Hybrid Entity Retrieval for Hierarchical Knowledge-Graph-Based Retrieval-Augmented Generation via Reciprocal Rank Fusion

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) systems based on hierarchical knowledge graphs, such as HiRAG, currently rely exclusively on vector cosine similarity to retrieve relevant entities prior to context assembly. This single-signal approach cannot simultaneously capture semantic, lexical, and structural relevance, leaving useful entities poorly ranked. We propose replacing vector-only retrieval with a three-signal hybrid module that fuses dense vector similarity, sparse BM25 keyword matching, and graph-structural PageRank scores via Reciprocal Rank Fusion (RRF), requiring no additional large language model calls, no training data, and no score normalization. Evaluated on the UltraDomain Mix benchmark (130 queries, 625,948-token corpus) using GPT-4o as a large-language-model-as-judge across 260 pairwise comparisons per experiment, hybrid RRF at k=10 achieves an overall win rate of 0.585 (p=0.0075, Cohen's h=0.170), a +17.0% improvement over the vector-only baseline. Quality peaks at k=10 rather than k=20, indicating a noise-inducing saturation effect beyond ten entities. Ablation confirms super-additive synergy: BM25+PageRank combined yields +10.0%, exceeding the sum of individual contributions (+3.1% + 1.5% = 4.6%). Maximal Marginal Relevance diversity reranking consistently degrades quality, establishing that entity redundancy is not the bottleneck in this architecture.

1 Introduction

Large language models (LLMs) suffer from hallucinations, fixed knowledge cutoffs, and inability to incorporate domain-specific information without fine-tuning. Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) addresses these limitations by grounding model inference in dynamically retrieved external context. Knowledge-graph-based RAG approaches extract structured entity graphs from source documents and use graph structure for both retrieval and context assembly. GraphRAG (Edge et al., 2025) pioneered community-level summarization over knowledge graphs; HiRAG (Huang et al., 2025) extended this with a four-layer hierarchical context—global community reports, bridge paths, local entity descriptions, and evidence chunks—achieving strong performance on complex multidocument benchmarks.

Despite HiRAG's architectural sophistication, its entity retrieval stage relies solely on cosine similarity between query and pretrained dense entity embeddings (Neelakantan et al., 2022). This single-signal approach cannot simultaneously capture semantic paraphrase relationships, lexical precision for exact keyword matches, and graph-structural importance. Entities that are lexical matches or structural hubs may score poorly on vector similarity alone. Lexical signals (BM25 (Robertson and Zaragoza, 2009)) and structural signals (PageRank (Page et al., 1999)) capture orthogonal relevance dimensions that vector similarity systematically ignores.

This paper replaces the single-signal retrieval step with a lightweight, parameter-free hybrid module based on Reciprocal Rank Fusion (RRF) (Cormack et al., 2009). Specific contributions include the following:

1. a hybrid retrieval module fusing dense vector, sparse BM25, and precomputed PageRank via RRF without additional LLM calls or training data;

2. controlled experiments on 130 queries varying entity budget k to characterize the quality–quantity tradeoff;

3. a four-way ablation demonstrating super-additive signal synergy; and

4. a negative result with Maximal Marginal Relevance (MMR) diversity reranking establishing

that redundancy is not the primary retrieval bottleneck.

2 Related Work

The fundamental motivation for hybrid retrieval is the empirically documented complementarity of dense and sparse signals. Dense retrieval excels at semantic paraphrase matching but underperforms on exact keyword tasks; BM25 (Robertson and Zaragoza, 2009) provides precise lexical matching at the cost of vocabulary-mismatch sensitivity. Lin (2022) observed that dense and sparse representations offer complementary relevance signals and that combining them yields further effectiveness gains. Hsu and Tzeng (2025) further characterize the divide: sparse retrieval depends on term overlap while dense retrieval depends on embedding alignment—orthogonal properties exercised differently across query types. Entity-centric retrieval particularly benefits from lexical signal, as exact entity identifiers in knowledge graphs are reliably matched by BM25 even when embedding spaces fail to capture rare terms (Cheekuri, 2025). The standard retrieve-then-merge two-channel architecture fuses BM25 and dense candidates into a single pool for ranking (Arabzadeh et al., 2021).

Graph-structural signals provide a third orthogonal relevance dimension suited to knowledge-graph settings. PageRank and Personalized PageRank (PPR) capture structural centrality and enable multihop reasoning unavailable to single-hop vector search (Raja and Vats, 2025; Vujić et al., 2025). Graph-augmented retrieval significantly outperforms dense-only retrievers on complex question answering requiring multidocument reasoning (Agrawal et al., 2025). PageRank has a demonstrated track record in entity disambiguation: applied over entity co-occurrence graphs it achieves 87% accuracy on named entity mention resolution (Alhelbawy and Gaizauskas, 2014). However, unconstrained graph traversal can introduce irrelevant content through over-expansion, and spurious edges in automatically constructed graphs risk concept drift (Dietz et al., 2018). We use static PageRank computed at index time rather than query-time traversal, avoiding these pitfalls while cleanly combining with lexical and semantic signals through rank fusion.

RRF (Cormack et al., 2009) is a parameter-free method that combines ranked lists without score normalization, consistently matching or exceeding alternative fusion methods including Condorcet Fuse and CombMNZ. Its key advantage is operating on ranks rather than scores, making it suitable for fusing signals with fundamentally different score distributions (cosine similarities, BM25 scores, PageRank probabilities). Mourão et al. (2014) propose Inverse Square Rank (ISR) fusion as an alternative with quadratic decay; we adopt RRF for its zero-parameter calibration requirement. Weighted and query-adaptive fusion variants have also been explored (Mala et al., 2025; Hsu and Tzeng, 2025). Fusion gains can be neutralized after reranking and truncation (Medrano et al., 2026), and NDCG is sensitive to RRF parameter choices (Bruch et al., 2023), underscoring the importance of end-to-end empirical evaluation.

Knowledge-graph-based RAG (KG-RAG) architectures vary in how they retrieve and assemble graph evidence. HybridRAG (Sarmah et al., 2024) concatenates VectorRAG and GraphRAG contexts, outperforming either alone on financial question answering (QA), while noting that KG-first retrieval fails when no entity is explicitly mentioned—motivating dense retrieval as a fallback. KG^{2}RAG (Zhu et al., 2025) combines semantic retrieval with graph-guided expansion, reporting 8%+ improvements, but finds that graph expansion without careful context organization degrades quality—consistent with HiRAG's hierarchical four-layer assembly design. End-to-end benefits of hybrid graph-plus-dense retrieval have been demonstrated across multiple architectures (Sun, 2025).

Despite extensive work on hybrid retrieval in information retrieval (IR), three-signal fusion combining dense vector similarity, BM25, and PageRank has not been applied to entity retrieval in hierarchical KG-RAG. Prior hybrid work targets passage-level retrieval in flat-chunk RAG or open-domain QA; the entity-level retrieval stage unique to HiRAG's architecture, where entity selection cascades through all four context layers, has not been studied. This paper addresses that gap.

3 Materials and Methods

3.1 HiRAG Architecture Overview

HiRAG (Huang et al., 2025) constructs a knowledge graph from a document corpus using LLM-based extraction: documents are chunked, an LLM identifies named entities and relational triplets, and mentions are canonicalized into graph nodes with directed edges annotated by description and

confidence weight. Community structure is identified using the Leiden algorithm (Traag et al., 2019), and an LLM generates natural-language community reports for each cluster. At query time, HiRAG assembles a four-layer hierarchical context: (1) global community reports for retrieved entities' communities, (2) bridge shortest paths between retrieved entities, (3) local entity descriptions, and (4) evidence source chunks. The entity retrieval step determines which entities seed all four layers. In the original HiRAG, retrieval uses cosine similarity between query and entity embeddings produced by a pretrained OpenAI `text-embedding-3-small` encoder (Neelakantan et al., 2022), selecting the top-k entities. All components except entity retrieval are held constant across experiments.

3.2 Hybrid Entity Retrieval via Reciprocal Rank Fusion

We replace the single-signal retrieval step with a three-signal hybrid module. Each signal independently ranks all knowledge graph entities with respect to the query; the three ranked lists are fused via RRF (Cormack et al., 2009) to produce a unified ranking, and the top-k entities are passed to HiRAG's context assembly pipeline unchanged.

Dense vector similarity. Each entity description is encoded at index time using a pretrained transformer-based text encoder (Neelakantan et al., 2022); at query time cosine similarity is computed between the query embedding and all entity embeddings (Karpukhin et al., 2020). This signal captures semantic paraphrase relationships and handles natural-language variation.

BM25 keyword retrieval. A BM25 index (Robertson and Zaragoza, 2009) over entity name-and-description text is queried with the tokenized query string. BM25 assigns higher scores to entities containing query terms at higher frequency with document-length normalization, providing lexical precision that dense embeddings lack for exact keywords, proper nouns, and domain-specific terminology.

PageRank structural importance. The precomputed PageRank score (Page et al., 1999) of each entity node, computed once at index time with damping factor $\alpha = 0.85$, measures global structural centrality. Hub entities connecting multiple communities receive high scores and are likely relevant to broad synthesis queries, capturing a relevance dimension orthogonal to both semantic and lexical matching.

The three ranked lists are fused via Reciprocal Rank Fusion. For each entity e:

$$\text{score}_{\text{RRF}}(e) = \sum_{i \in \{\text{vec, bm25, pr}\}} \frac{1}{k_{\text{RRF}} + \text{rank}_i(e)} \tag{1}$$

where $\text{rank}_i(e)$ is the 1-indexed rank of entity e in list i, and $k_{\text{RRF}} = 60$ following Cormack et al. (2009). Entities absent from a list receive a penalty rank of (total entities + 1). RRF requires no score normalization across heterogeneous signals, is robust to outlier scores, and adds negligible latency compared to LLM inference costs.

3.3 MMR Reranking as an Evaluated Alternative

We evaluate MMR reranking (Carbonell and Goldstein, 1998) applied after RRF retrieval. MMR iteratively selects entities balancing query relevance against similarity to already-selected entities:

$$\text{MMR}(e) = \lambda \cdot \text{sim}(e, q) - (1 - \lambda) \cdot \max_{e' \in S} \text{sim}(e, e') \tag{2}$$

where S is the selected set and λ controls the relevance–diversity tradeoff. We test $\lambda = 0.7$ and $\lambda = 0.9$, applying MMR to the top-30 RRF candidates to select the final $k = 10$ entities. The hypothesis is that RRF may retrieve redundant entities whose removal via MMR could improve context diversity and response quality.

3.4 Dataset and Evaluation Protocol

All experiments use the UltraDomain Mix benchmark (Huang et al., 2025): five documents spanning agriculture, computer science, law, medicine, and finance (625,948 tokens total). The knowledge graph contains 668 entity nodes and 587 relationship edges. The evaluation set comprises 130 queries spanning factoid, comparative, multihop, and synthesis question types.

We use a pairwise LLM-as-judge protocol (Zheng et al., 2023) with GPT-4o as judge. For each query, both system responses are presented in randomized order and evaluated on four criteria: *Comprehensiveness* (all aspects addressed), *Empowerment* (actionable, useful information), *Diversity* (multiple perspectives), and *Overall Winner*. Each query is evaluated twice with swapped response order to control for position bias, yielding 260 judge records per experiment. Win rate is the

proportion of comparisons in which the treatment system is judged superior.

Statistical significance is assessed using a one-sided binomial test against the null of equal win rates ($p = 0.5$) (Zheng et al., 2023). Effect size is quantified via Cohen's h on arcsine-transformed proportions. Wilson score 95% confidence intervals (half-width ± 6.1 percentage points at $n = 260$) are reported. The baseline is vector-only cosine similarity at $k = 20$, the default HiRAG configuration (Huang et al., 2025).

4 Experimental Results

4.1 Hybrid Retrieval vs. Vector-Only Baseline (k=20)

Table 1 shows per-criterion win rates for hybrid RRF vs. vector-only baseline, both at $k = 20$. Hybrid retrieval outperforms across all four criteria (+7.0% to +9.2%), demonstrating consistent quality gains from three-signal RRF fusion. However, hybrid k=20 produces 2.04× more context tokens (21,197 vs. 10,419), potentially confounding retrieval quality with context quantity; Section 4.2 addresses this with token-normalized comparisons.

4.2 Token-Normalized Evaluation

Table 2 compares hybrid RRF at $k = 7$, $k = 10$, and $k = 20$, showing mean context tokens, token ratio versus baseline, and overall win rates. The results reveal a nonmonotonic relationship: quality peaks at $k = 10$ (+17.0%), outperforming both $k = 7$ (+8.4%) and $k = 20$ (+9.2%). Entities ranked 11–20 appear less relevant on average, diluting context quality.

The k=7 result is practically significant: at only 1.14× token ratio, hybrid retrieval still achieves +8.4% improvement, confirming that quality gains reflect improved entity ranking rather than simply larger context. Table 3 gives the per-criterion breakdown at the optimal $k = 10$.

Empowerment shows the largest gain (+22.4%), indicating hybrid retrieval is particularly effective at surfacing specific, authoritative entities needed for actionable responses. Comprehensiveness also improves substantially (+13.0%), while Diversity shows the smallest gain (+10.8%), consistent with the finding in Section 4.4 that redundancy is not the primary bottleneck.

4.3 Ablation Study: Individual Signal Contributions (k=10)

Table 4 presents a four-way ablation at $k = 10$: vector-only baseline, vector+BM25, vector+PageRank, and full three-signal RRF. BM25 adds +3.1%; PageRank adds +1.5%; under an additive model one would predict +4.6% combined. The full fusion achieves +10.0%—more than double the additive prediction, confirming super-additive synergy.

BM25 is most effective for keyword-rich queries; PageRank boosts structurally central entities that vector similarity may underrank; vector similarity provides the semantic backbone neither BM25 nor PageRank can supply. RRF fusion—aggregating rank positions rather than raw scores—enables complementary signals to reinforce each other rather than cancel, producing the observed super-additive effect.

4.4 Negative Result: MMR Diversity Reranking

Table 5 shows that both MMR settings produce negative results, with win rates below 0.5 (the vector-only baseline level). Stronger diversity ($\lambda = 0.7$) causes more degradation (−4.6%) than mild diversity ($\lambda = 0.9$, −2.3%), consistent with the interpretation that MMR demotes relevant entities in favor of less relevant diverse ones when redundancy is not the bottleneck.

4.5 Statistical Significance

Table 6 summarizes statistical properties of key comparisons. Only hybrid RRF at k=10 reaches conventional significance ($\alpha = 0.05$), with $p = 0.0075$ and Cohen's $h = +0.170$ (small-to-medium effect). The 95% confidence interval $[0.524, 0.643]$ lies entirely above 0.5. Other configurations show positive trends without reaching significance at $n = 260$, reflecting power limitations of LLM-as-judge evaluation. No experiment shows significant degradation.

5 Discussion

5.1 Complementary Signals and Super-Additive Synergy

BM25 captures exact keyword matches that embedding spaces miss; PageRank provides structural priors identifying knowledge-graph hubs relevant to synthesis queries; RRF fuses all three signals without score normalization or calibration. The

Table 1: Win Rates: Hybrid RRF (k=20) vs. Vector-Only Baseline (k=20)

Criterion	**Vector-only** (k=20)	**Hybrid RRF** (k=20)	**Δ**
Comprehensiveness	0.465	0.535	+7.0%
Empowerment	0.454	0.546	+9.2%
Diversity	0.462	0.538	+7.7%
Overall Winner	0.454	0.546	+9.2%

Table 2: Overall Win Rate by Hybrid Entity Budget k vs. Vector-Only Baseline (k=20)

Hybrid k	**Mean tokens**	**Token ratio**	**Overall win rate**	**Δ**
$k = 7$	11,881	1.14×	0.542	+8.4%
$k = 10$	15,312	1.47×	0.585	+17.0%
$k = 20$	21,197	2.04×	0.546	+9.2%

Table 3: Per-Criterion Win Rates: Hybrid RRF (k=10) vs. Vector-Only Baseline (k=20)

Criterion	**Vector-only** (k=20)	**Hybrid RRF** (k=10)	**Δ**
Comprehensiveness	0.435	0.565	+13.0%
Empowerment	0.388	0.612	+22.4%
Diversity	0.446	0.554	+10.8%
Overall Winner	0.415	0.585	+17.0%

Table 4: Ablation Study: Signal Contributions at k=10 (Overall Win Rate vs. Vector-Only)

Configuration	**Compr.**	**Empow.**	**Diver.**	**Overall**	**Δ**
Vector only (baseline)	—	—	—	0.500	—
+ BM25	0.523	0.523	0.500	0.515	+3.1%
+ PageRank	0.492	0.504	0.477	0.508	+1.5%
+ BM25 + PageRank (full RRF)	0.573	0.538	0.562	0.550	+10.0%

Table 5: MMR Reranking Results at k=10 (vs. Vector-Only Baseline)

λ (MMR)	**Overall**	**Δ vs. baseline**	**Verdict**
0.7	0.477	−4.6%	Negative
0.9	0.488	−2.3%	Negative

ablation confirms that the +10.0% combined improvement far exceeds the +4.6% additive prediction, providing strong evidence that the signals are complementary on different, nonoverlapping subsets of the query distribution. The RRF smoothing constant $k_{\text{RRF}} = 60$ prevents overweighting rank-1 entities from any single signal, allowing the fusion to reliably surface entities that rank first on one signal but lower on others.

5.2 Quality Over Quantity: The Optimal Entity Budget

Quality peaks at k=10 rather than k=20: entities ranked 11–20 introduce more noise than signal into the assembled context. This is consistent with the "lost in the middle" effect in long-context LLMs (Liu et al., 2024), where information appended to an already-substantial context window receives less effective attention. Community reports for lower-ranked entities may also introduce topically unrelated information diluting primary signal. The practical implication is that practitioners should not simply maximize entity budgets; rather, a moderate budget calibrated on a held-out validation set is preferable, with $k \approx 10$ as a reasonable starting point for benchmarks of similar scale.

5.3 Why MMR Reranking Fails

Entity redundancy is not a significant problem in HiRAG's retrieval pipeline for several reasons: entity canonicalization during graph construction merges duplicate mentions, Leiden community detection (Traag et al., 2019) naturally introduces topical diversity at the community level, and with 668 nodes and k=10, the system selects only 1.5% of all entities, making near-identical top-10 retrievals unlikely. MMR addresses a failure mode absent

Table 6: Statistical Significance of Key Experimental Comparisons (n=260)

Experiment	Win rate	95% CI	p-value	Cohen's h
Hybrid RRF k=10 vs. Vector-only	0.585	[0.524, 0.643]	0.0075	+0.170
Hybrid RRF k=20 vs. Vector-only	0.546	[0.485, 0.606]	0.1536	+0.092
BM25+PageRank vs. Vector-only	0.550	[0.489, 0.609]	0.1209	+0.100
Hybrid RRF k=7 vs. Vector-only	0.542	[0.482, 0.602]	0.1927	+0.085

here, and its diversity term primarily demotes relevant entities in favor of less relevant diverse ones—directly degrading context quality.

5.4 Relation to Prior Work

To our knowledge, this paper is the first to apply three-signal RRF fusion (dense+BM25+PageRank) to entity retrieval in a hierarchical KG-RAG architecture. Prior hybrid retrieval work (Luan et al., 2021; Lin et al., 2021) targets passage-level retrieval in flat-chunk RAG; our results confirm these findings generalize to entity-level graph retrieval and extend them with a graph-structural third signal. GraphRAG (Edge et al., 2025) operates at the community report level via vector similarity; HiRAG's entity-level retrieval stage, where selection cascades through all four context layers (Huang et al., 2025), is distinct and has not previously been targeted with multisignal fusion. The LLM-as-judge evaluation (Zheng et al., 2023) with order-swapping and 260 comparisons provides a robust basis for conclusions despite known biases toward verbosity.

6 Conclusion

This paper proposed and evaluated a hybrid entity retrieval module for HiRAG that fuses dense vector similarity, sparse BM25 keyword retrieval, and graph-structural PageRank via Reciprocal Rank Fusion. At the optimal budget of k=10, hybrid RRF achieves an overall win rate of 0.585 against the vector-only baseline ($p = 0.0075$, Cohen's $h = +0.170$, +17.0% improvement), with the largest per-criterion gain on Empowerment (+22.4%). Quality peaks at k=10 rather than k=20—a saturation effect where additional entities introduce noise—and even at a near-equal token budget (k=7, 1.14× tokens) the improvement reaches +8.4%, confirming that gains reflect retrieval quality rather than context size. Ablation reveals super-additive synergy: BM25+PageRank combined yields +10.0%, more than double the +4.6% additive prediction, establishing genuine signal complementarity. MMR diversity reranking consistently degrades quality, showing that entity redundancy is not the retrieval bottleneck. The module requires no additional LLM calls, no training data, and no score normalization. Future work should evaluate on larger and more diverse knowledge graphs and investigate query-adaptive signal weighting strategies.

Limitations

All experiments use a single benchmark (UltraDomain Mix, five domains); generalizability of the k=10 optimum and improvement magnitudes to other corpora, languages, and domain distributions requires further study. The GPT-4o judge may introduce verbosity bias and self-enhancement bias not fully mitigated by order-swapping. The knowledge graph is relatively small (668 nodes, 587 edges), potentially limiting PageRank's ability to differentiate entity importance; larger graphs may show different signal balance. The RRF smoothing constant $k_{\mathrm{RRF}} = 60$ and equal fusion weights are fixed; learned or query-adaptive weighting could potentially improve upon uniform RRF.

Ethical Considerations

This work proposes a retrieval improvement for an existing hierarchical knowledge-graph-based RAG architecture and does not introduce new training data, models, or human-subject procedures. We use only publicly available scientific and professional-domain corpora (UltraDomain Mix) and commercial language-model APIs (GPT-4o) accessed under their published terms of service. We see no direct risk of producing harmful content beyond risks inherent to the underlying LLMs themselves, which are well documented elsewhere. The proposed hybrid retrieval module adds negligible computational cost over the vector-only baseline (no additional LLM calls, no training), so it does not materially increase the environmental footprint of running HiRAG. Because the technique can in principle be applied to any knowledge

graph—including ones built from biased or low-quality source documents—downstream applications should evaluate the retrieved entity set for representational and factual bias before user-facing deployment. No human annotators or crowdworkers were used; all evaluation is performed by an automated LLM-as-judge protocol (Zheng et al., 2023).

A Reproducibility Details

This appendix records implementation details, package versions, computational budget, and dataset/license information sufficient to reproduce the experiments reported in the main text.

A.1 Software Artifacts and Licenses

All software and data artifacts used in this work are publicly available and were used in accordance with their published terms.

- **HiRAG** (Huang et al., 2025): open-source implementation released under the MIT License; used as the base architecture, modified only to replace the entity-retrieval module.
- **UltraDomain Mix benchmark** (Huang et al., 2025): five public scientific and professional-domain corpora (agriculture, computer science, law, medicine, finance) distributed for research use. The corpora consist of expository, non-conversational text and contain no personally identifying information about private individuals.
- **OpenAI GPT-4o**: used both as the answer-generating language model in HiRAG and as the LLM-as-judge for pairwise evaluation, accessed via the OpenAI commercial API under the OpenAI Usage Policies.
- **OpenAI `text-embedding-3-small`**: 1536-dimensional embeddings for entity descriptions, accessed via the OpenAI API.
- **`rank_bm25`** v0.2.2 (Apache 2.0): BM25 implementation; default Okapi parameters ($k_1 = 1.5, b = 0.75$).
- **NetworkX** v3.3 (BSD 3-Clause): PageRank implementation with damping factor $\alpha = 0.85$, 10^{-6} tolerance.
- **`tiktoken`** v0.7.0 (MIT): token counting for context-budget normalization.

All artifacts are used for their intended research purpose. Derived artifacts (retrieval indices, intermediate ranked lists, evaluation logs) are produced from research-licensed inputs and are themselves intended only for academic reproduction.

A.2 Data Statement

The UltraDomain Mix benchmark used for evaluation comprises 625,948 tokens of English-language text drawn from five public expository corpora spanning agriculture, computer science, law, medicine, and finance. The constructed knowledge graph contains 668 entity nodes and 587 relationship edges, extracted automatically via LLM-based entity and relation recognition over the corpus. The 130 evaluation queries cover factoid, comparative, multihop, and synthesis question types and were released with the original HiRAG benchmark; they do not target or reference identifiable individuals. No additional data were collected or annotated for this paper.

A.3 Computational Budget

The proposed retrieval module itself is lightweight: BM25 indexing and PageRank precomputation over 668 nodes complete in seconds on a single CPU core, and per-query retrieval (vector similarity, BM25 scoring, RRF fusion, optional MMR) runs in well under one second. The dominant cost is OpenAI API usage. Across the nine pairwise comparisons reported in the main text—hybrid RRF at $k \in \{7, 10, 20\}$ versus baseline, the BM25-only and PageRank-only ablations, the full RRF ablation, and the two MMR settings ($\lambda = 0.7, 0.9$)—each experiment uses a fixed set of 130 queries evaluated twice in randomized order, producing $9 \times 260 = 2{,}340$ GPT-4o judge calls. Answer generation contributes roughly an additional 1,500–2,000 GPT-4o completion calls (per-configuration answers for the 130 queries, with the vector-only k=20 baseline reused across multiple comparisons). End-to-end reproduction of the retrieval ablations therefore consumes on the order of 4,000 GPT-4o API calls and 10–15 hours of wall-clock time on commodity hardware, with no local GPU required for the retrieval module itself. One-time knowledge graph construction (LLM-based entity and relation extraction over the 625,948-token corpus) is performed once by the public HiRAG release and reused unchanged across all reported experiments.

A.4 Hyperparameters

All hyperparameters are reported in the main text alongside the experiments that use them, and we restate them here for convenience: $k_{\mathrm{RRF}} = 60$ (Section 3.2), PageRank damping $\alpha = 0.85$ (Section 3.2), MMR diversity coefficient $\lambda \in \{0.7, 0.9\}$ (Section 3.3), entity budgets $k \in \{7, 10, 20\}$ (Section 4.2). No hyperparameter search was performed beyond the reported sweep over k and λ; the RRF smoothing constant follows Cormack et al. (2009) without tuning.

References

V. Agrawal, F. Wang, and R. Puri. 2025. Query-aware graph neural networks for enhanced retrieval-augmented generation. *Preprint*, arXiv:2508.05647.

A. Alhelbawy and R. Gaizauskas. 2014. Graph ranking for collective named entity disambiguation. In *Proc. 52nd Annu. Meeting Assoc. Comput. Linguist. (ACL)*, pages 75–80, Baltimore, MD, USA.

N. Arabzadeh, X. Yan, and C. L. A. Clarke. 2021. Predicting efficiency/effectiveness trade-offs for dense vs. sparse retrieval strategy selection. In *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, pages 2862–2866.

S. Bruch, S. Gai, and A. Ingber. 2023. An analysis of fusion functions for hybrid retrieval. *ACM Trans. Inf. Syst.*, 42(1). Art. 20.

J. Carbonell and J. Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, pages 335–336, Melbourne, Australia.

K. Cheekuri. 2025. Semantic–lexical fusion: Improving retrieval accuracy for AI-driven knowledge systems. *Int. J. Comput. Exp. Sci. Eng.*, 11.

G. V. Cormack, C. L. A. Clarke, and S. Buettcher. 2009. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proc. 32nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, pages 758–759, Boston, MA, USA.

L. Dietz, M. Verber, F. Arber, and K. Berberich. 2018. Utilizing knowledge graphs for text-centric information retrieval. In *Proc. SIGIR*, pages 1387–1390.

D. Edge and 1 others. 2025. From local to global: A graph RAG approach to query-focused summarization. *Preprint*, arXiv:2404.16130.

H.-L. Hsu and J. Tzeng. 2025. DAT: Dynamic alpha tuning for hybrid retrieval in retrieval-augmented generation. *Preprint*, arXiv:2503.23013.

H. Huang, Z. Zheng, Y. Zhao, and J. Zhang. 2025. Retrieval-augmented generation with hierarchical knowledge. In *Findings of ACL: EMNLP 2025*, pages 6044–6060.

V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-T. Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proc. EMNLP*, pages 6769–6781.

P. Lewis and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proc. 34th Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, BC, Canada. Art. 793.

J. Lin. 2022. A proposed conceptual framework for a representational approach to information retrieval. *SIGIR Forum*, 55(2). Art. 4.

J. Lin, X. Ma, S.-C. Lin, J.-H. Yang, M. Pradeep, and R. Nogueira. 2021. Pyserini: A Python toolkit for reproducible IR research. In *Proc. SIGIR*, pages 2356–2362.

N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. 2024. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguist.*, 12:157–173.

Y. Luan, J. Eisenstein, K. Toutanova, and M. Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Trans. Assoc. Comput. Linguist.*, 9:329–345.

C. S. Mala, G. Gezici, and F. Giannotti. 2025. Hybrid retrieval for hallucination mitigation in large language models: A comparative analysis. *Preprint*, arXiv:2504.05324.

L. Medrano, A. Verma, and M. Chhabra. 2026. Scaling retrieval augmented generation with RAG fusion: Lessons from an industry deployment. *Preprint*, arXiv:2603.02153.

A. Mourão, F. Martins, and J. Magalhães. 2014. Inverse square rank fusion for multimodal search. In *Proc. 12th Int. Workshop Content-Based Multimedia Indexing (CBMI)*, pages 1–6.

A. Neelakantan, T. Xu, R. Puri, A. Radford, J. M. Han, J. Tworek, Q. Yuan, N. Tezak, J. W. Kim, C. Hallacy, J. Heidecke, P. Shyam, B. Power, T. E. Nekoul, G. Sastry, G. Krueger, D. Schnurr, F. P. Such, K. Hsu, and 6 others. 2022. Text and code embeddings by contrastive pre-training. *Preprint*, arXiv:2201.10005. OpenAI `text-embedding` model family.

L. Page, S. Brin, R. Motwani, and T. Winograd. 1999. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, Stanford, CA, USA.

A. Raja and P. Vats. 2025. Beyond nearest neighbors: Semantic compression and graph-augmented retrieval for enhanced vector search. In *Proc. 3rd Int. Workshop Deep Multimodal Generation Retrieval*.

S. Robertson and H. Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

B. Sarmah, D. Mehta, B. Hall, R. Rao, S. Patel, and S. Pasquali. 2024. HybridRAG: Integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction. In *Proc. 5th ACM Int. Conf. AI Finance (ICAIF)*, pages 608–616, Brooklyn, NY, USA.

Y. Sun. 2025. Adaptive hybrid retrieval-augmented generation with context-aware confidence control. In *Proc. 4th Int. Conf. Electron. Inf. Technol. (EIT)*, pages 891–895.

V. A. Traag, L. Waltman, and N. J. van Eck. 2019. From Louvain to Leiden: Guaranteeing well-connected communities. *Sci. Rep.*, 9. Art. 5233.

D. Vujić, A. Njeguš, and N. Bacanin. 2025. Enhancing retrieval-augmented generation with graph-based retrieval and generative modeling. In *SINTEZA*.

L. Zheng and 1 others. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*.

X. Zhu, Y. Xie, Y. Liu, Y. Li, and W. Hu. 2025. Knowledge graph-guided retrieval augmented generation. In *Proc. NAACL*, pages 8912–8924, Albuquerque, NM, USA.