sum_{c\in S} u_\alpha(c \mid q)$. Here α controls reliance on latent intent and sensitivity to its estimation error.

Theorem 2 (Intent-estimation error controls retrieval regret)**.** *Let $S_q^\star$ and $\widehat{S}_q$ be the top-b sets under u_α and $\widehat{s}_\alpha$. Then*

$$|\widehat{s}_\alpha(c \mid q) - u_\alpha(c \mid q)| \le \alpha\eta_q, \qquad U_q(S_q^\star) - U_q(\widehat{S}_q) \le 2b\alpha\eta_q. \tag{8}$$

Furthermore, if $\Delta_b(q) := \min_{c\in S_q^\star} u_\alpha(c \mid q) - \max_{d\notin S_q^\star} u_\alpha(d \mid q) > 2\alpha\eta_q$, then $\widehat{S}_q = S_q^\star$.

Proof sketch. The query component cancels, leaving $\alpha(\widehat{z}_q - z_q)$; Cauchy–Schwarz gives the score bound. Applying it to both top-b sets yields the factor $2b\alpha\eta_q$. At the selection boundary, two scores can move by at most $2\alpha\eta_q$ in total, so a larger margin preserves the ranking. See Appendix A.4. □

Hence $\alpha\eta_q$ quantifies how intent error propagates to retrieval. This motivates EchoRAG's anchor mechanism, which retains the explicit query while incorporating induced intent. Appendix A.5 proves strict improvement under the margin condition when query-only retrieval selects a different set, and Appendix A.6 extends the guarantee to fused scores.

3 Methodology

Motivated by the theoretical insights in Section 2, we propose EchoRAG, a self-evolving retrieval-augmented generation framework. As shown in Figure 1, EchoRAG decouples the problem into two complementary levels: **Knowledge Organization** and **Knowledge Utilization**.

3.1 Initialization (Offline)

3.1.1 Topology-Based Warm Start

Before serving any queries, the organization operator $\mathcal{O}$ must construct a robust initial knowledge base $\mathcal{K}^{(0)}$ from the raw corpus $\mathcal{G}$. We construct $\mathcal{G}$ from the corpus and run the Leiden algorithm to

establish a topology-based initial partition. Let $m = |\mathcal{E}|$, d_i denote the degree of node i, and c_i its community assignment. The structural cohesion of a partition $\mathcal{C}$ is measured by its modularity:

$$Q(\mathcal{C}) = \frac{1}{2m} \sum_{i,j \in \mathcal{V}} \left(\mathcal{A}_{ij} - \frac{d_i d_j}{2m} \right) \mathbb{I}[c_i = c_j]. \tag{9}$$

Leiden returns an initial partition $\mathcal{C}^{(0)}$ by approximately maximizing $Q(\mathcal{C})$. Based on this partition, EchoRAG initializes each community's membership, summary, pooled representation, and evolution memory. The memory stores its topological and semantic state to inform future maintenance. Isolated but informative nodes are preserved in a residual community to prevent information loss.

3.1.2 Collaborative Node-Movement Engine

Relying solely on the initial structural modularity is demonstrably suboptimal, as established by Proposition 1. To correct these topological misalignments, the knowledge base requires continuous refinement. Consequently, we define a core evolution engine that acts as a standalone, callable tool ready to be triggered during subsequent maintenance cycles.

An intuitive approach to graph refinement would be invoking the language model directly to reorganize the entire knowledge base based on changing demands. However, such global LLM intervention is computationally prohibitive. To circumvent this exorbitant cost, we formulate a collaborative node-movement engine. For any candidate node movement $\delta = (v : c_s \to c_d)$, let $\mathcal{C}_\delta$ denote the partition after moving v. Its structural gain is the resulting change in modularity:

$$\Delta Q(\delta) = Q(\mathcal{C}_\delta) - Q(\mathcal{C}). \tag{10}$$

Simultaneously, EchoRAG computes a lightweight vector-based semantic gain $\Delta\mathcal{S}_0(\delta)$. We introduce a wake-up indicator $g(\delta) = 1$ strictly for cases where these two signals disagree, signifying a structural-semantic conflict. When awakened, the LLM acts as an accurate referee to produce a refined semantic gain $\Delta\mathcal{S}_{\mathrm{LLM}}(\delta)$. A binary verifier $v_{\mathrm{sem}}(\delta)$ filters out potential LLM hallucinations using historical reliability heuristics. The effective semantic gain $\Delta\mathcal{S}(\delta)$ is thus intelligently gated:

$$\Delta\mathcal{S}(\delta) = \begin{cases} \Delta\mathcal{S}_0(\delta), & \text{if } g(\delta) = 0, \\ \Delta\mathcal{S}_{\mathrm{LLM}}(\delta), & \text{if } g(\delta) = 1 \ \wedge \ v_{\mathrm{sem}}(\delta) = 1, \\ 0, & \text{if } g(\delta) = 1 \ \wedge \ v_{\mathrm{sem}}(\delta) = 0. \end{cases} \tag{11}$$

After normalizing the gains over current candidates, we formulate a collaborative objective parameterized by $\lambda \in [0, 1]$ that elegantly balances topology and semantics:

$$\Delta\mathcal{J}(\delta) = (1 - \lambda)\widetilde{\Delta Q}(\delta) + \lambda\widetilde{\Delta\mathcal{S}}(\delta). \tag{12}$$

A movement is accepted iff $\Delta\mathcal{J}(\delta) > 0$. During initialization, EchoRAG iterates these movements globally until convergence, lazily regenerating mutated summaries to finalize $\mathcal{K}^{(0)}$.

3.2 Phase I: Intent-Harmonized Online Serving (Online)

With the initialization, the online utilization operator $\mathcal{U}$ handles individual user queries. Guided by Theorem 2, this phase aims to strictly bound retrieval regret by estimating the latent user intent.

3.2.1 Intent Induction and Anchor Mechanism

Raw user queries are often brief and lexically ambiguous, while their true latent intent $I(q)$ is unobservable. EchoRAG therefore approximates it by prompting the LLM to produce a structured intent blueprint $\widehat{I}(q)$ (e.g., answer type and target entities) together with a hypothetical answer template h_q (Gao et al., 2023a). Both outputs serve only as retrieval proxies and are never treated as factual evidence. We combine these inferred intent signals with the explicit query to form a robust intent representation:

$$\mathbf{e}_I = \mathrm{Norm}\Big((1 - \alpha)\,\mathrm{Enc}(q) + \alpha\,\mathrm{Enc}(\mathrm{Serialize}(\widehat{I}(q), h_q)) \Big). \tag{13}$$

This convex combination acts as a critical safety mechanism. Relying entirely on LLM expansion risks semantic drift, whereas relying solely on the raw query suffers from vocabulary mismatch. The parameter $\alpha \in [0, 1]$ dictates the balance: the explicit query firmly anchors the search, while the inferred intent pulls the direction toward deeper semantic needs.

3.2.2 RETRIEVAL AND DEMAND LOGGING

EchoRAG fuses the semantic intent score with a deterministic linked-entity coverage score to retrieve the most valuable communities. However, beyond merely driving retrieval, the intent vector $\mathbf{e}_I$ silently operates as a continuous monitor of community demand.

For each query q_t, we compute its semantic similarity to community representations and distribute exactly one unit of query mass among the top-k candidates via a temperature-scaled softmax $a_{t,c}$. Averaging these query-level assignments over a temporal window of B queries naturally yields the empirical demand profile:

$$\boldsymbol{\pi}^{(r+1)}(c) = \frac{1}{B} \sum_{t=rB+1}^{(r+1)B} a_{t,c}. \tag{14}$$

This mechanism tracks user interests at no additional computational cost. Rather than relying on an independent monitoring module, the online retrieval process inherently tallies community usage as a byproduct, safely batching it as the quantitative signal required for the evolution in Phase II.

3.2.3 GROUNDED EVIDENCE HARMONIZATION

Naively concatenating the retrieved community summaries would bombard the generator with broad, unfocused topics. To provide precise context, EchoRAG transforms the raw retrieval into an intent-aligned view through grounded knowledge harmonization.

The harmonizing model extracts fine-grained evidence cards from the raw material based strictly on the intent blueprint $\widehat{I}(q)$. To eliminate hallucination risks, a strict verification function discards any card lacking an exact text-span citation from the source view. A ranking operator then orders the surviving cards by relevance. The final generator consequently consumes a perfectly harmonized and factually anchored context to synthesize the response.

3.3 PHASE II: DEMAND-TRIGGERED GRAPH EVOLUTION (ONLINE)

As user interests shift over time, the optimal graph structure drifts. Rather than rebuilding the graph blindly, EchoRAG utilizes the empirical demand profile $\boldsymbol{\pi}^{(r+1)}$ logged during Phase I to determine exactly where and when to trigger graph maintenance.

Backed by the selective-update certificate from Theorem 1, we compute a symmetric demand shift score that acts as a community-level update thermometer:

$$u_c^{(r)} = \mathbb{I}[N_c^{(r)} \geq N_{\min}] \frac{|\boldsymbol{\pi}^{(r+1)}(c) - \boldsymbol{\pi}^{(r)}(c)|}{\boldsymbol{\pi}^{(r+1)}(c) + \boldsymbol{\pi}^{(r)}(c) + \varepsilon}. \tag{15}$$

Here, $N_c^{(r)}$ counts the queries in window r for which community c appears among the top-k retrieval candidates. The support gate ensures that the intensity is set to zero for insufficiently observed communities, preventing sparse traffic from triggering costly updates (as theoretically supported by Corollary 1). Given two thresholds $\tau_D < \tau_H$, EchoRAG applies a tiered maintenance policy governed by the continuous score $u_c^{(r)}$, seamlessly deriving the next reusable knowledge base $\mathcal{K}^{(r+1)}$:

$$\mathcal{K}^{(r+1)} = \mathrm{Evolve}_{\Delta\mathcal{J}}\Big(\mathrm{Refresh}(\mathcal{K}^{(r)}; \mathbf{u}^{(r)}, \mathcal{Z}^{(r+1)}); \mathcal{G}, \mathbf{u}^{(r)}\Big). \tag{16}$$

Depending on the shift intensity, the system executes one of three actions: (1) **Keep ($u_c^{(r)} \leq \tau_D$):** The community topology, summary, representation, and memory remain unchanged, without invoking the LLM. (2) **Refresh ($\tau_D < u_c^{(r)} < \tau_H$):** EchoRAG keeps the community membership fixed but regenerates its summary from the source evidence conditioned on recent intents $\mathcal{Z}^{(r+1)}$, then updates its representation and memory. (3) **Evolve ($u_c^{(r)} \geq \tau_H$):** EchoRAG applies the node-movement engine from Section 3.1 to the affected subgraph, performs the required splits or merges, and regenerates the summaries and representations.

This decoupled, feedback-driven design ensures the graph topology dynamically and safely mirrors real-world usage without exorbitant computational overhead.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTING

We evaluate EchoRAG on Multihop-RAG (Tang & Yang, 2024), HotpotQA (Yang et al., 2018), and NarrativeQA (Kočiský et al., 2017), covering multi-hop and long-document QA, against inference-only, retrieval-only, and representative graph-based RAG baselines. Full dataset statistics and evaluated subsets, baseline descriptions, metric definitions, and implementation details are provided in Appendix C. Note that for a comprehensive evaluation of EchoRAG, additional experiments are presented in Appendix D.

4.2 OVERALL EFFECTIVENESS

We benchmark EchoRAG against state-of-the-art baselines across three diverse datasets, shown in Tables 1 and 2. From these results, we derive three key observations.

Table 1: Performance on conventional QA metrics (%).

Method	Multihop-RAG			HotpotQA			NarrativeQA		
	Accuracy	F1	Hit	Accuracy	F1	Hit	Accuracy	BLEU-1	METEOR
Zero-shot	52.1	52.5	N/A	55.4	56.3	N/A	10.5	7.6	12.8
CoT	47.2	46.4	N/A	55.5	58.6	N/A	11.7	9.1	12.7
BM25	21.6	23.4	33.2	22.0	49.6	76.1	3.0	2.7	3.1
Vanilla RAG	32.6	28.2	35.3	38.0	37.7	67.5	16.6	9.5	12.8
RAPTOR	55.1	43.7	38.7	49.6	49.2	18.4	23.2	8.7	13.0
GraphRAG	66.8	45.8	39.1	74.9	80.2	76.4	39.4	7.5	10.8
LightRAG	69.2	54.3	45.9	82.0	78.8	76.2	49.9	9.1	13.7
HippoRAG	67.3	**71.3**	65.7	78.2	75.2	93.3	27.4	7.2	9.8
ArchRAG	51.0	58.9	63.2	59.0	63.6	60.5	37.7	7.8	16.4
DA-RAG	60.9	44.4	42.5	73.5	76.1	87.3	35.7	7.0	9.7
LinearRAG	60.3	65.6	61.1	79.8	84.7	69.1	21.7	6.2	7.7
HyperRAG	58.1	67.3	65.4	48.1	52.1	88.3	14.4	9.3	11.8
CogRAG	51.3	35.8	31.5	69.7	71.9	43.1	13.2	5.9	7.4
EchoRAG (Ours)	**73.7**	69.1	**68.2**	**85.1**	**91.6**	**96.0**	**51.2**	**10.2**	**17.3**

Table 2: Performance under LLM-based evaluation (%).

Method	Multihop-RAG				HotpotQA				NarrativeQA			
	Comp.	Div.	Emp.	Over.	Comp.	Div.	Emp.	Over.	Comp.	Div.	Emp.	Over.
Zero-shot	38.6	28.2	28.5	34.4	31.6	20.8	30.7	52.5	26.6	25.6	24.9	24.8
CoT	37.0	26.0	26.1	32.3	46.5	29.8	32.4	40.2	24.6	22.1	24.3	27.4
BM25	23.4	26.8	26.4	23.3	26.5	27.0	26.3	25.0	20.9	21.9	21.2	20.8
Vanilla RAG	43.1	29.9	30.0	33.5	40.4	23.2	40.1	47.1	36.1	27.5	27.0	27.6
RAPTOR	53.5	51.0	50.8	49.9	54.2	46.3	46.0	47.9	29.0	32.2	28.3	26.9
GraphRAG	54.0	60.3	58.5	51.5	69.2	70.2	68.3	63.7	35.3	45.2	40.1	32.2
LightRAG	56.5	58.7	57.0	53.4	75.5	71.6	70.4	69.1	37.5	44.2	40.1	33.7
HippoRAG	51.7	52.8	51.0	47.9	64.1	57.6	56.7	56.4	27.8	32.8	29.3	26.5
ArchRAG	57.8	58.5	56.3	54.0	54.6	59.2	55.7	51.9	30.4	40.2	35.6	29.5
DA-RAG	49.4	59.4	56.3	46.4	61.5	67.6	63.9	55.7	34.2	46.1	40.1	30.9
LinearRAG	51.9	48.2	46.9	47.5	66.1	59.8	59.4	58.1	26.4	30.8	27.5	25.3
HyperRAG	41.3	32.3	32.0	37.6	49.6	33.7	34.7	39.9	26.2	22.9	22.9	23.7
CogRAG	45.5	57.7	55.0	42.1	56.9	63.3	59.8	49.9	26.3	37.0	32.2	24.4
EchoRAG (Ours)	**62.2**	**67.8**	**65.7**	**58.0**	**76.8**	**78.3**	**75.9**	**69.8**	**38.1**	**47.9**	**43.4**	**34.6**

Obs. 1. EchoRAG consistently improves answer correctness across datasets. EchoRAG achieves state-of-the-art performance across almost all conventional metrics (Table 1). Notably, it outperforms the strongest baseline in Accuracy by 6.5%, 3.8%, and 2.6% on Multihop-RAG, HotpotQA, and NarrativeQA, respectively. These consistent gains across diverse reasoning settings (from multi-hop queries to long-document QA) demonstrate the superiority of EchoRAG's query-driven design. By simultaneously adapting the knowledge to demand distributions and anchoring the

online retrieval to latent user intents, EchoRAG provides a far more robust foundation for accurate fact retrieval than static, query-blind knowledge bases.

Obs. 2. EchoRAG produces higher-quality responses beyond exact-match correctness. EchoRAG dominates all LLM-based qualitative evaluations in Table 2. This superiority is particularly pronounced on Multihop-RAG, where it surpasses the best baseline by a striking 7.4%–12.4% across all four evaluation dimensions. Unlike traditional methods restricted to surface-level lexical matching, EchoRAG deeply infers and harmonizes the latent intent behind the query. This intent-aware mechanism ensures that the generated responses are not just factually accurate, but also comprehensively well-rounded, informative, and structurally aligned with user expectations.

Obs. 3. Effective evidence utilization matters more than raw retrieval coverage. Retrieval metrics do not fully determine response quality. On HotpotQA, EchoRAG improves Hit over HippoRAG by only 2.9%, but improves Accuracy and Overall Quality by 8.8% and 23.8%, respectively. On Multihop-RAG, it achieves 9.5% higher Accuracy despite a 3.1% lower F1 score. These results suggest that query-adaptive organization and intent-aware filtering improve how retrieved evidence is used, rather than simply maximizing coverage.

4.3 Ablation Study

We ablate query-guided community evolution (QCE) and intent-harmonized knowledge utilization (IHU) to assess their contributions. Table 3 reports results across three datasets; component-level ablations are provided in the Appendix. Removing either module reduces all metrics, showing that both contribute to performance. On Multihop-RAG, QCE has a larger effect on F1, Diversity, and Overall Quality, whereas IHU contributes more to Accuracy. This suggests complementary roles: QCE organizes dispersed multi-step evidence, while IHU filters it for answer relevance. On HotpotQA, removing IHU causes a larger and more consistent decline, indicating that intent alignment is particularly important for distinguishing distractors among topically similar entities.

Table 3: Module-level ablation results on Multihop-RAG, HotpotQA, and NarrativeQA.

Method	Multihop-RAG				HotpotQA				NarrativeQA			
	Acc.	F1	Div.	Over.	Acc.	F1	Div.	Over.	Acc.	F1	Div.	Over.
EchoRAG	**73.7**	**69.1**	**67.8**	**58.0**	**85.1**	**91.6**	**78.3**	**69.8**	**51.2**	**52.5**	**48.4**	**36.8**
w/o QCE	69.5	63.6	60.4	54.1	80.6	88.1	74.1	62.3	43.2	43.2	41.9	29.9
w/o IHU	66.4	67.5	67.2	57.4	77.4	85.0	73.4	61.9	45.1	45.5	43.1	30.4

4.4 Motivation Validation

Under matched retrieval and generation settings, we compare structure-only with query-guided organization, query-only with intent-harmonized utilization, and their 2×2 combinations. Figure 2 reports the main results, with additional analyses in the Appendix. In Figure 2(a), query-guided evolution improves Evidence Recall by 50.1% and more than doubles Semantic Coherence and Answer Sufficiency, indicating that structural connectivity alone does not ensure generation-ready evidence. In Figure 2(b), intent harmonization improves Supporting-Fact Recall by 21.4%, together with Exact Match and Answer F1, suggesting more effective intent-specific evidence selection. The full combination performs best in Figure 2(c); its positive interaction of 0.010 on both metrics supports the complementary roles of organization and utilization.

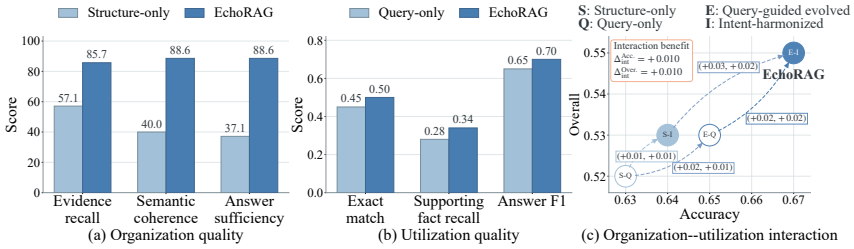
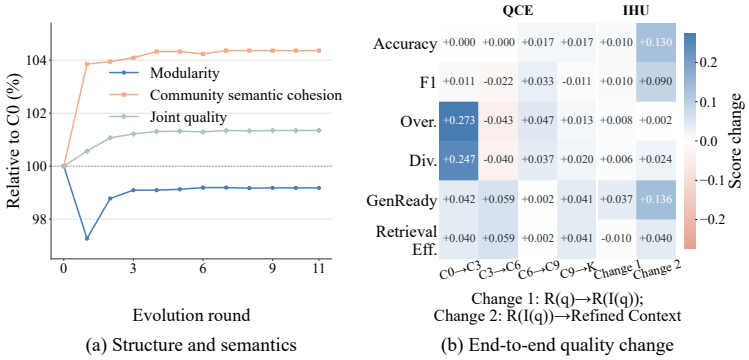


Figure 2: Results of the experiments validating the motivation of EchoRAG.

4.5 Self-Evolving Analysis

We checkpoint community states across QCE rounds and trace each query through direct retrieval $R(q)$, intent-conditioned retrieval $R(I(q))$, and grounded refinement. Figure 3(a) shows that Semantic Cohesion and Joint Quality increase, while Modularity briefly declines before recovering near its initial level. Thus, QCE improves semantic organization with limited structural loss. In Figure 3(b), QCE achieves its largest Overall and Diversity gains during $C^0 \rightarrow C^3$, followed by smaller updates. Within IHU, intent-conditioned retrieval improves generation readiness despite slightly lower retrieval effectiveness, while grounded refinement yields the largest gains in Accuracy, F1, and generation readiness. Overall, QCE improves community organization, and IHU converts it into query-specific evidence.



(a) Structure and semantics (b) End-to-end quality change

Figure 3: Self-evolution of EchoRAG. (a) Structural and semantic quality across QCE rounds. (b) Downstream quality across QCE and IHU stages.

4.6 Cost-Effectiveness Analysis

Figure 4 compares token consumption and monetary cost across graph-based RAG methods on Multihop-RAG. EchoRAG incurs more overhead than lightweight baselines because of localized LLM calls for community evolution and intent harmonization, but uses fewer tokens than SOTA methods that reason over large retrieved contexts. Together with its accuracy and quality gains, these results indicate a favorable performance–cost trade-off.

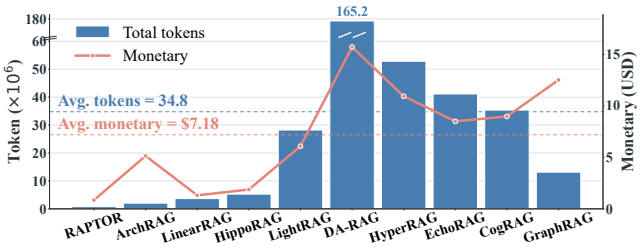


Figure 4: Token consumption and monetary cost of RAG methods on Multihop-RAG.

5 Conclusion

This work presents EchoRAG, a self-evolving graph-based RAG framework that transforms knowledge organization and utilization from isolated stages into a unified adaptive process. By continuously incorporating query feedback into knowledge-base maintenance and reshaping retrieved knowledge around the information required for generation, EchoRAG produces reusable communities and grounded evidence that become progressively better aligned with downstream needs, without reconstructing the knowledge base for every query. Extensive experiments and analyses across three benchmarks demonstrate consistent improvements over state-of-the-art baselines, validating adaptive knowledge evolution as an effective direction for graph-based RAG.

AI USE STATEMENT

In this work, we used generative AI tools as writing aids to improve the clarity and readability of author-written text and identify grammatical, typographical, and stylistic errors. Generative AI tools were not used to generate data, conduct experiments, determine reported numerical results, implement the proposed method, or formulate mathematical claims and proofs. We manually reviewed and revised this passage and verified its technical and factual content against the cited literature, derivations, code, tables, figures, and experimental records. We take full responsibility for the accuracy, originality, and final content of this paper.

ETHICS STATEMENT

This work does not involve human-subject studies or the collection of private or sensitive personal data; the experiments use established research benchmarks and their associated corpora. Nevertheless, the source corpora and underlying language models may contain factual errors, copyrighted material, harmful content, or societal biases, which may be reproduced or amplified in generated answers. EchoRAG's source-grounded evidence mechanism is intended to improve traceability and reduce unsupported generation, but it does not eliminate these risks. The method should therefore not be used for high-stakes decisions without domain-specific validation and appropriate human oversight. Deployments on proprietary or user-provided corpora should additionally enforce applicable licenses, consent requirements, privacy protections, and access controls. We conducted this research in accordance with the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

To facilitate reproduction, Section 1 defines the graph-based RAG setting and motivates EchoRAG; related work is discussed in Appendix B. Section 2 states the theoretical results and gives proof sketches, with full proofs in Appendix A. Section 3 describes initialization, intent-harmonized online serving, and demand-triggered graph evolution. Section 4 reports the experimental comparisons and analyses. Appendix C describes the datasets and evaluated subsets, baselines, metrics, model and decoding settings, computing environment, and random seed. Appendix D contains additional effectiveness results, fine-grained ablations, indexing and query-time measurements, hyperparameter sensitivity, backbone comparisons, and case studies. Appendix E provides prompt templates for semantic node-movement judgment, intent induction, grounded evidence refinement, and final answer generation. Code is linked at `https://anonymous.4open.science/r/EchoRAG`.

REFERENCES

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113, 2023.

Anjali de Silva, Gang Chen, Hui Ma, Seyed Mohammad Nekooei, and Xingquan Zuo. Advancing community detection with graph convolutional neural networks: Bridging topological and attributive cohesion. *arXiv preprint arXiv:2505.10197*, 2025.

Xinyi Duan, Yuanrong Tang, and Jiangtao Gong. Grouprag: Cognitively inspired group-aware retrieval and reasoning via knowledge-driven problem structuring. *arXiv preprint arXiv:2603.26807*, 2026.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Junfeng Fang, Yukai Wang, Ruipeng Wang, Zijun Yao, Kun Wang, An Zhang, Xiang Wang, and Tat-Seng Chua. Safemlrm: Demystifying safety in multi-modal large reasoning models. *arXiv preprint arXiv:2504.08813*, 2025.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1762–1777, 2023a.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1), 2023b.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*, 2024.

Bernal J Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in neural information processing systems*, 37:59532–59569, 2024.

Rujun Han, Yuhao Zhang, Peng Qi, Yumo Xu, Jenyuan Wang, Lan Liu, William Yang Wang, Bonan Min, and Vittorio Castelli. Rag-qa arena: Evaluating domain robustness for long-form retrieval augmented question answering. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 4354–4374, 2024.

Hao Hu, Yifan Feng, Ruoxue Li, Rundong Xue, Xingliang Hou, Zhiqiang Tian, Yue Gao, and Shaoyi Du. Cog-rag: Cognitive-inspired dual-hypergraph with theme alignment retrieval-augmented generation. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence*, 2026.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. *arXiv preprint arXiv:2403.14403*, 2024.

Songwei Jia, Lin Gao, Yong Gao, James Nastos, Yijie Wang, Xindong Zhang, and Haiyang Wang. Defining and identifying cograph communities in complex networks. *New Journal of Physics*, 17 (1):013044, 2015.

Houcheng Jiang, Junfeng Fang, Ningyu Zhang, Guojun Ma, Mingyang Wan, Xiang Wang, Xiangnan He, and Tat-seng Chua. Anyedit: Edit any knowledge encoded in language models. *arXiv preprint arXiv:2502.05628*, 2025a.

Wenqi Jiang, Shuai Zhang, Boran Han, Jie Wang, Bernie Wang, and Tim Kraska. Piperag: Fast retrieval-augmented generation via adaptive pipeline parallelism. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 589–600, 2025b.

Di Jin, Zhizhi Yu, Pengfei Jiao, Shirui Pan, Dongxiao He, Jia Wu, Philip S Yu, and Weixiong Zhang. A survey of community detection approaches: From statistical modeling to deep learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(2):1149–1170, 2021.

Imran Kabir, Md Alimoor Reza, and Syed Billah. Logic-rag: Augmenting large multimodal models with visual-spatial knowledge for road scene understanding. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10585–10591. IEEE, 2025.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The narrativeqa reading comprehension challenge, 2017. URL `https://arxiv.org/abs/1712.07040`.

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

Yuhan Li, Zhixun Li, Peisong Wang, Jia Li, Xiangguo Sun, Hong Cheng, and Jeffrey Xu Yu. A survey of graph meets large language model: Progress and future directions. *arXiv preprint arXiv:2311.12399*, 2023.

Zhaodonghui Li, Haitao Yuan, Huiming Wang, Gao Cong, and Lidong Bing. Llm-r2: A large language model enhanced rule-based rewrite system for boosting query efficiency. *arXiv preprint arXiv:2404.12872*, 2024.

Wen-Sheng Lien, Yu-Kai Chan, Hao-Lung Hsiao, Bo-Kai Ruan, Meng-Fen Chiang, Chien-An Chen, Yi-Ren Yeh, and Hong-Han Shuai. Hyperrag: Reasoning n-ary facts over hypergraphs for retrieval augmented generation. In *Proceedings of the ACM Web Conference 2026*, pp. 2465–2476, 2026.

Bo-Young Lim, Jeong-Ha Park, Kisung Lee, and Hyuk-Yoon Kwon. Multi-level graph representation learning through predictive community-based partitioning. *Proceedings of the ACM on Management of Data*, 3(1):1–27, 2025.

Qing Liu, Minjun Zhao, Xin Huang, Jianliang Xu, and Yunjun Gao. Truss-based community search over large directed graphs. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pp. 2183–2197, 2020.

Shengjie Ma, Chengjin Xu, Xuhui Jiang, Muzhi Li, Huaren Qu, Cehao Yang, Jiaxin Mao, and Jian Guo. Think-on-graph 2.0: Deep and faithful large language model reasoning with knowledge-guided retrieval augmented generation, 2025. URL https://arxiv.org/abs/2407.10805.

Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

Li Ni, Rui Ye, Wenjian Luo, Yiwen Zhang, Lei Zhang, and Victor S Sheng. Slrl: Semi-supervised local community detection based on reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 631–639, 2025.

Qiuyi Qi, Tian Liang, Jiamu Wang, Jinjian Zhang, Wei Zhou, Pengcheng Zhu, Linjian Mo, Ming Kong, Jie Liu, and Qiang Zhu. Metarag: Belief-action aligned policy optimization for agentic rag. *arXiv preprint arXiv:2608.24214*, 2026.

Hongjin Qian, Zheng Liu, Peitian Zhang, Kelong Mao, Defu Lian, Zhicheng Dou, and Tiejun Huang. Memorag: Boosting long context processing with global memory-enhanced retrieval augmentation. In *Proceedings of the ACM on Web Conference 2025*, pp. 2366–2377, 2025.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*, volume 500-225 of *NIST Special Publication*, pp. 109–126. National Institute of Standards and Technology, 1994.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher Manning. Raptor: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations*, volume 2024, pp. 32628–32649, 2024a.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. Raptor: Recursive abstractive processing for tree-organized retrieval, 2024b. URL https://arxiv.org/abs/2401.18059.

Mauro Sozio and Aristides Gionis. The community-search problem and how to plan a successful cocktail party. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 939–948, 2010.

Yixuan Tang and Yi Yang. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries, 2024. URL https://arxiv.org/abs/2401.15391.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233, 2019.

Jinyu Wang, Jingjing Fu, Rui Wang, Lei Song, and Jiang Bian. Pike-rag: specialized knowledge and rationale augmented generation. *arXiv preprint arXiv:2501.11551*, 2025a.

Shu Wang, Yixiang Fang, Yingli Zhou, Xilin Liu, and Yuchi Ma. Archrag: Attributed community-based hierarchical retrieval-augmented generation. *arXiv preprint arXiv:2502.09891*, 2025b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Shangyu Wu, Ying Xiong, Yufei Cui, Haolun Wu, Can Chen, Ye Yuan, Lianming Huang, Xue Liu, Tei-Wei Kuo, Nan Guan, et al. Retrieval-augmented generation for natural language processing: A survey. *arXiv preprint arXiv:2407.13193*, 2024.

Yilin Xiao, Junnan Dong, Chuang Zhou, Su Dong, Qian-wen Zhang, Di Yin, Xing Sun, and Xiao Huang. Graphrag-bench: Challenging domain-specific reasoning for evaluating graph retrieval-augmented generation. *arXiv preprint arXiv:2506.02404*, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering, 2018. URL https://arxiv.org/abs/1809.09600.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

Xingyuan Zeng, Zuohan Wu, Yue Wang, Chen Zhang, Quanming Yao, Libin Zheng, and Jian Yin. Da-rag: Dynamic attributed community search for retrieval-augmented generation. In *Proceedings of the ACM Web Conference 2026*, pp. 2195–2206, 2026.

Anran Zhang, Xingfen Wang, and Yuhan Zhao. Hacd: Harnessing attribute semantics and mesoscopic structure for community detection. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pp. 616–624, 2025a.

Qinggang Zhang, Junnan Dong, Hao Chen, Daochen Zha, Zailiang Yu, and Xiao Huang. Knowgpt: Knowledge graph based prompting for large language models. *Advances in neural information processing systems*, 37:6052–6080, 2024.

Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Hao Chen, Yilin Xiao, Chuang Zhou, Junnan Dong, et al. A survey of graph retrieval-augmented generation for customized large language models. *arXiv preprint arXiv:2501.13958*, 2025b.

Zhuoping Zhou, Davoud Ataee Tarzanagh, Sima Didari, Wenjun Hu, Baruch Gutow, Oxana Verkholyak, Masoud Faraki, Heng Hao, Hankyu Moon, and Seungjai Min. Query-aware flow diffusion for graph-based rag with retrieval guarantees, 2026. URL https://arxiv.org/abs/2605.18775.

Luyao Zhuang, Shengyuan Chen, Yilin Xiao, Huachi Zhou, Yujing Zhang, Hao Chen, Qinggang Zhang, and Xiao Huang. Linearrag: Linear graph retrieval augmented generation on large-scale corpora. *arXiv preprint arXiv:2510.10114*, 2025.

A Detailed Proofs for the Theoretical Analysis

A.1 Proof of Theorem 1

Proof. We prove the three claims in order.

1. Nonnegativity and simultaneous optimality. For any $\pi \in \Delta^{C-1}$, $\mathcal{K}_0$ is feasible. Therefore,

$$\min_{\mathcal{K} \in \mathfrak{K}_B(\mathcal{G})} \mathcal{R}_\pi(\mathcal{K}) \le \mathcal{R}_\pi(\mathcal{K}_0),$$

which proves $\Gamma_{\mathrm{org}}(\pi; \mathcal{K}_0) \ge 0$.

Suppose first that $\mathcal{K}_0$ simultaneously minimizes every coordinate loss: for all $c \in [C]$ and all feasible $\mathcal{K}$, $L_c(\mathcal{K}_0) \le L_c(\mathcal{K})$. Multiplying these inequalities by the nonnegative weights $\pi(c)$ and summing gives

$$\mathcal{R}_\pi(\mathcal{K}_0) = \sum_{c=1}^{C} \pi(c) L_c(\mathcal{K}_0) \le \sum_{c=1}^{C} \pi(c) L_c(\mathcal{K}) = \mathcal{R}_\pi(\mathcal{K})$$

for every feasible $\mathcal{K}$. Thus $\Gamma_{\mathrm{org}}(\pi; \mathcal{K}_0) = 0$ for every π.

Conversely, assume $\Gamma_{\mathrm{org}}(\pi; \mathcal{K}_0) = 0$ for every $\pi \in \Delta^{C-1}$. Fix an arbitrary $c_0 \in [C]$ and choose the simplex vertex $\pi = e_{c_0}$. Then $\mathcal{R}_{e_{c_0}}(\mathcal{K}) = L_{c_0}(\mathcal{K})$, so zero regret implies

$$L_{c_0}(\mathcal{K}_0) = \min_{\mathcal{K} \in \mathfrak{K}_B(\mathcal{G})} L_{c_0}(\mathcal{K}).$$

Because c_0 was arbitrary, $\mathcal{K}_0$ simultaneously minimizes all coordinate losses.

2. Stability under workload drift. Let $\delta_c := L_c(\mathcal{K}) - L_c(\mathcal{K}')$. Since both losses lie in $[0, L]$, $|\delta_c| \le L$. Expanding the difference gives

$$D_{\pi_1}(\mathcal{K}, \mathcal{K}') - D_{\pi_0}(\mathcal{K}, \mathcal{K}') = \sum_{c=1}^{C} [\pi_1(c) - \pi_0(c)] \delta_c.$$

Hence, by the triangle inequality (equivalently, Hölder's inequality),

$$\begin{aligned} |D_{\pi_1}(\mathcal{K}, \mathcal{K}') - D_{\pi_0}(\mathcal{K}, \mathcal{K}')| &\le \sum_{c=1}^{C} |\pi_1(c) - \pi_0(c)| \, |\delta_c| \\ &\le L \sum_{c=1}^{C} |\pi_1(c) - \pi_0(c)| \\ &= L \|\pi_1 - \pi_0\|_1. \end{aligned}$$

3. Selective-update certificate. For every $c \notin \mathcal{A}_\tau$, the definition of the selected set implies

$$|\pi_1(c) - \pi_0(c)| \le \tau [\pi_1(c) + \pi_0(c) + \epsilon].$$

Using again $|\delta_c| \le L$, we obtain

$$\begin{aligned} &\left| \sum_{c \notin \mathcal{A}_\tau} [\pi_1(c) - \pi_0(c)][L_c(\mathcal{K}) - L_c(\mathcal{K}')] \right| \\ &\quad \le L \sum_{c \notin \mathcal{A}_\tau} |\pi_1(c) - \pi_0(c)| \\ &\quad \le L\tau \sum_{c \notin \mathcal{A}_\tau} [\pi_1(c) + \pi_0(c) + \epsilon] \\ &\quad \le L\tau \left(\sum_{c=1}^{C} \pi_1(c) + \sum_{c=1}^{C} \pi_0(c) + C\epsilon \right) \\ &\quad = L\tau(2 + C\epsilon), \end{aligned}$$

where the final equality uses $\pi_0, \pi_1 \in \Delta^{C-1}$. □

A.2 Proof of Proposition 1

For a partition $\mathcal{K}$ and required evidence set $T(q)$, the normalized evidence-fragmentation risk under a one-community retrieval budget is

$$\mathcal{F}_\mu(\mathcal{K}) := \mathbb{E}_{Q\sim\mu}\left[\mathbf{1}\{\nexists K \in \mathcal{K} : T(Q) \subseteq K\}\right] \in [0,1]. \tag{17}$$

Proof. Consider the undirected graph with vertex set $\{a,b,c,d\}$ and edge set $\{(a,b),(c,d)\}$. It has $m = 2$ edges, and every vertex has degree one. For any partition $\mathcal{K}$, Newman–Girvan modularity can be written as

$$Q(\mathcal{K}) = \sum_{K\in\mathcal{K}} \left(\frac{\ell_K}{m} - \left(\frac{d_K}{2m}\right)^2 \right),$$

where ℓ_K is the number of edges internal to K and d_K is the sum of degrees in K. In this graph, $d_K = |K|$, and therefore

$$Q(\mathcal{K}) = \frac{k(\mathcal{K})}{2} - \frac{1}{16}\sum_{K\in\mathcal{K}} |K|^2,$$

where $k(\mathcal{K}) \in \{0,1,2\}$ is the number of graph edges whose endpoints lie in the same community.

If $k(\mathcal{K}) = 2$, either the two edges form separate size-two communities, which gives

$$Q = 1 - \frac{2^2+2^2}{16} = \frac{1}{2},$$

or all four vertices lie in a single community, which gives $Q = 1 - 4^2/16 = 0$. If $k(\mathcal{K}) = 1$, the sum of squared community sizes is minimized by one size-two community and two singletons; hence

$$Q(\mathcal{K}) \le \frac{1}{2} - \frac{2^2+1^2+1^2}{16} = \frac{1}{8}.$$

If $k(\mathcal{K}) = 0$, then $Q(\mathcal{K}) \le 0$. Consequently, up to the irrelevant ordering of community labels, the unique modularity-maximizing partition is

$$\mathcal{K}_{\mathrm{str}} = \big\{\{a,b\},\{c,d\}\big\}.$$

Now let the workload put probability $1/2$ on each of two queries with required evidence sets $T(q_1) = \{a,c\}$ and $T(q_2) = \{b,d\}$. Neither required pair is contained in a community of $\mathcal{K}_{\mathrm{str}}$. Thus, under the one-community retrieval budget in Eq. equation 17, $\mathcal{F}_\mu(\mathcal{K}_{\mathrm{str}}) = 1$. For the workload-aware partition

$$\mathcal{K}_\mu = \big\{\{a,c\},\{b,d\}\big\},$$

each required pair is itself a community, so $\mathcal{F}_\mu(\mathcal{K}_\mu) = 0$. □

A.3 Finite-Sample Stability of the Demand Trigger

Corollary 1 (Finite-sample stability of the demand trigger)**.** *For $w \in \{0,1\}$, let $Q_{w,1},\dots,Q_{w,n} \overset{\text{i.i.d.}}{\sim} \mu_w$, write $\pi_w := \pi_{\mu_w}$, and let $\widehat{\pi}_w(c)$ be the empirical mean of $a_c(Q)$. For $\delta \in (0,1)$, define $\eta_n := \sqrt{\log(4C/\delta)/(2n)}$. With probability at least $1-\delta$,*

$$\max_{c\in[C]} |\widehat{d}_\epsilon(c) - d_\epsilon(c)| \le \frac{2\eta_n}{\epsilon}, \tag{18}$$

where $\widehat{d}_\epsilon$ replaces π_w by $\widehat{\pi}_w$. Hence the empirical and population threshold decisions agree for every c satisfying $|d_\epsilon(c) - \tau| > 2\eta_n/\epsilon$. If the trigger is additionally gated by a fixed frequency mask, the same statement holds conditional on that mask.

Proof. For every $w \in \{0,1\}$ and $c \in [C]$, the random variable $a_c(Q_{w,i})$ lies in $[0,1]$. Hoeffding's inequality therefore yields

$$\Pr(|\widehat{\pi}_w(c) - \pi_w(c)| > t) \le 2e^{-2nt^2}.$$

A union bound over the $2C$ pairs (w, c) gives

$$\Pr\left(\max_{w\in\{0,1\},\, c\in[C]} |\widehat{\pi}_w(c) - \pi_w(c)| > t\right) \le 4Ce^{-2nt^2}.$$

Setting $t = \eta_n = \sqrt{\log(4C/\delta)/(2n)}$ shows that, with probability at least $1-\delta$,

$$\max_{w,c} |\widehat{\pi}_w(c) - \pi_w(c)| \le \eta_n. \tag{19}$$

It remains to propagate this coordinatewise error through the normalized shift score. Define

$$f_\epsilon(x, y) := \frac{|x-y|}{x+y+\epsilon}, \qquad x, y \ge 0.$$

For fixed y, the derivative with respect to x exists away from $x = y$ and has absolute value at most $1/\epsilon$; the same is true with the two coordinates interchanged. Continuity at $x = y$ then implies the coordinatewise Lipschitz bound

$$|f_\epsilon(x', y') - f_\epsilon(x, y)| \le \frac{|x'-x| + |y'-y|}{\epsilon}.$$

Applying this inequality with $(x, y) = (\pi_1(c), \pi_0(c))$ and $(x', y') = (\widehat{\pi}_1(c), \widehat{\pi}_0(c))$, and then using Eq. equation 19, yields simultaneously for all c,

$$|\widehat{d}_\epsilon(c) - d_\epsilon(c)| \le \frac{2\eta_n}{\epsilon}.$$

Let $\rho := 2\eta_n/\epsilon$. If $d_\epsilon(c) > \tau + \rho$, then $\widehat{d}_\epsilon(c) > \tau$; if $d_\epsilon(c) < \tau - \rho$, then $\widehat{d}_\epsilon(c) < \tau$. Thus the empirical and population decisions agree whenever $|d_\epsilon(c) - \tau| > \rho$. Any additional frequency mask that is held fixed merely restricts the set of communities to which this comparison is applied and hence does not change the argument. □

A.4 Proof of Theorem 2

Proof. From Eq. equation 7,

$$\widehat{v}_q - v_q^\star = \alpha(\widehat{z}_q - z_q).$$

Thus, for every community c, Cauchy–Schwarz and $\|m_c\|_2 \le 1$ give

$$\begin{aligned} |\widehat{s}_\alpha(c \mid q) - u_\alpha(c \mid q)| &= |\langle \widehat{v}_q - v_q^\star, m_c\rangle| \\ &\le \alpha \|\widehat{z}_q - z_q\|_2 \|m_c\|_2 \\ &\le \alpha\eta_q. \end{aligned}$$

This proves the first claim.

For a size-b set S, define $\widehat{U}_q(S) := \sum_{c\in S} \widehat{s}_\alpha(c \mid q)$. Summing the per-community bound over S yields

$$|\widehat{U}_q(S) - U_q(S)| \le b\alpha\eta_q. \tag{20}$$

Because $\widehat{S}_q$ is a top-b set under $\widehat{s}_\alpha$, $\widehat{U}_q(S_q^\star) \le \widehat{U}_q(\widehat{S}_q)$. Consequently,

$$\begin{aligned} U_q(S_q^\star) &\le \widehat{U}_q(S_q^\star) + b\alpha\eta_q \\ &\le \widehat{U}_q(\widehat{S}_q) + b\alpha\eta_q \\ &\le U_q(\widehat{S}_q) + 2b\alpha\eta_q, \end{aligned}$$

where the first and third inequalities use Eq. equation 20. Rearranging proves Eq. equation 8.

Finally, take any $c \in S_q^\star$ and $d \notin S_q^\star$. The score perturbation bound gives

$$\begin{aligned} \widehat{s}_\alpha(c \mid q) - \widehat{s}_\alpha(d \mid q) &\ge u_\alpha(c \mid q) - u_\alpha(d \mid q) - 2\alpha\eta_q \\ &\ge \Delta_b(q) - 2\alpha\eta_q. \end{aligned}$$

When $\Delta_b(q) > 2\alpha\eta_q$, this difference is strictly positive for every such pair (c, d). Therefore all communities in $S_q^\star$ rank above all communities outside it under the estimated score, implying $\widehat{S}_q = S_q^\star$. □

A.5 Strict Separation from Query-Only Retrieval

Corollary 2 (Strict separation from query-only retrieval)**.** *Let $s_{\mathrm{qry}}(c \mid q) := \langle e_q, m_c \rangle$ be the query-only score, and let S_q^{qry} be its top-b set. Assume $\Delta_b(q) > 2\alpha\eta_q$. If $S_q^{\mathrm{qry}} \neq S_q^\star$, then*

$$\widehat{S}_q = S_q^\star, \qquad U_q(\widehat{S}_q) - U_q(S_q^{\mathrm{qry}}) \geq \Delta_b(q) > 0. \tag{21}$$

Proof. The condition $\Delta_b(q) > 2\alpha\eta_q$ and Theorem 2 imply $\widehat{S}_q = S_q^\star$.

Let

$$A := S_q^\star \setminus S_q^{\mathrm{qry}}, \qquad B := S_q^{\mathrm{qry}} \setminus S_q^\star.$$

The two sets being compared both have cardinality b, so $|A| = |B|$. Since they are not equal, this common cardinality is some $r \geq 1$. Every $a \in A$ lies in $S_q^\star$ and every $d \in B$ lies outside it. Therefore,

$$u_\alpha(a \mid q) \geq \min_{c \in S_q^\star} u_\alpha(c \mid q), \qquad u_\alpha(d \mid q) \leq \max_{c \notin S_q^\star} u_\alpha(c \mid q).$$

Pair the r elements of A and B arbitrarily. Each pair contributes at least $\Delta_b(q)$ to the utility difference, and hence

$$\begin{aligned} U_q(S_q^\star) - U_q(S_q^{\mathrm{qry}}) &= \sum_{a \in A} u_\alpha(a \mid q) - \sum_{d \in B} u_\alpha(d \mid q) \\ &\geq r\Delta_b(q) \\ &\geq \Delta_b(q) > 0. \end{aligned}$$

Substituting $\widehat{S}_q = S_q^\star$ gives Eq. equation 21. □

A.6 Extension to Fused Retrieval Scores

Remark 1 (Extension to fused retrieval scores)**.** *Suppose the system uses* $\mathrm{Fuse}(s_{\mathrm{ent}}, s)$*, where the entity-side score is identical for the ideal and estimated rankings and the fusion map is nondecreasing and κ-Lipschitz in its second argument. Then the per-community fused-score error is at most $\kappa\alpha\eta_q$. Consequently, the fused top-b regret is at most $2b\kappa\alpha\eta_q$, and exact recovery holds whenever the ideal fused-score boundary margin exceeds $2\kappa\alpha\eta_q$.*

Proof. Let

$$\begin{aligned} u_{\mathrm{fuse}}(c \mid q) &:= \mathrm{Fuse}\big(s_{\mathrm{ent}}(c \mid q), u_\alpha(c \mid q)\big), \\ \widehat{s}_{\mathrm{fuse}}(c \mid q) &:= \mathrm{Fuse}\big(s_{\mathrm{ent}}(c \mid q), \widehat{s}_\alpha(c \mid q)\big). \end{aligned}$$

Because the entity-side score is fixed and the fusion map is κ-Lipschitz in its second argument,

$$|\widehat{s}_{\mathrm{fuse}}(c \mid q) - u_{\mathrm{fuse}}(c \mid q)| \leq \kappa |\widehat{s}_\alpha(c \mid q) - u_\alpha(c \mid q)| \leq \kappa\alpha\eta_q.$$

Applying the proof of Theorem 2 with the uniform perturbation radius $\kappa\alpha\eta_q$ gives fused top-b regret at most $2b\kappa\alpha\eta_q$. The same pairwise argument yields exact recovery when the ideal fused-score boundary margin is greater than $2\kappa\alpha\eta_q$. □

B Related Work

Large language models (LLMs) have substantially advanced natural language understanding and generation, enabling strong performance across a wide range of knowledge-intensive tasks (Li et al., 2023). However, LLMs implicitly learn knowledge from parameters (Zeng et al., 2022; Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023) and are prone to hallucination (Jiang et al., 2025a; Fang et al., 2025). Retrieval-Augmented Generation (RAG) alleviates these limitations by enabling LLMs to move beyond their internal parametric knowledge and dynamically access external corpora during inference (Ma et al., 2025; Gao et al., 2023b). The retrieved evidence is incorporated into the generation context, thereby grounding model outputs in accessible and updatable external knowledge (Xiao et al., 2025; Zhang et al., 2025b). RAG (Lewis et al., 2020) combines pre-trained generative models with non-parametric external knowledge and demonstrates that retrieval-grounded generation can produce more factual and less hallucinatory responses than purely parametric generation.

B.1 Retrieval-Augmented Generation

RAG has gradually been widely studied and applied to various tasks, such as open-domain question answering (Jeong et al., 2024; Han et al., 2024), SQL rewrite (Li et al., 2024), and autonomous driving (Kabir et al., 2025). PipeRAG (Jiang et al., 2025b) proposes an algorithm-system co-design approach to reduce generation latency and enhance generation quality. Memorag (Qian et al., 2025) introduces a framework for long-context processing on top of global-memory enhanced retrieval augmentation. In general, conventional RAG follows a pipeline in which it segments external knowledge into independent text chunks and retrieves query-relevant chunks according to vector similarity (Wu et al., 2024; Wang et al., 2025a). Recently, many studies argue that external knowledge should not be treated merely as isolated text fragments, but should instead be organized into a structured knowledge space. RAPTOR (Sarthi et al., 2024b) constructs a tree with differing levels of chunk summarization from the bottom up. GraphRAG (Edge et al., 2024) further suggests that the knowledge units in RAG can be naturally represented by graph-structured knowledge graphs, where nodes denote entities and edges represent relations between entities (Zhang et al., 2024). In this way, the retrieval process can be extended from simple chunk matching to relation-aware exploration (Ni et al., 2025; de Silva et al., 2025; Lim et al., 2025). This structural perspective extends the conventional RAG pipeline into two key stages: knowledge organization and knowledge retrieval.

B.2 Knowledge Organization in Graph-based RAG

Existing work typically uses LLMs to construct a knowledge graph by abstracting entities and relationships from corpora. LinearRAG (Zhuang et al., 2025) investigates an alternative way of constructing knowledge graphs compared with GraphRAG, while a large body of subsequent work focuses more on how to build effective graph-based indexes. LightRAG (Guo et al., 2024) constructs hierarchical indexes (Jia et al., 2015) to improve retrieval performance and efficiency. ArchRAG (Wang et al., 2025b) further employs attributed community detection (Jin et al., 2021; Zhang et al., 2025a) to discover hierarchical subgraphs enriched with attribute information. DA-RAG (Zeng et al., 2026) proposes that the indexing stage should not be a static process, but rather dynamically extract relevant subgraphs. It leverages a community search strategy (Liu et al., 2020) to identify query-relevant communities. HyperRAG (Lien et al., 2026) introduces hypergraph structure to preserve high-order relational integrity. In addition to these structure-oriented methods (Newman & Girvan, 2004; Sozio & Gionis, 2010), some studies have explored index construction from the perspective of cognitive science. HippoRAG (Gutiérrez et al., 2024) synergistically orchestrates LLMs, knowledge graphs, and the Personalized PageRank algorithm to mimic the different roles of the neocortex and hippocampus in human memory. Inspired by the top-down cognitive process of human reasoning, Cog-RAG (Hu et al., 2026) introduces dual hypergraphs to align topics and entities, enabling semantic alignment and coherent generation from global themes to local details. MetaRAG (Qi et al., 2026) reformulates the search decision quality as belief-action alignment. GroupRAG (Duan et al., 2026) simulates the way humans structurally decompose questions and discovers latent structural groups to support fine-grained retrieval. These studies enrich the structural representation of external knowledge, but generally treat graph partitioning as a task-independent preprocessing step. Although QAFD-RAG (Zhou et al., 2026) attempts to adapt graph traversal to each query's holistic semantics, the significant overhead resulting from this cannot be ignored.

C Reproducibility Details

C.1 Datasets

We evaluate EchoRAG on Multihop-RAG (Tang & Yang, 2024), HotpotQA (Yang et al., 2018), and NarrativeQA (Kočiský et al., 2017). Following the evaluation protocol used by ArchRAG (Wang et al., 2025b), all methods use the same raw corpus and 1,000 questions per dataset. Unless otherwise specified, other supplementary experiments use 100 questions randomly sampled from the corresponding subset, with the same questions used across methods or variants within each comparison. The hyperparameter experiments show that 100 questions are sufficient to reflect the overall performance trends. Table 4 and Table 5 list the statistics of the three datasets.

Table 4: Statistics of the full datasets.

Dataset	# Docs	# Questions	Corpus Words
Multihop-RAG	609	2,556	1,073,454
HotpotQA	482,021	90,447	39,432,799
NarrativeQA	1,572	46,765	85,499,660

Table 5: Statistics of the 1,000-question experimental subsets.

Dataset	# Docs	# Questions	Corpus Words	# Nodes	# Edges
Multihop-RAG	609	1,000	1,073,454	13,791	11,408
HotpotQA	9,754	1,000	869,589	20,573	11,824
NarrativeQA	23	1,000	1,197,157	6,075	6,603

(i) Multihop-RAG: Multihop-RAG evaluates retrieval and reasoning across documents in RAG systems. The complete benchmark contains 2,556 queries over 609 English news articles, with evidence for each answerable query distributed across two to four documents. It includes inference, comparison, temporal, and null queries, covering relational reasoning, attribute comparison, event ordering, and unanswerable-question detection. Our subset retains all 609 articles and contains 308 inference, 338 comparison, 223 temporal, and 131 null queries.

(ii) HotpotQA: HotpotQA is a multi-hop QA benchmark containing approximately 113k question–answer pairs constructed from English Wikipedia. It includes bridge questions requiring inference through intermediate entities and comparison questions involving properties of multiple entities. Sentence-level supporting facts are provided to evaluate both answer prediction and evidence identification. Our subset contains 1,000 questions and 9,754 Wikipedia paragraphs, comprising 696 inference and 304 comparison questions.

(iii) NarrativeQA: NarrativeQA is a long-document reading-comprehension benchmark containing 46,765 human-authored question–answer pairs associated with 1,572 books and movie scripts. Its full-story setting requires models to integrate information across complete narratives, while its free-form answers are not restricted to extractive spans. Our subset contains 1,000 questions associated with 23 complete stories or screenplays.

C.2 Baselines Details

In our experiments, we compare our method against widely used methods, including inference-only methods (Zero-shot and CoT), retrieval-only methods (BM25 and Vanilla RAG), and graph-based RAG methods (RAPTOR, GraphRAG, LightRAG, HippoRAG, ArchRAG, DA-RAG, LinearRAG, HyperRAG, and CogRAG). All retrieval-based methods are initialized with the same corpus, and each baseline follows the experimental settings of its original paper and public implementation.

(i) Zero-shot directly prompts the LLM to answer a question using only its parametric knowledge, without retrieving external evidence or providing task-specific demonstrations.

(ii) CoT (Wei et al., 2023) is also retrieval-free, but elicits intermediate reasoning steps before the LLM produces its final answer.

(iii) BM25 (Robertson et al., 1994) is a sparse lexical retriever that ranks corpus passages according to their term-level relevance to the query and supplies the top-ranked passages as generation context.

(iv) Vanilla RAG (Lewis et al., 2020) performs dense semantic retrieval over independently encoded text chunks and provides the top-k retrieved chunks to the LLM for answer generation.

(v) RAPTOR (Sarthi et al., 2024a) recursively embeds, soft-clusters, and summarizes text chunks to construct a bottom-up hierarchical tree, enabling retrieval of relevant information at different levels of abstraction.

(vi) GraphRAG (Edge et al., 2024) extracts entities and relationships to construct a knowledge graph, organizes them into hierarchical Leiden communities, and answers global queries through map-reduce generation over community summaries.

(vii) LightRAG (Guo et al., 2024) constructs a graph-based index from extracted entities and relationships and employs dual-level retrieval to capture both fine-grained entity information and high-level thematic knowledge.

(viii) HippoRAG (Gutiérrez et al., 2024) builds a schemaless knowledge graph from OpenIE triples and applies Personalized PageRank with query concepts as seeds, supporting single-step retrieval of evidence distributed across multiple passages.

(ix) ArchRAG (Wang et al., 2025b) detects attributed communities by considering both graph structure and textual attributes, and organizes communities and entities into a C-HNSW index for efficient multi-level hierarchical retrieval.

(x) DA-RAG (Zeng et al., 2026) formulates retrieval as Embedding-Attributed Community Search and adopts a coarse-to-fine strategy over semantic chunk, knowledge-graph, and similarity layers to dynamically identify query-relevant cohesive subgraphs.

(xi) LinearRAG (Zhuang et al., 2025) constructs an information-preserving graph of entity, sentence, and passage nodes, then combines local semantic bridging for entity activation with Personalized PageRank for passage ranking.

(xii) HyperRAG (Lien et al., 2026) represents complex facts as n-ary hyperedges and retrieves query-conditioned relational chains using either a learned structural-semantic retriever or LLM-guided beam search over the hypergraph.

(xiii) Cog-RAG (Hu et al., 2026) constructs complementary theme and entity hypergraphs and performs two-stage retrieval, first identifying query-relevant global themes and then using them to guide fine-grained entity recall and diffusion.

C.3 METRICS

For HotpotQA and Multihop-RAG, we evaluate answer and retrieval quality using **Accuracy**, **Best-Span F1**, **Recall**, and **Hit** (Hit is undefined for retrieval-free methods). Accuracy measures semantic consistency between a generated response and the reference answer and is determined by an LLM judge to accommodate different valid expressions. Best-Span F1 computes token-level F1 against the best matching contiguous span of a response, reducing the influence of additional explanatory text. Recall measures the proportion of reference-answer content recovered by the response; when multiple references are available, we retain the best score. Hit indicates whether any retrieved context contains a reference answer.

Following the standard evaluation protocol for NarrativeQA, we report **BLEU-1**, **BLEU-4**, **METEOR**, and **ROUGE-L**. BLEU measures unigram and longer n-gram precision, METEOR supports flexible unigram alignment, and ROUGE-L captures longest-common-subsequence overlap.

Following the evaluation protocol of LightRAG (Guo et al., 2024), we further assess four semantic dimensions: **Comprehensiveness**, which measures the coverage of relevant facts and relationships; **Diversity**, which reflects the breadth and non-redundancy of the information presented; **Empowerment**, which evaluates whether an answer helps the reader understand and reason about the question; and **Overall**, which provides a holistic assessment of answer quality across these complementary aspects. These four semantic scores and Accuracy are assessed by an LLM judge, while the remaining metrics are computed automatically.

For the diagnostic and module-level analyses, we additionally report metrics of knowledge organization, retrieval, and evidence quality. **Modularity** measures the structural cohesion of a community partition relative to a degree-preserving null model, while **Semantic Coherence** is the micro-average cosine similarity between node embeddings and their assigned community representations. **Evidence Recall@**k and **Evidence Precision@**k measure the coverage and concentration of gold evidence among the retrieved contexts, while **Supporting-Fact Recall** measures the proportion of annotated supporting facts recovered. We also report **Exact Match**. For evidence- and stage-level analyses, an LLM judge evaluates **Evidence Coherence**, **Answer Sufficiency**, **Generation Readiness**, **Evidence Sufficiency**, and **Retrieval Effectiveness**, capturing how coherent, complete, focused, and relevant the retrieved evidence is for answering the question.

C.4 IMPLEMENTATIONS

We use GPT-4o-mini for knowledge-graph construction, answer generation, and LLM-based evaluation, and `text-embedding-3-small` (1,536 dimensions) for embedding-based indexing and retrieval. GPT-4o is used in supplementary experiments that require a stronger model. Unless required otherwise by an original implementation, the temperature is 0.3 and the maximum output length is 2,048 tokens. Shared model backbones enable a controlled comparison of retrieval and knowledge-organization strategies.

Experiments run on Linux with Python 3.10 and CUDA acceleration where required, while the generation and embedding models are accessed via APIs. We use public baseline implementations with identical corpora and question sets, compute lexical metrics with NLTK and ROUGE, and set the random seed to 42.

D ADDITIONAL EXPERIMENTS

D.1 EFFECTIVENESS RESULTS

Table 6 complements the main effectiveness results by reporting Recall on Multihop-RAG and HotpotQA, as well as BLEU-4 and ROUGE-L on NarrativeQA. EchoRAG achieves the highest Recall on both Multihop-RAG and HotpotQA, while also obtaining the best BLEU-4 and ROUGE-L scores on NarrativeQA. Graph-based methods also generally outperform retrieval-only baselines, especially on HotpotQA, confirming the value of relational knowledge organization for covering information distributed across multiple evidence units. Although EchoRAG's Recall gains over HippoRAG are modest, they are consistent across both benchmarks. More importantly, when considered together with its leading generation-quality results on NarrativeQA and the substantially larger Accuracy and response-quality improvements in the main results, these findings show that EchoRAG preserves strong evidence coverage while making the retrieved information more useful for generation. Its advantage therefore stems not simply from retrieving more relevant content, but from evolving and harmonizing that content into a form that better supports the intended answer.

Table 6: Recall on Multihop-RAG and HotpotQA, and BLEU-4 and ROUGE-L on NarrativeQA. Recall and ROUGE-L are reported as percentages (%), while BLEU-4 is reported in per mille (‰).

Method	**Multihop-RAG**	**HotpotQA**	**NarrativeQA**	
	Recall (%)	Recall (%)	BLEU-4 (‰)	ROUGE-L (%)
Zero-shot	53.2	54.5	4.63	14.44
CoT	47.4	57.1	<u>4.91</u>	15.39
BM25	25.0	48.8	0.04	5.35
Vanilla RAG	26.2	37.0	4.60	15.91
RAPTOR	48.5	55.6	4.56	15.58
GraphRAG	51.3	80.5	2.22	14.80
LightRAG	55.3	83.2	3.31	16.92
HippoRAG	<u>70.0</u>	<u>91.0</u>	1.69	13.08
ArchRAG	60.4	70.2	3.52	<u>18.73</u>
DA-RAG	46.7	76.8	1.55	13.96
LinearRAG	66.3	78.6	1.92	11.25
HyperRAG	65.3	50.8	4.85	14.92
CogRAG	38.5	72.8	0.92	11.70
EchoRAG	**70.5**	**91.4**	**5.44**	**19.16**

D.2 ABLATION RESULTS

To complement the module-level ablation in the main text, we ablate fine-grained components one at a time while keeping the remaining pipeline and evaluation settings unchanged. Within Query-Guided Collaborative Community Evolution, we ablate query-distribution feedback (qdf), LLM semantic feedback (lsf), and the semantic evolution verifier (sev). Within Intent-Harmonized Knowl-

edge Utilization, we ablate entity anchoring (ea), coverage-aware selection (cas), and grounded harmonization (gh). Table 7 reports results on all three datasets.

The results show that every component contributes to response quality: removing any component lowers both Diversity and Overall on all three datasets, while all ablations also reduce Accuracy and F1 on HotpotQA and NarrativeQA. Among the community-evolution components, qdf has the most consistent effect, with its removal reducing Overall across the three datasets, confirming that accumulated query demand is important for deciding where the reusable knowledge organization should evolve. Removing lsf also causes substantial losses, particularly on HotpotQA, whereas removing sev produces smaller but systematic declines in semantic quality, indicating that the verifier mainly improves the reliability of LLM-guided evolution. The utilization-side components play complementary roles: gh is especially important for answer correctness and evidence use, ea preserves precise graph entities when translating inferred intent into evidence, and cas improves the coverage and diversity of the selected knowledge. On Multihop-RAG, several ablated variants obtain higher Accuracy or F1 but consistently lower Diversity and Overall, showing that lexical answer overlap alone does not capture whether dispersed evidence is organized into a complete and useful response. Overall, qdf, lsf, and sev jointly maintain a demand-aware and reliable knowledge organization, whereas ea, cas, and gh transform that organization into grounded, intent-aligned evidence for generation.

Table 7: Fine-grained ablation results on Multihop-RAG , HotpotQA, and NarrativeQA (%).

Method	**Multihop-RAG**				**HotpotQA**				**NarrativeQA**			
	Acc.	F1	Div.	Over.	Acc.	F1	Div.	Over.	Acc.	F1	Div.	Over.
EchoRAG	**73.7**	**69.1**	**67.8**	**58.0**	**85.1**	**91.6**	**78.3**	**69.8**	**51.2**	**52.5**	**48.4**	**36.8**
w/o qdf	69.0	69.7	58.0	51.0	76.0	84.3	72.4	58.6	45.0	48.5	43.8	31.6
w/o lsf	71.0	71.7	60.0	53.6	77.0	82.8	72.6	58.6	48.0	48.6	44.0	33.2
w/o sev	73.0	72.0	63.8	56.6	77.0	84.2	73.0	60.6	49.0	50.9	45.4	33.8
w/o ea	75.0	73.3	59.6	53.4	76.0	82.5	70.8	55.6	49.0	50.5	44.8	35.0
w/o cas	74.0	72.3	61.6	54.4	78.0	82.7	71.6	58.2	45.0	50.6	43.0	33.6
w/o gh	64.0	66.0	62.8	54.6	71.0	81.0	63.8	57.6	45.0	50.8	46.2	32.4

D.3 Visualization of Intent-Harmonized Retrieval

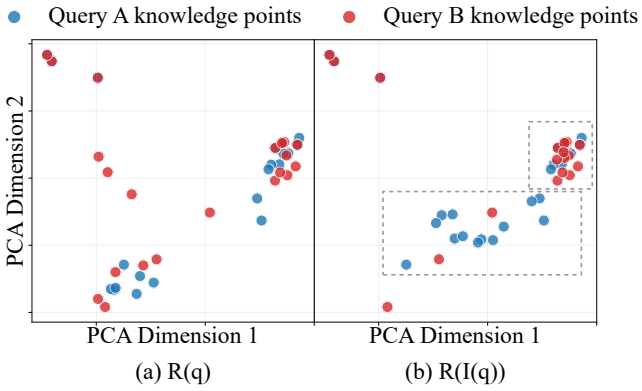


Figure 5: PCA visualization of knowledge points retrieved for two queries using query-only retrieval $R(q)$ and intent-harmonized retrieval $R(I(q))$. The dashed boxes highlight the improved separation between query-specific evidence after intent harmonization.

To provide an intuitive complement to the quantitative motivation validation in Figure 2(b), we select queries with similar surface semantics but different latent information needs and compare the knowledge retrieved by query-only retrieval $R(q)$ with that retrieved after intent harmonization $R(I(q))$. Both settings use the same corpus, retrieval budget, and encoder. We embed the retrieved knowledge units in a shared representation space and jointly project them into two dimensions using PCA, with colors denoting the query for which each knowledge point is retrieved.

As shown in Figure 5(a), query-only retrieval produces substantial overlap between the knowledge points associated with the two queries, indicating that surface-level similarity directs both queries toward largely indistinguishable evidence. After intent harmonization, Figure 5(b) shows clearer query-specific concentrations: the knowledge points for Query A form a distinct lower region, while those for Query B are concentrated primarily in the upper-right region. Limited overlap remains, as expected for queries sharing a common topic, but the stronger separation shows that the inferred intent changes the retrieval focus toward evidence specific to each information need rather than merely expanding the original query. This visualization supports the quantitative gains in Supporting-Fact Recall, Exact Match, and Answer F1 reported in the main text, and provides direct qualitative evidence that intent harmonization reduces cross-intent interference in retrieved knowledge.

D.4 Indexing and Query-Time Efficiency

To complement the token and monetary cost analysis in Section 4, we compare the wall-clock efficiency of all graph-based RAG methods under the common implementation environment described in Appendix C. Table 8 reports two complementary costs in seconds: *Index* is the one-time cost of constructing each method's reusable retrieval index, and *Query* is its average latency per question. All methods use the same corpus and evaluated question set for each dataset.

Table 8: Index construction time and average query latency across three datasets, in seconds.

Method	Multihop-RAG		HotpotQA		NarrativeQA	
	Index (s)	Query (s)	Index (s)	Query (s)	Index (s)	Query (s)
RAPTOR	11 118	3.75	7 086	3.88	8 437	2.70
GraphRAG	516 419	30.13	49 908	8.35	188 992	9.80
LightRAG	32 304	17.97	11 592	7.89	9 106	12.18
HippoRAG	26 909	11.53	29 926	7.99	10 859	7.20
ArchRAG	39 669	5.16	58 490	2.62	23 756	3.40
DA-RAG	3 075	162.61	3 422	82.26	2 644	103.96
LinearRAG	6 997	11.79	5 578	6.74	1 653	4.45
HyperRAG	25 292	364.94	22 861	75.72	23 566	95.91
CogRAG	10 574	31.51	9 558	18.33	10 022	18.63
EchoRAG	33 145	33.12	18 378	13.63	6 238	13.79

The results reveal that no method dominates both offline construction and online serving. EchoRAG has a moderate indexing cost because it constructs graph communities and their reusable representations, but remains consistently less expensive to index than GraphRAG, ArchRAG, and HyperRAG across all three datasets. In particular, it reduces indexing time by approximately 63%–97% relative to GraphRAG, showing that adaptive community organization does not require its very large preprocessing overhead. At query time, EchoRAG is slower than lightweight retrieval pipelines because intent induction, complementary community selection, and grounded evidence harmonization introduce additional processing. Nevertheless, it is substantially faster than the higher-latency DA-RAG and HyperRAG on every dataset, while remaining close to GraphRAG and CogRAG on Multihop-RAG and achieving lower latency than CogRAG on HotpotQA and NarrativeQA. Together with the effectiveness and cost results in the main text, these measurements show that EchoRAG achieves its quality gains with a practical balance between reusable offline organization and per-query latency rather than relying on unrestricted computation.

D.5 Hyperparameter Sensitivity

We conduct hyperparameter sensitivity studies on Multihop-RAG. We examine four main hyperparameters: (i) the structure–semantics balance weight λ, which controls the relative contributions of topological and semantic gains during community evolution; (ii) the query–intent coordination weight α, which balances the original query representation against the inferred intent representation; (iii) the coupled demand-shift thresholds (τ_D, τ_H), which determine whether a community is retained, refreshed, or structurally evolved; and (iv) the query count n, which controls the amount of observed query evidence used to estimate the empirical demand distribution. The last factor measures sensitivity to query-observation scale rather than constituting a model hyperparameter in the

strict sense. We report Accuracy and Recall together with the LLM-based Overall and Comprehensiveness scores in Figure 6.

EchoRAG exhibits robust stability across a wide spectrum of hyperparameter configurations. The non-monotonic dynamics of both the structure-semantic balance parameter and the intent expansion weight reveal that over-indexing on either isolated signal (graph topology or semantic inference) disrupts the system and inevitably degrades holistic performance. Furthermore, the downstream results remain highly stable across a broad, practical range of evolution trigger thresholds, proving that EchoRAG obviates the need for brittle, dataset-specific tuning heuristics. Finally, as the query observation window expands, both accuracy and overall generation quality scale accordingly. This observation demonstrates that accumulating denser, richer distributional feedback inherently drives more reliable and precise knowledge graph evolution.

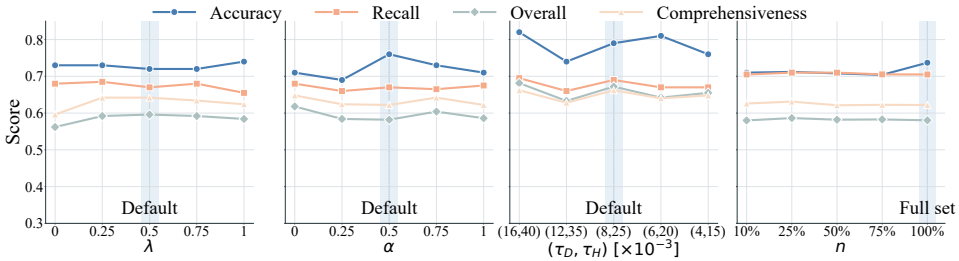


Figure 6: Hyperparameter sensitivity of EchoRAG.

D.6 Robustness to LLM Backbones

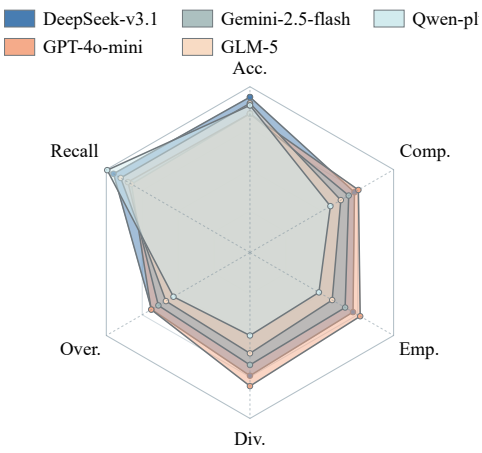


Figure 7: Performance of EchoRAG with different LLM backbones. Each axis reports the corresponding metric on the scale shown in brackets.

To examine whether EchoRAG depends on a particular LLM, we instantiate the framework with five representative backbones: DeepSeek-v3.1, GPT-4o-mini, Gemini-2.5-flash, GLM-5, and Qwen-plus. All variants use the same dataset split, knowledge base, retrieval budget, prompts, hyperparameters, and evaluation protocol, with only the LLM backbone replaced. Figure 7 compares answer Accuracy and Recall together with the four semantic response-quality dimensions.

Figure 7 shows that EchoRAG maintains a broadly consistent performance across all five backbones, indicating that its benefits are not tied to a single proprietary or open model. DeepSeek-v3.1 and Gemini-2.5-flash remain competitive on answer correctness and retrieval coverage, whereas GPT-4o-mini performs particularly strongly on Comprehensiveness, Diversity, Empowerment, and Overall quality. GLM-5 achieves balanced performance across the evaluated dimensions, while Qwen-plus exhibits a more visible decline in the semantic quality metrics. The smaller variation in Recall than in response-quality dimensions suggests that EchoRAG's structured retrieval remains

relatively stable across backbones, whereas the final synthesis of retrieved evidence is more sensitive to the reasoning and generation capabilities of the underlying LLM.

D.7 Case Study

Case 1: Correcting a semantically inconsistent community assignment. Experimental setup. We select a representative node-movement example from the Multihop-RAG evaluation and trace it under the same system configuration used in the main experiments. For the candidate entity MLB, we record its graph neighbors, the dominant themes and representative entities of its source and destination communities, its entity–community similarities, the LLM judgment, the joint-objective gain, and the stability of the resulting assignment. **Analysis.** As shown in Figure 8, MLB is initially assigned to an NFL-oriented community despite its baseball-specific description and graph context. Reassigning it to the MLB-oriented community increases its semantic compatibility from 0.3383 to 0.6219, agrees with the semantic judge, and yields a positive joint gain. The new assignment remains unchanged through convergence, indicating that conflict-triggered semantic feedback can correct a structurally plausible but semantically inconsistent community assignment without causing oscillatory updates.

Candidate entity	**MLB**: Major League Baseball, the governing body of professional baseball in the United States.
Graph context	Houston Astros; PitchCom Devices; Prop Bets.

	Source: Community 34	**Destination: Community 35**
Dominant theme	NFL teams, players, injuries, and the 2023 NFL season	MLB teams, players, competition, contracts, and league events
Representative entities	NFL, Philadelphia Eagles, Carolina Panthers, NFL Draft, NFL Network, MLB	Houston Astros, Shohei Ohtani, Mike Trout, New York Yankees, Los Angeles Dodgers, World Series, MLB
Similarity	0.3383	0.6219
Diagnosis	MLB is a cross-sport semantic outlier.	MLB is consistent with the community topic and its graph neighbors.

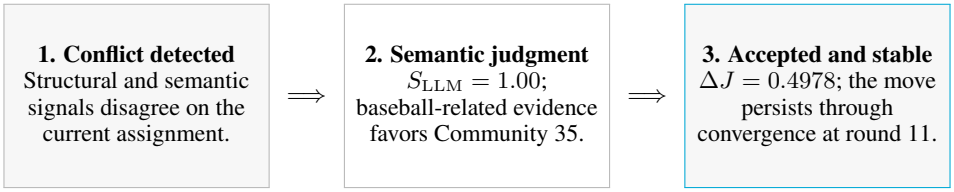


Figure 8: Community-evolution case on Multihop-RAG. The comparison table exposes the semantic mismatch in the source community and the improved compatibility after reassignment; the decision trace summarizes how the conflict-triggered judge produces a stable update.

Case 2: Resolving multi-hop intent and route-related distractors. Experimental setup. We select a HotpotQA question that requires two dependent reasoning steps: identifying the final destination of Mount Lindesay Road and retrieving the population of that destination. Under the same evaluation setting as the main experiments, we present the output of every baseline separately and trace EchoRAG from its retrieved passage through query-constraint resolution and evidence harmonization to the final answer. **Analysis.** As shown in Figure 9, RAPTOR does not recover sufficient evidence; GraphRAG returns Woodenbong and its population; LightRAG returns Rathdowney and its population; HippoRAG identifies Tenterfield but omits the requested population; ArchRAG also selects Rathdowney instead of the final destination; DA-RAG correctly returns Tenterfield and its population; LinearRAG and HyperRAG each identify Tenterfield but provide incomplete answers; and CogRAG lacks sufficient evidence to resolve the full query. EchoRAG returns the correct answer and additionally exposes a grounded trace that jointly resolves the destination and population constraints while removing Rathdowney and Woodenbong as distractors. This case demonstrates that intent harmonization transforms retrieved knowledge into focused evidence that directly supports all constraints expressed in the query.

Question	What is the population of the town that the Mount Lindesay Road eventually leads to?
Ground truth	**6,986**

Method	Extracted answer	Observed failure or success
RAPTOR	Insufficient evidence	Fails to recover enough evidence to resolve the destination and its population.
GraphRAG	Woodenbong: 332	Confuses a nearby route location with the requested endpoint.
LightRAG	Rathdowney: 434	Selects an intermediate town on the route rather than its final destination.
HippoRAG	Tenterfield; population missing	Identifies the destination but does not provide the requested population.
ArchRAG	Rathdowney: 434	Treats a route-related intermediate town as the requested endpoint.
DA-RAG	Tenterfield: 6,986	Recovers the correct endpoint and population.
LinearRAG	Tenterfield; population missing	Recovers the destination entity but loses the population attribute needed for the answer.
HyperRAG	Tenterfield; population missing	Reaches the correct destination but returns an incomplete answer without its population.
CogRAG	Insufficient evidence	Misses at least one of the two constraints required to produce the answer.
EchoRAG	**Tenterfield: 6,986**	Resolves both constraints and explicitly rejects route-related distractors.

Retrieved source	"Mount Lindesay Road leads to Legume and eventually Tenterfield. At the 2015 census, Tenterfield had a population of 6,986."
Query constraints	Final destination = Tenterfield; requested attribute = population.
Harmonized evidence	Community 1859 (score: 0.9553; relevance: high) → Tenterfield is the destination of Mount Lindesay Road and has a population of 6,986.
Distractors removed	Rathdowney and Woodenbong are locations along or near the route, not the final destination requested by the query.
Final answer	**The town is Tenterfield, with a population of 6,986.**

Figure 9: Intent-harmonization case on HotpotQA. The upper comparison reports each competing method separately, while the lower evidence trace shows how EchoRAG resolves both query constraints and removes route-related distractors. Red denotes incorrect or incomplete evidence, and cyan denotes correct evidence.

E Prompt Templates

This section provides the complete prompts for the four LLM calls in EchoRAG: semantic move judgment, intent induction, grounded evidence refinement, and final answer generation. We retain all four templates because each controls a distinct stage whose output affects either knowledge evolution, retrieval, or the reported answer quality.

LLM Prompt: Semantic-Coherence Move Judgment

Role. Act as a domain-adaptive knowledge-graph expert who evaluates the semantic coherence of moving an entity between communities. Assess the proposed move using the communities' ontology, relationships, themes, and historical decision context.

Task. Determine whether moving an entity from its current source community to a proposed target community would improve semantic modularity.

Inputs.

1. `node_description`: a textual description of the entity being considered for relocation;
2. `node_neighbors`: attributes and neighboring entities associated with the entity;
3. `source_summary`: a summary of the entity's current source community;
4. `target_summary`: a summary of the proposed target community;
5. `target_theme`: the distribution of member types or core semantic theme of the proposed target community;
6. `memory_context`: historical information about previous gating decisions involving the entity and target community.

Prompt continued

Evaluation Criteria.

1. **Thematic Homogeneity.** Determine whether the entity shares the core topic or intent of the target community more strongly than that of its current source community.
2. **Structural Logic.** Determine whether the entity's neighbors and attributes are functionally or semantically compatible with members of the target community.
3. **Contextual Gain.** Determine whether the move would reduce semantic noise in the source community while adding meaningful information to the target community.

Scoring Rules. Assign a continuous `consistency_score` in $[-1, 1]$: `1.0` denotes a perfect and central fit to the target; `0.0` denotes a neutral move because the entity fits both or neither community; and `-1.0` denotes a conflict with the target theme. Positive values favor the target, whereas negative values favor the source.

Output. Return one JSON object containing `reasoning`, a concise analysis using the three criteria and limited to 50 words, and `consistency_score`, a floating-point value in $[-1, 1]$. Return JSON only, with no Markdown, conversational commentary, or additional fields.

LLM Prompt: Grounded Intent Induction

Role. Act as a knowledge-graph query expert who infers the latent information need behind a user query. Produce retrieval guidance, not factual evidence, and do not invent entities or facts that are not stated or safely implied by the query.

Task. Transform the query into a structured intent blueprint $\widehat{I}(q)$ and a hypothetical answer form h_q for retrieving relevant knowledge-graph communities.

Input. `query`: the user's original natural-language question.

Induction Rules.

1. Identify all explicit constraints, including named entities, actions, relationships, dates, numbers, locations, comparisons, and source restrictions.
2. Infer the expected answer type, target entities or entity types, required facets, and salient relations needed to answer the query.
3. Write a concise retrieval-intent description that preserves query-specific entities and keywords.
4. Express the hypothetical answer only as a semantic template with placeholders for unknown values; do not treat it as evidence or assert unsupported facts.

Output. Return one JSON object with an `intent_blueprint` containing `expected_answer_type`, `target_entities`, `required_facets`, `salient_relations`, `constraints`, and a 40–80 word `retrieval_intent`; also return `hypothetical_answer`, a concise answer template using placeholders for unknown facts. Return JSON only, with no Markdown, commentary, or additional fields.

LLM Prompt: Single-Community Evidence Refinement

Role. Act as a precision information extractor who distinguishes answer-relevant evidence from broad topical similarity within a retrieved knowledge-graph community.

Task. Assess the community's relevance to the original query and extract atomic, query-relevant facts grounded in its raw source chunks.

Inputs. The prompt receives:

1. the original `query` and its structured `intent_blueprint`;
2. retrieval metadata, including `community_id`, `retrieval_rank`, and `similarity_score`;
3. structured community information, including the `community_summary`, query-ranked `key_entities`, and `key_relationships`;
4. the original `raw_text_chunks` and their exact `chunk_id` values.

Evidence Extraction Rules.

1. Scan every raw text chunk rather than stopping at the first relevant passage. Prioritize raw chunks over compressed entity descriptions and community summaries.
2. Identify explicit query entities, direct relationships, and answer-critical information such as actions, comparisons, dates, numbers, and locations. Broad topical overlap without such information constitutes low relevance.
3. For multi-hop or comparative queries, retain complementary facts supporting different required hops, even when an individual chunk cannot answer the complete query. Connect facts across chunks only when the connection is explicitly supported.
4. Extract atomic subject–action–object facts in the form `Entity: query-relevant fact or relationship`. Do not return isolated entity names, infer unstated relationships, fabricate missing hops, or introduce external knowledge.

Prompt continued

5. Ground every evidence card in one or more exact input `chunk_id` values and copy the minimal supporting `source_span` verbatim from those chunks. Information found only in the community summary may inform relevance assessment but must not be presented as grounded evidence.

Relevance Levels. Assign `high` when the community contains query entities and specific answer-critical facts; `medium` when it provides a concrete but incomplete fact or reasoning hop; `low` when it has only broad topical overlap; and `none` when it is unrelated.

Output. Return one JSON object with:

- `community_id`, `relevance_level`, a continuous `relevance_score` in $[0, 1]$, and a one-sentence `justification`;
- a concise `query_focused_summary`;
- `key_entities`, expressed as entity–fact or entity–relationship statements;
- `evidence_cards`, each containing an atomic `fact`, exact `source_span`, `source_chunk_ids`, an `intent_relevance` score in $[0, 1]$, and a `high` or `medium` confidence label.

For `low` or `none` relevance, return an empty `evidence_cards` array. Return JSON only. Do not include Markdown, conversational commentary, unsupported facts, or identifiers absent from the input.

LLM Prompt: Evidence-Grounded Final Answer Generation

Role. Act as an evidence-grounded answering assistant. Answer the query using only the supplied verified evidence cards and their supporting source passages.

Task. Generate a direct, comprehensive, and evidence-grounded answer by synthesizing the query-relevant facts extracted from the retrieved knowledge-graph communities.

Inputs. The prompt receives the original `query`, the structured `intent_blueprint`, relevance-ordered `verified_evidence_cards`, and their `source_passages`. Community identifiers and retrieval metadata are supplied for provenance checking but are not answer content.

Answer Generation Rules.

1. Begin with a direct and concise answer, such as a name, value, "Yes," "No," or "I don't know."
2. After the direct answer, provide only the explanation needed to connect the supporting facts. Synthesize complementary evidence instead of listing retrieved items.
3. Include relevant names, dates, numbers, actions, relationships, conditions, consequences, and differing perspectives only when explicitly supported by the supplied evidence.
4. Use only the supplied evidence cards and source passages. Connect facts across sources only when that connection is explicitly supported; never use the hypothetical answer as factual evidence.
5. If the evidence is incomplete, answer the supported portion and state the remaining limitation. If no verified evidence supports an answer, respond that the retrieved evidence is insufficient.
6. Unless the query explicitly requests provenance, do not expose internal community IDs, retrieval scores, or processing labels in the answer.

Output. Return a concise natural-language answer whose first sentence directly addresses the query. Begin with an insufficiency statement only when no verified evidence supports an answer, and do not use filler expressions such as "Based on the provided contexts."