MULTIRAG-BENCH: BENCHMARKING JOINT RETRIEVAL VERSUS KNOWLEDGE GRAPH FUSION FOR MULTI-DOMAIN RAG

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph Retrieval-Augmented Generation (RAG) systems are widely deployed, yet a fundamental question remains unexamined: when multiple domain-specific Graph RAGs are independently constructed over the same corpus using different extraction models and prompts, should their knowledge graphs be fused into one unified structure or queried jointly while preserving domain separation? We present MULTIRAG-BENCH, the first benchmark for this problem. Our dataset comprises five domain-specific knowledge graphs built from a common document corpus, a multi-model heterogeneity analysis demonstrating that different LLM architectures produce substantially divergent graph structures even under identical prompts, and a stratified question set whose cross-domain requirements are each verified via direct Neo4j graph traversal, ensuring that cross-domain questions genuinely require evidence from multiple graphs. We establish baselines for both graph fusion and joint retrieval paradigms. We find that the merge penalty flips sign with label quality: naive fusion introduces hub noise under text-level cross-domain labels that are not graph-verified, whereas on our graph-verified benchmark naive merging unlocks genuine cross-domain evidence. Pooling independent domain retrievals accounts for most of the achievable gain over single-domain baselines, with cache-based enrichment adding only a small bridge-recovery increment and absolute retrieval quality remaining far from ceiling on the hardest multi-hop scenarios. We release all code, data, and baselines.

1 INTRODUCTION

Graph Retrieval-Augmented Generation (Graph RAG) enhances Large Language Model (LLM) reasoning by extracting structured knowledge graphs (KGs) from text and leveraging graph traversal for evidence retrieval. In practice, organizations routinely construct *multiple domain-specific* KGs, each tailored to a knowledge vertical through customized extraction prompts, entity vocabularies, and relation schemas. Even a single corpus processed through different extraction configurations yields substantially divergent graph structures, a phenomenon documented in the LLM-based KG construction literature (Zhu et al., 2024). We demonstrate empirically that different extraction prompts on identical documents produce KGs whose pairwise relation overlap averages only 5.7%, and that switching the underlying LLM architecture drives this overlap down further to 1.1%. This heterogeneity is a structural property of any deployment where multiple teams independently configure their Graph RAG pipelines.

Despite this ubiquity, the resulting retrieval challenge has not been benchmarked. Graph RAG systems (Fan et al., 2024; Edge et al., 2024; Guo et al., 2025; Chen et al., 2026; Gutiérrez et al., 2024; Ma et al., 2025; Zhu et al., 2025; Mavromatis & Karypis, 2024; Liang et al., 2024; Li et al., 2024) remain confined to single-graph operation, and existing benchmarks (Es et al., 2024; Xiao et al., 2025; Tang & Yang, 2024; Yang et al., 2024a; Trivedi et al., 2022; Yang et al., 2018; Ho et al., 2020) measure retrieval within a single KG or over flat document collections. Moreover, the fused architecture is often infeasible: storage compounds with each new domain, distributed ownership precludes centralized merging, and compliance may prohibit data from crossing organizational boundaries. The central question of this work, illustrated in Figure 1, is therefore: given N heterogeneous domain-specific KGs that cannot be physically merged, how can we enable effective cross-domain retrieval?

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.

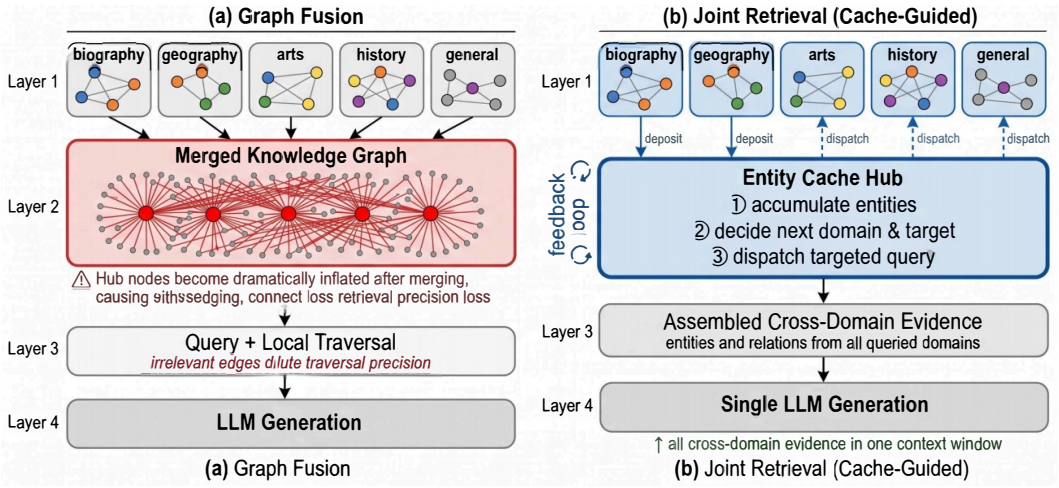


Figure 1: Two competing paradigms for multi-domain Graph RAG retrieval. **(a) Graph Fusion:** KGs are physically merged; hub inflation may degrade traversal precision, though whether this penalty materializes depends on label validity. **(b) Joint Retrieval:** an entity cache hub coordinates retrieval across independent domain graphs.

Answering this question requires understanding the cost of each paradigm. We find that graph fusion, even when feasible, incurs a measurable hub noise penalty: merging domain graphs inflates high-degree hub nodes several-fold, causing local traversal to follow irrelevant edges and degrading retrieval precision; as our experiments show, however, whether this penalty materializes depends critically on whether the evaluation labels themselves are graph-verified. Joint retrieval, the alternative, is in its infancy: cache-based methods that propagate entity evidence across sequential domain queries, inspired by retrieval-augmented reasoning (Yao et al., 2023; Shi et al., 2024; Trivedi et al., 2023), can surpass mean single-domain performance on moderate cross-domain questions, but absolute retrieval quality remains far from ceiling, and harder multi-hop regimes remain open.

To address this problem, we present MULTIRAG-BENCH, the first benchmark for fusion-vs-joint retrieval over LLM-extracted, prompt-divergent domain KGs. The dataset consists of five domain-specific KGs built from the same 21,100-document corpus under different extraction prompts. Cross-domain questions are mined from verified multi-hop chains and validated through direct Neo4j graph traversal, guaranteeing that no single-domain graph contains a complete answer path; applying the same graph-traversal check to MuSiQue shows that 96% of its degree-2 questions are solvable within a single domain graph under our construction. The evaluation framework, overviewed in Figure 2, compares graph fusion against joint retrieval across stratified hop counts using four basic retrieval metrics (PHR, ER, SNR, BR) computed from known gold chains, plus two derived diagnostics (MP, Δ) that isolate the fusion-vs-joint trade-off. We release all code, data, and baselines. Our instantiation is deliberately narrow (one shared corpus, one Graph RAG framework, one extractor for the main graphs); we quantify extractor divergence structurally (§4.3) and leave multi-framework and disjoint-corpus instantiations to future work. In summary, our contributions are:

1. **A benchmark with graph-verified cross-domain requirements.** MULTIRAG-BENCH is the first to verify cross-domain questions by direct multi-graph traversal under edge-exclusivity constraints, guaranteeing the gold path exists only across graphs; applying the same check to MuSiQue degree-2 labels shows 96% are solvable within a single domain graph under our construction, motivating graph-verified construction.

2. **Joint-retrieval baselines with headroom analysis.** Cache-based joint retrieval surpasses the mean and best single-domain baselines at 1-hop and 2-hop, primarily through result pooling (§4.3), but absolute PHR remains far from ceiling, establishing multi-domain retrieval as an open challenge.

3. **Multi-model heterogeneity quantification.** Different LLM architectures under identical prompts produce KGs with lower relation overlap than different prompts on the same model, confirming structural inevitability of KG divergence.

4. **A cache-based retrieval baseline with an honest decomposition.** A no-cache union control quantifies how much of joint retrieval's gain derives from result pooling versus cache propagation, establishing a calibrated reference point for future multi-domain methods.

2 Related Work

Graph RAG retrieval methods. Graph RAG systems construct knowledge graphs from text and retrieve evidence via graph traversal (Peng et al., 2025; Fan et al., 2024), with community-wide benchmarking accelerated by the CRAG challenge (Yang et al., 2024a; Xie & Liang, 2024). GraphRAG (Edge et al., 2024) combines community detection with hierarchical summaries; LightRAG (Guo et al., 2025) mixes keyword-guided local traversal with global vector search; GNN-RAG (Mavromatis & Karypis, 2024), KAG (Liang et al., 2024), StructRAG (Li et al., 2024), PathRAG (Chen et al., 2026), HippoRAG (Gutiérrez et al., 2024) and its memory-based successor (Jiménez Gutiérrez et al., 2025), Think-on-Graph (Sun et al., 2024) and its successor ToG 2.0 (Ma et al., 2025), KG2RAG (Zhu et al., 2025), and path-grounded retrievers such as RoG (Luo et al., 2024) and G-Retriever (He et al., 2024) refine single-graph retrieval along orthogonal axes. All share a common assumption: retrieval operates over a single unified KG. None evaluate retrieval quality when multiple LLM-extracted, relation-disjoint KGs coexist and must be queried jointly, the scenario MultiRAG-Bench targets.

Knowledge graph fusion and federated access. Entity alignment methods (Xin et al., 2022; Yang et al., 2024b; Sun et al., 2018; Suchanek et al., 2011) identify equivalent entities across KGs, from trained matchers to unsupervised probabilistic alignment; LLM-based construction (Zhu et al., 2024; Zhang & Soh, 2024) builds unified graphs end-to-end; KG embeddings (Bordes et al., 2013; Sun et al., 2019; Trouillon et al., 2016) learn cross-schema representations; QMKGF (Wei et al., 2026) fuses multi-source subgraphs in a query-aware manner for RAG. Federated SPARQL engines such as FedX (Schwarte et al., 2011) and SPLENDID (Görlitz & Staab, 2011) studied source selection and join planning over autonomous endpoints that cannot be merged, benchmarked by FedBench (Schmidt et al., 2011), the classical analogue of our un-mergeable regime. This literature asks *how* to fuse or federate KGs effectively; MultiRAG-Bench instead asks *whether* fusion benefits downstream retrieval when the sources are LLM-extracted, structurally divergent views whose relation vocabularies were never designed to interoperate, and isolates that question with dedicated diagnostics.

Multi-source RAG. RAG-Fusion (Rackauckas, 2024) applies reciprocal rank fusion at the chunk level; DPR (Karpukhin et al., 2020) and ColBERT (Khattab & Zaharia, 2020) retrieve over flat indexes (Zhao et al., 2024b; Robertson & Zaragoza, 2009); Self-RAG (Asai et al., 2024) and IR-CoT (Trivedi et al., 2023) interleave retrieval with self-reflection or chain-of-thought reasoning; HybridRAG (Sarmah et al., 2024) and HeteRAG (Yang et al., 2025) combine vector and single-graph or heterogeneous chunk signals. These methods combine evidence at the chunk, vector, or single-graph level, but none traverse graph structures residing in separate databases with different schemas and extraction biases.

Evaluation benchmarks and label validity. RAGAS (Es et al., 2024), ARES (Saad-Falcon et al., 2024), RAGBench (Friel et al., 2024), GraphRAG-Bench (Xiao et al., 2025), MultiHop-RAG (Tang & Yang, 2024), and CRAG (Yang et al., 2024a) evaluate single-RAG or flat-corpus retrieval, as documented in evaluation surveys (Yu et al., 2024; Zhao et al., 2024a). MuSiQue (Trivedi et al., 2022), HotpotQA (Yang et al., 2018), and 2WikiMQA (Ho et al., 2020) target multi-hop QA (Mavi et al., 2024); KG-grounded QA work (Cattaneo et al., 2025; Yin et al., 2025) advocates verified ground-truth subgraphs, a principle we adopt through our Neo4j-verified gold chains. Closer to our motivation, Min et al. (2019) show that compositional questions do not necessitate multi-hop reasoning, an early warning that text-level multi-hop labels can overstate task difficulty; our graph-traversal check extends this concern to the multi-graph setting and quantifies it on MuSiQue (Appendix A).

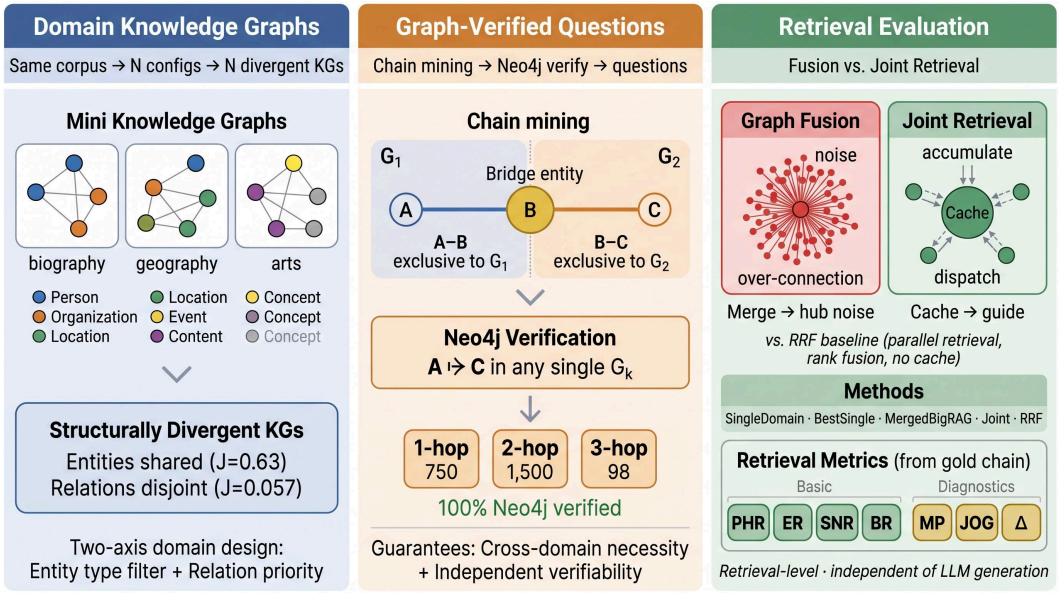


Figure 2: Overview of MULTIRAG-BENCH. **Left:** N domain-specific KGs are built from a shared corpus under different extraction configurations. **Center:** Cross-domain questions are mined via chain traversal with Neo4j-verified cross-domain guarantees. **Right:** Retrieval is evaluated under both graph fusion and cache-based joint retrieval using gold-chain metrics; whether the fusion hub-noise penalty materializes depends on label validity (§4.3).

3 MULTIRAG-BENCH

We present MULTIRAG-BENCH as a general methodology with a concrete instantiation. Figure 2 overviews the approach: domain-specific KGs are constructed from a shared corpus under different extraction configurations (§3.1), cross-domain questions are mined via graph-verified chain traversal (§3.2), and retrieval is evaluated under both graph fusion and joint retrieval (§3.3). We describe each component in general terms, then present our $N = 5$ instantiation in §3.5.

3.1 DOMAIN KNOWLEDGE GRAPH CONSTRUCTION

The first stage constructs N structurally divergent knowledge graphs from a shared document corpus. Let $\mathcal{D}$ be a document corpus and $\mathcal{C} = \{C_1, \ldots, C_N\}$ a set of extraction configurations, where each C_i specifies an entity type filter $\mathcal{E}_i$ and a relation type priority $\mathcal{R}_i$. Applying each C_i to $\mathcal{D}$ via a Graph RAG extraction pipeline $\mathcal{P}$ yields N domain-specific KGs $\mathcal{G} = \{G_1, \ldots, G_N\}$, where $G_i = \mathcal{P}(\mathcal{D}, C_i)$ has nodes V_i drawn from $\mathcal{E}_i$ and edges extracted under $\mathcal{R}_i$ constraints (Zhu et al., 2024; Zhang & Soh, 2024; Shomer et al., 2025; Touvron et al., 2023; OpenAI, 2023). Because each C_i encodes different domain priorities, $\mathcal{G}$ exhibits structural heterogeneity: entities are largely shared across graphs while relations are largely disjoint. Concretely, each domain is defined along two axes: an *entity-type axis* restricting extraction to a prioritized subset of an entity vocabulary $\mathcal{T}$, and a *relation-type axis* assigning prioritized relation categories biasing extraction toward domain-relevant connections. Table 2 defines our $N = 5$ instantiation over $\mathcal{T} = \{\text{Person}, \text{Organization}, \text{Content}, \text{Event}, \text{Location}, \text{Concept}\}$; each domain KG is extracted using LightRAG with a domain-specific prompt on the same 21,100-document MuSiQue corpus (extraction model and configuration in Appendix A).

3.2 GRAPH-VERIFIED QUESTION GENERATION

This stage constructs cross-domain questions with a guarantee that no single domain graph contains a complete answer path. Given N domain graphs and a hop bound h, a valid h-hop cross-domain chain is a triple (A, B, C) satisfying (i) $A \in V_i$ and edge A–B exists exclusively in G_i, (ii) $B \in$

Table 1: Evaluation metrics: basic (top) and derived diagnostics (bottom).

Basic Metric	**Definition**	**Dimension**
PHR (Path Hit Rate)	all gold chain entities retrieved	Completeness
ER (Entity Recall)	fraction of gold entities in $\mathcal{R}(Q)$	Recall
SNR (Signal-to-Noise)	gold entities / total retrieved	Precision
BR (Bridge Recovery)	bridge entity retrieved	Cross-domain link
Derived Diagnostic	**Definition**	**Reveals**
MP (Merge Penalty)	$\text{PHR}_{\text{Best}} - \text{PHR}_{\text{Merged}}$	Fusion cost (sign label-dependent)
Δ (Cross-Domain Gain)	$\text{PHR}_{\text{Joint}} - \text{PHR}_{\text{Mean}}$	Multi-domain benefit

$V_i \cap V_j$ is a bridge entity, (iii) $C \in V_j$ and edge B–C exists exclusively in G_j, and (iv) $A \not\to C$ in every single G_k within h hops. For example, the question "Which country's election did a candidate from a Belarusian protest movement participate in?" yields the chain (March 19, 2006 Belarusian election; Alyaksandr Kazulin; 2006 Belarusian election), where A–B is exclusive to the arts domain and B–C to the history domain.

The mining algorithm executes three steps for each ordered domain pair (d_1, d_2): (1) *bridge identification*, enumerating all $B \in V_{d_1} \cap V_{d_2}$; (2) *exclusive edge filtering*, retaining only neighbors whose edges to B are domain-exclusive; (3) *reachability verification*, discarding chains where A reaches C in any single graph. The process extends to three-hop chains across domain triples, and verified chains are converted to natural-language questions by an LLM. This guarantees *cross-domain necessity* (no single graph contains a complete answer path) and *verifiability* (every chain is independently auditable via the same graph queries used during construction).

3.3 Evaluation Framework

Task definition. The benchmark compares two retrieval paradigms across stratified hop counts (1-hop, 2-hop, 3-hop). *Graph Fusion* merges all G_i into $G_{\text{merged}} = \bigcup_i G_i$ and queries this single graph; *Joint Retrieval* queries each G_i separately, propagating cross-domain evidence via a shared entity cache. Evaluating the same questions under both paradigms isolates the cost of the fusion-vs-joint design choice. Four methods are evaluated: **SingleDomain** (per-domain independent hybrid retrieval, with the per-domain mean MeanSingle as a no-cross-domain-evidence baseline), **BestSingle** (a reference selecting the best per-question single-domain result, an oracle domain selector though not an upper bound on single-domain retrieval), **Merged** (physical graph fusion via name-based entity deduplication), and **Joint** (the cache-based pipeline of §3.4).

Metrics. Every question carries a known gold chain verified via Neo4j traversal (§3.2), so evaluation is performed at the retrieval level. For an h-hop question with gold chain entities $E^* = \{e_1, \ldots, e_{h+1}\}$, let $\mathcal{R}(Q)$ denote the set of entities returned by a retrieval method for question Q. Table 1 summarizes the four basic metrics and two derived diagnostics. **Path Hit Rate** (PHR_h) is the fraction of h-hop questions where all gold chain entities appear in $\mathcal{R}(Q)$, a strict completeness measure. **Entity Recall** (ER_h) is the average fraction of gold chain entities appearing in $\mathcal{R}(Q)$. **Signal-to-Noise Ratio** (SNR_h) is the fraction of retrieved entities belonging to the gold chain. **Bridge Recovery** (BR_h) is the fraction of questions where the bridge entity appears in $\mathcal{R}(Q)$, isolating the cross-domain linking step; BR uses the chain's b-slot entity throughout (the bridge on the 2-hop stratum, the target entity on the 1-hop control tier, the first intermediate entity at 3-hop). Two derived diagnostics reveal the fusion-vs-joint trade-off: the **Merge Penalty** $\text{MP}_h = \text{PHR}_h^{\text{BestSingle}} - \text{PHR}_h^{\text{Merged}}$ (positive when name-deduplicated merging misses gold evidence that the best single domain finds; negative when merging unlocks cross-domain edges unavailable to any single graph), and the **Cross-Domain Gain** $\Delta_h = \text{PHR}_h^{\text{Joint}} - \text{PHR}_h^{\text{MeanSingle}}$ (positive when joint retrieval exploits multi-domain evidence beyond any single domain). Entity matching uses exact name equality on MULTIRAG-BENCH (gold chains share the graphs' node vocabulary) and fuzzy substring matching on MuSiQue (Wikipedia-derived surface variants); each hybrid query returns its top-5 entities per queried index. Full configuration and per-tier metric selection rules ship with the released evaluation code (Appendix A).

Table 2: Domain definitions for our $N = 5$ instantiation.

Domain	Entities	Relations
biography	Person, Org, Content	born_in, spouse_of, member_of
geography	Location, Org, Event	located_in, borders, capital_of
arts	Content, Person, Org	directed_by, performed_by, released_by
history_politics	Org, Person, Event	governed, ruled, signed
general	all six types (baseline)	(no priority)

Table 3: MultiRAG-Bench question distribution.

	1-hop	2-hop	3-hop	Total
Questions	750	1,500	98	**2,348**
Domain coverage	5 domains $\times$ 150	10 pairs $\times$ 150	all triples	—
Neo4j-verified	✓	✓	✓	**100%**

3.4 Cache-Based Joint Retrieval

The Joint Retrieval baseline propagates cross-domain evidence without physical merging. Each domain graph G_i is queried through a retrieval function $\mathcal{R}_{d_i} : Q \to 2^{V_i}$. *Stage 1 (cache initialization)* queries the first domain d_1 with the original question Q, initializing the cache $\mathcal{C}_{\text{prev}} \leftarrow \mathcal{R}_{d_1}(Q)$. *Stage 2 (sequential enrichment)* issues to each subsequent domain d_i ($i = 2, \ldots, N$) an enriched query

$$Q_i^{\text{enriched}} = Q \ \oplus \ \text{top}_k(\mathcal{C}_{\text{prev}}, k = 5),$$

where top_k selects up to k entities from the previous domain's result, ranked by entity name length (longer names are more specific retrieval anchors), and $\oplus$ denotes string concatenation. *Stage 3 (assembly)* aggregates all retrieved sets: $\mathcal{R}_{\text{joint}}(Q) = \bigcup_{i=1}^{N} \mathcal{R}_{d_i}(Q_i^{\text{enriched}})$ with $Q_1^{\text{enriched}} = Q$. The key mechanism is that entities discovered in early domains bias subsequent retrieval toward adjacent graph regions: for LightRAG's hybrid retrieval, appending entity names increases keyword match density, so a bridge entity found in G_i becomes a keyword anchor when querying G_j, increasing the probability that G_j's result includes entities connected to the bridge. Ablations show the pipeline is insensitive to domain order and cache ranking (PHR within 1pp, BR within 1.6pp of random replacements; Appendix B), and a no-cache union control isolates cache propagation versus result pooling (§4.3).

3.5 Dataset

We instantiate the methodology with $N = 5$ domains (Table 2): biography focuses on person-centric relations, geography on spatial relations, arts on creative production relations, and history_politics on governance relations, with general as an unconstrained baseline. Table 3 reports the question distribution: **1-hop questions** (750) serve as a single-domain control group, **2-hop questions** (1,500, covering all 10 domain pairs) form the core cross-domain evaluation set, and **3-hop questions** (98) probe three-domain scenarios and are treated as exploratory throughout.

Our chain mining pipeline examined $\sim$390,000 candidate triples; the edge-exclusivity filter rejected 88%, and Neo4j reachability verification retained $\sim$5,200 two-hop chains and only 98 three-hop chains, reflecting the compounding difficulty of finding consecutive domain-exclusive edges. Chains are mined by seeded random sampling (fixed seed; shuffled bridge and neighbor candidates; 3$\times$ per-pair oversampling), the final 2-hop set is balanced at 150 questions per domain pair after question validation, and all 98 verified 3-hop chains are retained. Verified chains are converted to natural-language questions by DeepSeek-V4-Flash, and every generated question is programmatically validated against its Neo4j gold chain. All 2,348 questions are Neo4j-verified: the 1,598 cross-domain questions are guaranteed to have no complete answer path in any single domain graph, while the 750 1-hop questions serve as a single-domain control set. A question-quality audit (entity-leakage check over all questions; LLM-judged naturalness on a stratified sample) is reported in Appendix B.

Table 4: Retrieval performance: our benchmark vs. original MuSiQue, under two retrieval backends. MuSiQue uses fuzzy entity matching; MultiRAG-Bench uses exact matching (gold chains Neo4j-verified). The 0% single-domain solvability of the 1,598 cross-domain questions is by construction (§3.2); the 750 1-hop questions are single-domain controls. PHR measures entity co-presence, a weaker condition than path existence. By convention MP is printed on the BestSingle row and Δ on the Joint row; "—" denotes structurally undefined cells. Arrows mark the desirable direction (negative MP favors merging; positive Δ favors joint retrieval); **bold** is the per-column best within each panel. Joint unions five result sets ($|\mathcal{R}(Q)| \approx 39$ vs. ≈ 8 per domain; budget controls in §4.3). Wilson 95% CI half-widths for PHR/BR: $\leq$2.0pp overall, 2.4–3.5pp (1-hop), 0.9–2.5pp (2-hop), 7.7–9.7pp (3-hop, n=98). All metrics in %.

(a) LightRAG (hybrid) backend						
Method	**PHR↑**	**ER↑**	**SNR↑**	**BR↑**	**MP↓**	**Δ ↑**
MultiRAG-Bench (2,348 questions)						
MeanSingle	7.38	38.52	**6.52**	47.80	—	—
BestSingle	31.86	—	—	—	−22.02	—
Merged	53.88	73.02	5.79	73.68	—	—
Joint	**58.13**	**78.05**	3.87	**80.24**	—	+50.76
MuSiQue degree-2 (2,504 questions)						
MeanSingle	10.86	32.62	**3.96**	50.69	—	—
BestSingle	24.36	—	—	—	+10.06	—
Merged	14.30	37.52	2.46	56.67	—	—
Joint	**26.84**	**54.27**	1.90	**78.35**	—	+15.97
(b) PG-Direct (embedding) backend						
MeanSingle	6.62	37.79	13.43	47.37	—	—
BestSingle	27.34	—	—	—	−16.52	—
Merged	43.87	66.01	**13.97**	65.93	—	—
Joint	**50.51**	**70.55**	10.69	**73.13**	—	+43.89
MuSiQue degree-2 (2,504 questions)						
MeanSingle	6.15	28.28	**7.64**	47.14	—	—
BestSingle	10.94	—	—	—	+4.71	—
Merged	6.23	28.83	5.96	47.28	—	—
Joint	**11.34**	**36.70**	5.14	**58.87**	—	+5.19

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and backends. We evaluate on MULTIRAG-BENCH ($N = 5$, 2,348 questions) and MuSiQue (Trivedi et al., 2022) (2,504 degree-2 questions, text-level cross-domain labels), both constructed over the same 21,100-document Wikipedia corpus. We compare two retrieval architectures: **LightRAG** (Guo et al., 2025), which stores entities and relations in Neo4j with vector/keyword indices in PostgreSQL under hybrid mode, and **PG-Direct**, a graph-free ablation performing pure embedding-based cosine similarity retrieval, isolating the contribution of graph structure. All domain KGs and the merged graph are extracted with DeepSeek-V4-Flash using BGE-M3 embeddings (1024-d); each hybrid query returns its top-5 entities per index (Appendix A). For Joint Retrieval, cache size is $k = 5$ and domain query order follows the fixed sequence {biography, geography, arts, history_politics, general}.

Implementation. All experiments use asyncio-based parallelism with batch size 50. Retrieval metrics are computed per-question against known gold chains using exact entity-name matching on MULTIRAG-BENCH and fuzzy substring matching on MuSiQue (to account for entity name variations in Wikipedia-derived graphs). Evaluation is performed at the retrieval level; all reported numbers are computed by the released deterministic evaluation code from logged retrieval outputs.

4.2 Main Results: Retrieval Performance

Table 4 reports retrieval performance across two datasets and two backends; we structure the analysis around metric validity and data validity.

Benchmark validity. The four basic metrics reveal complementary dimensions that a single score conflates. SNR is consistently low (1.90–13.97% across methods and backends), meaning 86–98% of retrieved entities are irrelevant to the gold chain: hub nodes dominate local traversal, pulling in semantically adjacent but causally irrelevant entities. SNR is, however, not comparable across methods with different retrieved-set sizes (Joint pools $|\mathcal{R}(Q)| \approx 39$ vs. ≈ 8 per single domain, which mechanically deflates its SNR), so we treat cross-method SNR comparisons as confounded and use the budget-matched control (§4.3) to quantify size effects. BR isolates the cross-domain linking step: Joint reaches 80.24% BR versus MeanSingle's 47.80% under LightRAG (+32.44pp; §4.3 quantifies how much of this increment is pooling). The ER–PHR gap (78.05% vs. 58.13%) further decomposes failure into two regimes: most gold entities are found, but assembling the complete chain remains the bottleneck.

MeanSingle PHR on MultiRAG-Bench is only 7.38% (LightRAG) and 6.62% (PG-Direct), consistent with the verification guarantee: no single domain contains a complete gold path by design (PHR's co-presence semantics is weaker, so non-zero single-domain PHR does not contradict it). Joint achieves 58.13% PHR under LightRAG (7.9× MeanSingle), surpassing the best single-domain reference BestSingle (31.86%) by 26.27pp, quantifying the headroom that cross-domain evidence aggregation exploits. This headroom does not exist on MuSiQue, where Joint and BestSingle differ by only 2.48pp, confirming that our benchmark creates a fundamentally different retrieval challenge. Entity leakage tempers these absolutes (89.3% of questions name a gold-chain entity, 50.6% of 2-hop questions name the bridge), so PHR and BR are optimistic; the cross-domain gap survives on clean-bridge questions (43.7% vs. 9.9% PHR, cache $+0.27$pp; Appendix B).

Data validity. The Merge Penalty acts as a label-property diagnostic: its sign tracks the single-domain solvability of the labels rather than retrieval behavior per se. On MuSiQue, merging under name deduplication degrades retrieval (MP $= +10.06$ LightRAG, $+4.71$ PG-Direct) because its text-level labels are not graph-verified under our construction: 96% of questions are solvable within a single domain, so merging adds hub noise without cross-domain benefit. On MultiRAG-Bench the opposite holds (MP $= -22.02$ and -16.52): merging improves retrieval because questions genuinely require edges no single domain contains. Correspondingly, Δ is $+50.76$ (LightRAG) on our benchmark versus $+15.97$ on MuSiQue, reflecting the 0% vs. 96% solvability gap. Both backends agree on the MP sign reversal and Δ ordering, so these signals reflect benchmark properties rather than backend artifacts. The reversal also survives matcher unification: rerunning MuSiQue retrieval under the exact matcher (LightRAG backend) yields MP $= +5.83$ versus -22.02 on MultiRAG-Bench, so the sign flip is not a matcher artifact (solvability protocol in Appendix A).

4.3 Benchmark Design Validation

Structural inevitability of KG heterogeneity. To isolate model effects from prompt effects, we build four LightRAG instances with open-weight 7B–9B models (Llama-3.1-8B, Mistral-7B, Gemma-2-9B, Dolphin-Llama-3-8B, the last a fine-tune of Llama-3) using the identical general prompt on 1,000 documents (Table 5). Mean pairwise relation Jaccard is 0.011, versus 0.063 for the five prompt-domain graphs *restricted to the same 1,000 documents* (0.057 on the full corpus; Appendix C): model-architecture divergence exceeds prompt divergence at matched corpus scale, ruling out a corpus-size artifact, and entity overlap drops from 0.63 to 0.50. Model-architecture differences thus produce even more divergent graphs than prompt customization: any deployment of independently managed Graph RAG instances will face the fusion-vs-joint decision our benchmark evaluates.

Cross-domain barrier by hop count. Table 6 stratifies PHR across hop tiers. The 1-hop to 2-hop transition confirms the intended design: MeanSingle collapses from 13.47% to 3.57% and BestSingle from 60.00% to 16.40%, validating that 2-hop questions impose a barrier single-domain retrieval cannot overcome. MP is near zero at 1-hop ($+0.27$pp) and strongly negative at 2-hop (-34.47pp): merging unlocks cross-domain edges no single graph contains. Δ is robustly positive ($+48$–52pp) across the 1-hop and 2-hop tiers. The 3-hop column is *exploratory* (n=98): MeanSingle rises to 18.98% and MP shrinks to -2.04pp, reflecting higher entity co-occurrence in surviving chains (53%

Table 5: Multi-model heterogeneity (identical prompt, 1,000 documents).

Model	Entities	Relations
Llama-3.1-8B	5,567	2,490
Gemma-2-9B	5,858	2,166
Dolphin-Llama-3-8B	4,117	1,581
Mistral-7B	3,950	1,043
Mean pairwise relation Jaccard	**0.011**	
Mean pairwise entity Jaccard	0.50	

Table 6: Retrieval PHR stratified by hop count (LightRAG hybrid, MultiRAG-Bench). All values in %. 3-hop (n=98) is exploratory.

Method	1-hop	2-hop	3-hop
MeanSingle	13.47	3.57	18.98
BestSingle	60.00	16.40	53.06
Merged	59.73	50.87	55.10
Joint	61.60	55.20	76.53
Δ (Joint $-$ Mean)	+48.13	+51.63	+57.55
MP (Best $-$ Merged)	+0.27	-34.47	-2.04

of 3-hop chains have all four entities present in some single domain, vs. 16% for 2-hop); PHR measures entity co-presence, a weaker condition than our path-existence guarantee.

Joint Retrieval analysis and union control. A no-cache *union control* (querying all five domains independently with the original question and pooling the results) isolates how much of Joint's advantage derives from cache propagation versus mere result pooling. All variants share a common keyword extractor (DeepSeek-V4-Flash/local-gemma2), so contrasts are extractor-fair, though absolute values shift vs. Table 4 (Merged 53.88 here vs. 34.28 PHR; naive fusion is extractor-sensitive). The union control attains 54.64% PHR and 77.68% BR versus Joint's 56.52% and 80.03% (+1.87pp PHR, $p = 0.10$; +2.34pp BR, $p = 0.02$; paired bootstrap; the BR increment does not survive Holm correction): result pooling accounts for most of the gain, while cache propagation contributes a small, bridge-recovery-oriented increment concentrated at 1-hop (+3.60pp PHR), shrinking at 2-hop (+1.33pp), and negative at the exploratory 3-hop tier (-3.06pp; Appendix B). Retrieved-set size is itself a first-order factor: at a matched budget of 25 entities the union drops to 23.08% PHR. The fusion side is equally robust to baseline choice: replacing exact-name deduplication with embedding-similarity alignment merges 16.3% additional entities and moves MP from -6.94 (naive) to $+7.50$ (shared-extractor values, single tested threshold $\tau = 0.90$; Table 4 reports -22.02; Appendix B): at this threshold, aggressive alignment amplifies, rather than mitigates, the fusion cost. We therefore present Joint as a pooling-plus-cache baseline rather than evidence that current cache mechanisms solve multi-domain retrieval, and release the union control as a mandatory reference point (Appendix B).

5 Conclusion

We presented MultiRAG-Bench, the first benchmark for joint retrieval versus graph fusion across LLM-extracted domain-specific Graph RAG systems. The benchmark combines a general methodology (N extraction configurations over a shared corpus, graph-verified chain mining, fusion-vs-joint diagnostics) with a 5-domain, 2,348-question instantiation. Our experiments yield three findings: single-domain retrieval is structurally insufficient for cross-domain questions (MeanSingle PHR 13.47% vs. 3.57%); the Merge Penalty flips sign with label quality, tracking the construction-enforced solvability contrast; and the 7.9$\times$ PHR gain comes mostly from result pooling, cache propagation adding only +2.3pp BR that is not robust to multiple-comparison correction. Limitations include LLM-generated questions without human audit, a single corpus/framework/threshold, unranked logs (no size-controlled SNR), and the exploratory 3-hop tier (n=98). We release all code, data, prompts, graphs, and evaluation scripts.

AI USE STATEMENT

In this work, we used generative AI in two roles. First, DeepSeek-V4-Flash converted graph-verified chains into candidate natural-language questions (§3.5); every candidate was validated against its Neo4j gold chain, and the validation procedure is part of the released pipeline. Second, generative AI coding assistants assisted in drafting implementation and evaluation scripts; all such code was executed, tested, and reviewed by the authors. All reported numbers are computed by deterministic evaluation code from logged retrieval outputs. We did not use generative AI to originate the paper's scientific claims, and we have reviewed all AI-assisted content. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All benchmark components are released: the five domain-specific extraction prompts, the LightRAG-based graph construction pipeline, the Neo4j chain-mining and verification scripts, the 2,348 graph-verified questions with gold chains, all baseline implementations (SingleDomain, BestSingle, Merged, Joint), evaluation code for all metrics (PHR, ER, SNR, BR, MP, Δ) with per-question logging, and the ablation suite described in §3.4. Every question can be independently audited by re-running the graph queries used during construction. Experimental configuration (extraction model, embeddings, retrieval parameters, matching rules) is specified in Appendix A. The benchmark is derived from MuSiQue, whose source text is Wikipedia content distributed under CC BY-SA; the released graphs and questions therefore carry attribution and share-alike obligations, and MuSiQue's redistribution terms apply to the derived question set. All artifacts are released for research use, and practitioners applying the construction pipeline to their own corpora should verify that ownership, licensing, and compliance constraints permit the cross-graph queries the methodology requires.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, et al. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *ICLR*, 2024.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. Translating Embeddings for Modeling Multi-relational Data. In *NeurIPS*, 2013.

Alberto Cattaneo, Carlo Luschi, and Daniel Justus. Ground-Truth Subgraphs for Better Training and Evaluation of Knowledge Graph Augmented LLMs. arXiv:2511.04473, 2025.

Boyu Chen, Zirui Guo, et al. PathRAG: Pruning Graph-based Retrieval Augmented Generation with Relational Paths. In *AAAI*, 2026.

Darren Edge, Ha Trinh, Newman Cheng, et al. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. arXiv:2404.16130, 2024.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. RAGAS: Automated Evaluation of Retrieval Augmented Generation. In *EACL*, 2024.

Wenqi Fan et al. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In *KDD*, 2024.

Rylan Friel, Yichuan Yang, Kelly Marchisio, et al. RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems. arXiv:2407.11005, 2024.

Olaf Görlitz and Steffen Staab. SPARQL Endpoint Federation Exploiting VOID Descriptions. In *International Conference on Consuming Linked Data*, 2011.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. LightRAG: Simple and Fast Retrieval-Augmented Generation. In *Findings of the Association for Computational Linguistics: EMNLP*, 2025.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models. In *NeurIPS*, 2024.

Xiaoxin He, Yijun Tian, Yifei Sun, et al. G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering. In *NeurIPS*, 2024.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing A Multi-Hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. In *COLING*, 2020.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to Memory: Non-Parametric Continual Learning for Large Language Models. arXiv:2502.14802, 2025.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, et al. Dense Passage Retrieval for Open-Domain Question Answering. In *EMNLP*, 2020.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *SIGIR*, 2020.

Zixuan Li et al. StructRAG: Boosting Knowledge Intensive Reasoning of LLMs via Graph Knowledge Hybridization. arXiv:2410.08815, 2024.

Lei Liang, Mengshu Sun, Zhengke Gui, et al. KAG: Boosting LLMs in Professional Domains via Knowledge Augmented Generation. arXiv:2409.13731, 2024.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning. In *ICLR*, 2024.

Shengjie Ma, Chengjin Xu, Xuhui Jiang, et al. Think-on-Graph 2.0: Deep and Faithful Large Language Model Reasoning with Knowledge Graph-Guided Retrieval. In *ICLR*, 2025.

Vaibhav Mavi, Anubhav Jangra, and Adam Jatowt. Multi-hop Question Answering. *Foundations and Trends in Information Retrieval*, 2024.

Costas Mavromatis and George Karypis. GNN-RAG: Graph Neural Retrieval for Large Language Model Reasoning. arXiv:2405.20139, 2024.

Sewon Min, Eric Wallace, Sameer Singh, et al. Compositional Questions Do Not Necessitate Multi-hop Reasoning. In *ACL*, 2019.

OpenAI. GPT-4 Technical Report. arXiv:2303.08774, 2023.

Boci Peng, Yun Zhu, Yongchao Liu, et al. Graph Retrieval-Augmented Generation: A Survey. *ACM Transactions on Information Systems*, 2025.

Zack Rackauckas. RAG-Fusion: A New Take on Retrieval-Augmented Generation. *International Journal of Natural Language Computing*, 2024.

Stephen Robertson and Hugo Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 2009.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. In *NAACL*, 2024.

Bhaskarjit Sarmah, Dhagash Mehta, et al. HybridRAG: Integrating Knowledge Graphs and Vector Retrieval Augmented Generation for Efficient Information Extraction. In *ICAIF*, 2024.

Michael Schmidt, Olaf Görlitz, Peter Haase, Günter Ladwig, Andreas Schwarte, and Thanh Tran. FedBench: A Benchmark Suite for Federated Semantic Data Query Processing. In *International Semantic Web Conference*, 2011.

Andreas Schwarte, Peter Haase, Katja Hose, Ralf Schenkel, and Michael Schmidt. FedX: Optimization Techniques for Federated Query Processing on Linked Data. In *ISWC*, 2011.

Weijia Shi, Sewon Min, Michihiro Yasunaga, et al. REPLUG: Retrieval-Augmented Black-Box Language Models. In *ICML*, 2024.

Harry Shomer, Jay Revolinsky, and Jiliang Tang. Towards Better Benchmark Datasets for Inductive Knowledge Graph Completion. In *KDD*, 2025.

Fabian M. Suchanek, Serge Abiteboul, and Pierre Senellart. PARIS: Probabilistic Alignment of Relations, Instances, and Schema. *Proceedings of the VLDB Endowment*, 5(3), 2011.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, et al. Think-on-Graph: Deep and Responsible Reasoning of Large Language Model on Knowledge Graph. In *ICLR*, 2024.

Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu. Bootstrapping Entity Alignment with Knowledge Graph Embedding. In *IJCAI*, 2018.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *ICLR*, 2019.

Yixuan Tang and Yi Yang. MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries. arXiv:2401.15391, 2024.

Hugo Touvron et al. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971, 2023.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-hop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics*, 2022.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In *ACL*, 2023.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex Embeddings for Simple Link Prediction. In *ICML*, 2016.

Qikai Wei, Huansheng Ning, Chunlong Han, and Jianguo Ding. A Query-Aware Multi-Path Knowledge Graph Fusion Approach for Enhancing RAG in LLMs. *Expert Systems with Applications*, 2026.

Yilin Xiao et al. GraphRAG-Bench: Challenging Domain-Specific Reasoning in Graph-Based Retrieval-Augmented Generation. arXiv:2506.02404, 2025.

Weijian Xie and Xuefeng Liang. WeKnow-RAG: An Adaptive Approach for Retrieval-Augmented Generation Integrating Web Search and Knowledge Graphs. arXiv:2408.07611, 2024.

Kexuan Xin, Zequn Sun, Wen Hua, et al. Large-scale Entity Alignment via Knowledge Graph Merging, Partitioning and Embedding. In *CIKM*, 2022.

Peiru Yang, Xintian Li, Zhiyang Hu, et al. HeteRAG: A Heterogeneous Retrieval-Augmented Generation Framework with Decoupled Knowledge Representations. arXiv:2504.10529, 2025.

Xiao Yang, Kai Sun, et al. CRAG: Comprehensive RAG Benchmark. In *NeurIPS Datasets and Benchmarks Track*, 2024a.

Yaming Yang, Zhe Wang, Ziyu Guan, et al. Aligning Multiple Knowledge Graphs in a Single Pass. arXiv:2408.00662, 2024b.

Zhilin Yang, Peng Qi, Saizheng Zhang, et al. HotpotQA: A Dataset for Diverse, Explainable Multi-Hop Question Answering. In *EMNLP*, 2018.

Shunyu Yao, Jeffrey Zhao, Dian Yu, et al. ReAct: Synergizing Reasoning and Acting in Language Models. In *ICLR*, 2023.

Hang Yin, Zihao Wang, Weizhi Fei, and Yangqiu Song. EFOk-CQA: Towards Knowledge Graph Complex Query Answering beyond Set Operation. In *KDD*, 2025.

Hao Yu, Aoran Gan, Kai Zhang, et al. Evaluation of Retrieval-Augmented Generation: A Survey. arXiv:2405.07437, 2024.

Bowen Zhang and Harold Soh. Extract, Define, Canonicalize: An LLM-based Framework for Knowledge Graph Construction. In *EMNLP*, 2024.

Penghao Zhao, Hailin Zhang, Qinhan Yu, et al. Retrieval-Augmented Generation for AI-Generated Content: A Survey. arXiv:2402.19473, 2024a.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. Dense Text Retrieval Based on Pretrained Language Models: A Survey. *ACM Transactions on Information Systems*, 2024b.

Xiangrong Zhu, Yuexiang Xie, Yi Liu, Yaliang Li, and Wei Hu. Knowledge Graph-Guided Retrieval Augmented Generation. In *NAACL*, 2025.

Yuqi Zhu, Xiaohan Wang, Jing Chen, et al. LLMs for Knowledge Graph Construction and Reasoning: Recent Capabilities and Future Opportunities. *World Wide Web*, 2024.

A Implementation Details

Graph construction. All five domain KGs and the merged graph are built with LightRAG: entity and relation extraction with DeepSeek-V4-Flash (accessed via the `deepseek-chat` API endpoint, greedy decoding), domain-specific extraction prompts released with the code, and BGE-M3 embeddings (1024-dimensional, GPU) for the vector indices. Entities and relations are stored in Neo4j; vector and keyword indices in PostgreSQL.

Retrieval. LightRAG operates in hybrid mode (keyword-guided local traversal combined with global vector search); each query returns its top-5 entities per queried index (`top_k`=5). PG-Direct performs cosine-similarity retrieval over the same PostgreSQL vector indices without graph traversal. The Joint baseline uses cache size $k = 5$, entity selection ranked by name length, and the fixed domain order {biography, geography, arts, history_politics, general}.

Label solvability protocol. A question is *single-domain solvable* if there exists one domain graph in which every gold-chain entity can be matched (fuzzy substring matching for MuSiQue surface variants) and consecutive chain entities are connected within h hops under the same directed Neo4j traversal used for mining (§3.2). Applying this check to the 2,504 MuSiQue degree-2 questions over our five graphs finds 96% solvable. The statistic characterizes MuSiQue labels *under our specific graph construction*; it does not assess the correctness of MuSiQue's own composition annotations, and we do not extrapolate it to other datasets.

Evaluation. Retrieved entities are parsed from the returned context, and metrics are computed per question against gold chains: exact name equality on MultiRAG-Bench (gold chains share the graphs' node vocabulary) and fuzzy substring matching on MuSiQue. The released evaluation code additionally reports Wilson 95% confidence intervals for all proportion-valued metrics, per-method result-set sizes $|\mathcal{R}(Q)|$, per-tier bridge selection rules, and the ablation suite (no-cache union control, domain-order permutations, cache-ranking variants, exact/fuzzy matcher sensitivity on both datasets).

B Union Control and Sensitivity Ablations

Table 7 reports the union-control ablation on the full 2,348-question benchmark under a single shared keyword extractor (cached DeepSeek-V4-Flash extractions where available, a local gemma2-9B model otherwise; every method in the table uses the same extractor, so internal contrasts are unaffected by this substitution, though absolute values shift slightly relative to Table 4). **UnionSingle** queries all five domains independently with the original question and pools the results (no cache, no enrichment); **UnionSingle@25** is the same pooled set randomly subsampled to a 25-entity budget. All confidence intervals are Wilson 95%; contrast p-values use a paired bootstrap (10^4 resamples).

Table 7: Union-control ablation (LightRAG hybrid backend; shared keyword extractor; exact matching; all values %). Joint−Union contrasts: PHR +1.87pp ($p = 0.10$), BR +2.34pp ($p = 0.02$) overall; −3.06pp PHR / +2.04pp BR at the exploratory 3-hop tier. BestSingle's per-tier BR coincides exactly with the union control's: the bridge is a single entity, so the per-domain maximum of its recovery indicator equals the pooled indicator (its PHR remains strictly lower, since the full chain spans domains). UnionSingle@25 is reported overall only. Absolute PHR values differ slightly from Table 4 because of the shared-extractor substitution described in the text.

	Overall		**1-hop**		**2-hop**		**3-hop**	
Method	PHR	BR	PHR	BR	PHR	BR	PHR	BR
MeanSingle	6.15	44.78	11.89	16.37	2.65	57.89	15.71	61.43
BestSingle	27.34	77.68	54.53	62.27	12.53	84.80	45.92	86.73
UnionSingle	54.64	77.68	59.07	62.27	51.53	84.80	68.37	86.73
UnionSingle@25	23.08	54.47	—	—	—	—	—	—
Joint	56.52	80.03	62.67	66.13	52.87	86.40	65.31	88.78

Three observations follow. First, the union control closes most of the gap between single-domain baselines and Joint: pooling five domain result sets (each $|\mathcal{R}| \approx 8$; pooled ≈ 39) yields 54.64% PHR versus Joint's 56.52%, so cache propagation contributes +1.87pp PHR overall, a small increment

concentrated at 1-hop and not significant at 2-hop (+1.33pp, $p = 0.24$), while its bridge-recovery increment (+2.34pp BR, $p = 0.02$) is significant. Second, retrieved-set size is a first-order confound for both PHR and SNR: budget-matching the union to 25 entities halves PHR (54.64% $\rightarrow$ 23.08%), and Joint's low SNR in Table 4 partly reflects its larger pooled result set. Third, repeating all metrics under fuzzy substring matching leaves the ordering and gaps essentially unchanged (UnionSingle 71.76% vs. Joint 73.42% PHR), so the conclusions are not artifacts of the matching protocol.

Domain-order and cache-ranking sensitivity. On a 249-question stratified subsample (83 per hop tier) under the same extractor, replacing the fixed domain order with a random permutation changes PHR by -0.80pp (60.64% $\rightarrow$ 59.84%, $p = 0.44$) and BR by $+1.21$pp, and replacing the name-length cache ranking with random selection changes PHR by -0.40pp (60.24%, $p = 0.48$) and BR by -1.60pp. The joint pipeline is thus insensitive to these design choices, consistent with the union-control finding that pooling, rather than the enrichment details, drives performance.

Question audit. Two audits accompany the released dataset. *Entity leakage* (all 2,348 questions): 89.3% of questions contain a normalized surface form of at least one gold-chain entity. Most of this is inherent to multi-hop question composition: the question must name its topic entity (1-hop: 76.4%; 2-hop slot a: 82.6%), exactly as hand-written multi-hop datasets do. The 2-hop bridge, however, is additionally named in 50.6% of questions, while the 2-hop answer entity is rarely leaked (12.5%). Leaked entities act as free keyword anchors, so PHR and BR should be read as optimistic; the path-level Neo4j guarantee (§3.2) is topological and unaffected, and recovering the non-leaked answer entities still requires genuine retrieval. *LLM-judged quality* (qwen2.5:32b judge, 150-question stratified sample, temperature 0): mean naturalness 3.2/5 (no question judged below 2); 68% judged answerable from the gold chain and, on the independent ambiguity axis, 33% judged ambiguous. Both audits and per-question judgments ship with the dataset; a human audit at scale remains future work.

Leakage-stratified robustness. Because leaked entities act as free anchors, we recompute the 2-hop stratum (the tier where BR is defined) split by whether the bridge entity is named in the question: 759 leaked vs. 741 clean questions. On the clean subset, pooled retrieval still roughly triples the best single-domain PHR (UnionSingle 43.7% and Joint 44.0% vs. BestSingle 9.9% and MeanSingle 2.0%), so the benchmark's central cross-domain gap is not a leakage artifact. The cache increment, however, shrinks to $+0.27$pp PHR on clean bridges (vs. $+2.37$pp on leaked bridges, where the leaked bridge name directly enriches the enriched-query anchor set), reinforcing the pooling-first reading of §4.3. Stratified metrics for all methods and tiers are included in the released per-question logs.

Alignment-informed fusion. To test whether the fusion cost is an artifact of naive name matching, we rebuild the fused graph with embedding-similarity entity alignment: entities are clustered by cosine similarity of their name+description embeddings ($\tau = 0.90$, BGE-M3), merging 186,261 $\rightarrow$ 155,959 entities (an additional 16.3% reduction beyond exact-name matching), with all 292,788 remapped relations resolving to canonical entities. Under the shared extractor, this *degrades* fused retrieval: PHR falls from 34.28% [32.4, 36.2] to 19.85% [18.3, 21.5] and SNR from 16.2% to 13.2%, moving the Merge Penalty from -6.94 (naive, merging helps) to $+7.50$ (aligned; bootstrap $p < 0.001$ against BestSingle). This is consistent with hub-noise amplification (stronger alignment merges more variant entities into shared hubs, reflected in the SNR drop), though a false-merge audit would be required to separate hub growth from topological corruption of gold chains. Either way, the practical conclusion holds: the fusion-vs-joint trade-off does not depend on the deduplication strategy being weak, and if anything worsens as alignment strengthens. Threshold sweeps, hub-degree statistics for the aligned graph, and a merge-error analysis ship with the released construction code.

C Dataset Characterization

Table 8 reports per-domain graph statistics. Entity counts are stable across domains (113k–120k), while relation counts vary substantially from 135,853 (arts) to 207,945 (biography), confirming that domain-specific prompts produce structurally different graphs. The mean pairwise relation Jaccard across domains is 0.057 (range 0.044–0.066): different prompts extract largely disjoint relation sets from the same corpus. In contrast, mean pairwise entity Jaccard is 0.63: domains share nearly two-thirds of named entities but describe them with substantially different relations. This high-entity,

low-relation asymmetry is the structural basis for cross-domain retrieval: a bridge entity exists in multiple domain graphs, but its connecting edges are domain-specific, so no single graph captures the full cross-domain evidence chain. Throughout, relation Jaccard is computed over (head, relation type, tail) triples and entity Jaccard over entity identities; restricting the five domain graphs to the 1,000-document subset used by the model study leaves the prompt-induced relation Jaccard essentially unchanged ($0.057 \rightarrow 0.063$), confirming the prompt-vs-model comparison is not a corpus-size artifact.

Table 8: Domain knowledge graph statistics.

Domain	Entities	Relations	Degree
biography	115,306	207,945	3.61
geography	117,614	148,259	2.52
arts	113,698	135,853	2.39
history_politics	113,052	138,460	2.45
general	119,697	145,725	2.43

For the fusion baseline, we merge all five graphs via case-insensitive exact-name matching (580k $\rightarrow$ 186k entities, 67.9% removed; 776k $\rightarrow$ 347k relations, 55.3% removed). The merged graph exhibits substantial hub inflation: the number of hub nodes with degree ≥ 100 grows from 47–101 per domain to 229 in the merged graph (a 2.3×–4.9× increase in hub density; the maximum node degree grows 1.7×, from 1,946 to 3,341). This increase in hub density is a structural consequence of fusion that contributes to the low signal-to-noise ratios observed across all retrieval methods (§4).

Table 9 reports each domain's focus ratio and top-3 relation types. Focus ratios range from 5.3% (history_politics) to 18.4% (geography), indicating that domain prompts bias extraction rather than strictly constrain it, and the realized top-3 relations diverge from the declared priorities in Table 2: the extraction realizes a soft bias rather than the declared schema. The relation `located_in` appears in all five domains' top-3 due to Wikipedia's inherent geographic bias, but secondary preferences are domain-specific. This partial-but-measurable differentiation is the practical regime where the fusion-vs-joint question is non-trivial.

Table 9: Domain focus: top-3 relation types and focus ratio. Focus ratio is the fraction of a domain's relations falling under its declared priority types (Table 2); it is undefined (−−−) for general, which declares no priorities.

Domain	Focus	Top-3 relation types
biography	13.6%	located_in, member_of, starred_in
geography	18.4%	located_in, part_of, contains
arts	12.4%	located_in, stars, portrays
history_politics	5.3%	located_in, member_of, part_of
general	—	located_in, part_of, starred_in