BRIDGING THE COMPLETENESS GAP IN GRAPHRAG: A UNIFIED GNN ENHANCEMENT LAYER FOR RETRIEVAL AND GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

GraphRAG retrieves evidence from knowledge graphs automatically constructed from text, yet dense graphs can still omit the relations that downstream queries need: on GraphRAG-Bench Medical, a graph with an average degree of 6.86 covers only 14.5% of annotated evidence triples, and in 61.0% of cases the relevant entities are present but the relation between them is missing. We introduce a lightweight query-adaptive GNN layer that predicts missing relations, estimates their confidence and query relevance, and selects them for structural augmentation and verbal hints. A structural path inserts query-adaptively filtered relations into the graph before Personalized PageRank (PPR) retrieval, while a verbal path passes top-ranked relations to the language model as concise hints, exploiting relational knowledge without propagating noisy edges through the graph. Filtering is decisive: inserting all predicted edges lowers Medical accuracy by 5.23 points, whereas a learned query-adaptive gate turns this into a 3.47-point gain. The verbal path adds 4.95 points, and the two paths together gain 7.94 and 5.54 points on Medical and Novel, respectively. The layer improves six retrieval-system configurations across the evaluated benchmark subsets and three language-model backbones on Medical, supporting selective structural and verbal augmentation as a practical way to address missing connections in this setting.

1 INTRODUCTION

Retrieval-augmented generation connects language models to external evidence, making retrieval a central part of answering knowledge-intensive questions (Lewis et al., 2020; Karpukhin et al., 2020; Izacard & Grave, 2021). GraphRAG extends this interface by organizing a corpus into entities and relations. Community-based summaries expose global themes, while graph traversal connects evidence distributed across passages (Edge et al., 2024; Gutiérrez et al., 2024; 2025; Guo et al., 2025). These capabilities depend on what survives graph construction. If a relation required by a question is absent, retrieving its endpoint entities does not necessarily recover the connection needed to answer.

We study this gap between graph connectivity and evidence completeness. On the Medical subset of GraphRAG-Bench (Xiang et al., 2026), the constructed graph contains 15,789 edges, yet covers only 14.5% of annotated evidence triples. Approximately 61% of evidence occurrences have both entities represented but lack their connecting relation. The dominant failure in this setting therefore lies between existing entities. This diagnosis identifies a concrete target for enhancement: recover question-relevant connections that the extraction pipeline omitted.

Recovering a plausible connection is only the first step. An edge inserted before retrieval changes the routes through which relevance propagates; a mistaken shortcut can promote irrelevant passages. In our Medical experiment, inserting all predicted edges reduces the answer score by 5.23 points. A relation introduced after retrieval has a different effect: it becomes a contextual cue that the generator can interpret alongside retrieved passages. These interventions impose different demands on the same prediction. The central design problem is to determine both *which relations matter for a question* and *how those relations should be used*.

We develop a unified GNN enhancement layer around this distinction. A shared graph attention network predicts candidate relations and supports confidence estimation, query relevance, and query-

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.

adaptive gating (QAG). The structural path uses the gate to select additions to the retrieval graph. The verbal path ranks predicted relations and presents a compact set of hints to the frozen answer generator. The paths share representations but make separate selection decisions: the structural gate controls graph propagation, while verbal ranking selects the relations the generator sees. This division connects relation completion to the two stages at which missing evidence affects an answer.

Our evaluation connects the completeness diagnosis to downstream answer quality. QAG turns the negative effect of unfiltered insertion into a 3.47-point gain on Medical. Verbal augmentation gains 4.95 points, and the combined configuration gains 7.94 and 5.54 points on Medical and Novel. The enhancement also improves six retrieval-system configurations and three generator backbones. Together, these results show that recovering relations and selecting their downstream use are distinct requirements for improving GraphRAG.

We make three contributions. First, question-specific evidence alignment identifies missing relations between existing entities as a major source of incompleteness in the evaluated graphs. Second, a shared GNN recovers candidate relations and selects them for structural and verbal augmentation. Third, controlled experiments separate prediction quality from downstream selection and demonstrate the benefit of combining the two paths across the evaluated systems and generators.

2 Diagnosing the Completeness Gap

2.1 Question-Relevant Evidence Completeness

Let $\mathcal{D}$ be a text corpus and $G = (V, E)$ its extracted knowledge graph. Each entity is associated with text features and source passages. Given a question q, a graph retriever selects passages from $\mathcal{D}$, and an LLM generates an answer from the question and those passages. We consider predictions whose endpoints already belong to V; entity discovery is outside the scope of the enhancement layer.

To measure the relations preserved by G, we align annotated evidence triples with its entities and edges. For each question, let $\mathcal{T}_q$ contain its evidence-triple occurrences. An occurrence is *full* when both endpoints and the relation align with the graph, *partial* when the endpoints align but their relation is absent, and an *entity failure* when an endpoint cannot be correctly aligned. Let $M_G(t) \in \{0, 1\}$ indicate whether the full triple matches the graph. Coverage is

$$\text{Coverage}(G) = \frac{\sum_q \sum_{t \in \mathcal{T}_q} M_G(t)}{\sum_q |\mathcal{T}_q|}. \tag{1}$$

This measure evaluates the preservation of annotated evidence rather than the number of extracted edges. Entity collision is measured separately because it concerns whether a graph identifier refers to the intended entity. Appendix A summarizes the matching categories and reports the diagnostic counts.

Table 1: Graph statistics and question-relevant evidence. Edge precision is measured on 400 sampled edges per domain. Coverage and partial evidence measure relations; collision measures entity alignment.

	Medical	Novel
Questions in the study subset	509	609
Entities / edges	4,605 / 15,789	2,038 / 4,102
Sampled edge precision	85.8%	82.3%
Triple coverage	14.5%	43.2%
Partial evidence	≈61%	38.4%
Entity collision	0.0%	9.5%

2.2 Different Gaps Require Different Interventions

Table 1 shows why edge precision alone is insufficient to assess graph quality for question answering. Most sampled Medical edges are correct, but few annotated evidence relations are represented. Among 1,367 Medical evidence occurrences, 198 are full and 833 are partial. Predicting relations between

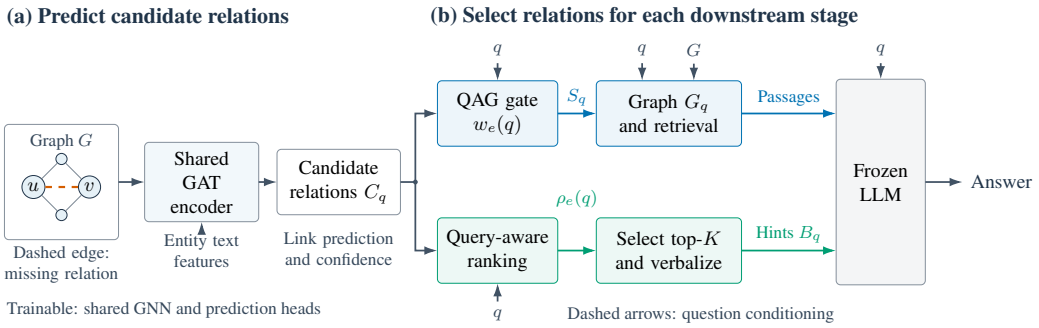


Figure 1: Shared relation prediction with separate structural and verbal selection. The structural path adds S_q to the original graph G and retrieves passages from G_q; the verbal path ranks candidates by $\rho_e(q)$ and converts the selected top-K relations H_q into hints B_q. Passages and hints enter the frozen generator through separate inputs. Repeated q symbols denote the same question. The dashed edge in (a) marks a missing relation; dashed arrows denote question conditioning.

existing entities directly targets this partial-evidence population. Novel has higher coverage but also exhibits entity collisions, making endpoint disambiguation important before predicted relations can be used.

The diagnosis determines where augmentation can help, while downstream selection determines whether it does help. A missing relation may connect a query entity to useful evidence, but a plausible prediction may also connect it to an irrelevant region. The structural and verbal paths address these consequences separately. Figure 1 summarizes how shared predictions enter the two stages.

3 A Unified GNN Enhancement Layer

3.1 Shared Representations and Prediction Heads

The encoder combines entity text with graph neighborhoods to represent candidate connections. We use a two-layer graph attention network (Veličković et al., 2018), with hidden dimension 128, four attention heads, and dropout 0.3. For entity v, let h_v be its encoded representation. A candidate $e = (u, r, v)$ receives an edge representation $z_e = \phi(h_u, r, h_v)$, while z_q represents the question. Entity features are initialized with the text embeddings produced during graph construction. The prediction heads then use the encoded endpoints to score connections between entities already in the graph.

Four heads assign different roles to these shared representations. Link prediction scores the plausibility of a relation, and confidence scoring estimates its quality. Query relevance and gating additionally take the question as input:

$$p_e = f_{\mathrm{LP}}(z_e), \qquad c_e = f_{\mathrm{conf}}(z_e), \tag{2}$$

$$s_e(q) = f_{\mathrm{rel}}([z_e \| z_q]), \qquad w_e(q) = \sigma(f_{\mathrm{gate}}([z_e \| z_q]))\,. \tag{3}$$

The distinction is functional: plausibility concerns whether a connection holds, whereas query-conditioned scores concern its usefulness for the current question. The gate is a $256 \to 64 \to 1$ MLP with a sigmoid output. It adds approximately 16K parameters to the 229K-parameter base model.

We jointly optimize the encoder and heads with

$$\mathcal{L} = \mathcal{L}_{\mathrm{LP}} + \mathcal{L}_{\mathrm{conf}} + \mathcal{L}_{\mathrm{rel}} + 0.5\mathcal{L}_{\mathrm{QAG}}. \tag{4}$$

Training evidence provides positive supervision for relation prediction, and the query-conditioned objectives connect candidate representations to their downstream selection roles. Joint training shares information across the four tasks; it does not require gradients through graph retrieval or the frozen generator. The loss weights balance relation prediction and quality estimation with the two query-conditioned tasks. Appendix B summarizes the architecture, optimization settings, and inference configurations.

3.2 Structural Augmentation for Evidence Discovery

Structural augmentation changes which passages the retriever can reach. For the candidate pool C_q, the gate selects

$$S_q = \{e \in C_q : w_e(q) \geq t_g\}, \qquad G_q = (V, E \cup S_q), \tag{5}$$

where the cutoff t_g is selected on validation data. The query dependence lies in $w_e(q)$: the same candidate can receive different scores for different questions. This allows selection to focus on the connections needed by the current retrieval task.

Our reference implementation applies personalized PageRank (PPR) (Haveliwala, 2002) to G_q. Let P_q be its transition matrix and v_q the seed distribution over query-linked entities. Retrieval uses the stationary relevance vector

$$\pi_q = \alpha v_q + (1 - \alpha) P_q^\top \pi_q, \tag{6}$$

where α is the restart probability. Selected edges alter P_q, allowing relevance to reach evidence through connections absent from the original graph. The same mechanism explains why indiscriminate insertion is risky: erroneous edges also change the propagation routes. Query-adaptive selection controls this intervention before passage retrieval.

3.3 Verbal Augmentation for Evidence Use

Verbal augmentation exposes predicted relations directly to the generator. For a query-conditioned ranking score $\rho_e(q)$, we select

$$H_q = \mathrm{TopK}_{e \in C_q} \rho_e(q), \qquad B_q = \{\mathrm{verbalize}(e) : e \in H_q\}. \tag{7}$$

We use $\rho_e(q) = s_e(q)$ for the query-relevance verbal configuration and $\rho_e(q) = w_e(q)$ for the QAG-ranked variant and the combined QAG configuration. We compare these ranking functions separately in Section 5. Each selected edge is verbalized as "A is related to B." This generic template supplies a connection for the LLM to interpret using retrieved evidence, without requiring the prompt to assert a predicted relation type.

Let $R(G, q)$ denote retrieved passages and F the frozen generator. Structural-only augmentation produces $F(q, R(G_q, q))$, verbal-only augmentation produces $F(q, R(G, q), B_q)$, and the combined configuration produces

$$\hat{a}_q = F(q, R(G_q, q), B_q). \tag{8}$$

The structural path changes the evidence discovered; the verbal path adds explicit relational cues to that evidence. Their candidate subsets need not coincide. This separation lets the two selection rules serve different stages while reusing the same learned representations.

3.4 Entity Identity and System Integration

Predicted relations are useful only when their endpoints refer to the intended entities. On Novel, repeated names across documents create entity collisions. We use document-qualified identifiers and same-document constraints, followed by anchor expansion and an edge-reward term. The identifiers and document constraint distinguish namesakes before augmentation; the additional expansion and reward stages improve the resulting candidate configuration. Appendix F reports the cumulative ablation, including the edge-reward coefficient $\lambda = 0.5$.

For integration with a retrieval system, structural augmentation is applied before evidence retrieval and verbal hints are appended to its generation context. The generator remains frozen. We evaluate a reranked text-RAG baseline and five GraphRAG systems: HippoRAG2 (Gutiérrez et al., 2025), G-reasoner (Luo et al., 2026), LightRAG (Guo et al., 2025), MS-GraphRAG (Edge et al., 2024), and Fast-GraphRAG (Circlemind, n.d.). PPR-based graph enhancement precedes each system's retrieval stage, and verbalized edges are appended to its generation prompt (Appendix G). This interface separates the shared relation model from each system's retrieval policy.

Table 2: Answer quality on all 509 Medical and 609 Novel questions in the study subsets. Gains are ACC points relative to the corresponding baseline. Bold marks the best score in each domain.

	Medical		Novel	
Configuration	ACC	Gain	ACC	Gain
RAG + rerank	.3917	—	.3833	—
Raw predicted edges into PPR	.3394	−5.23	.3845	+0.12
Fixed-confidence selection	.4198	+2.81	.4057	+2.24
QAG structural path	.4264	+3.47	.4106	+2.73
Query-relevance verbal path	.4412	+4.95	.4138	+3.05
Existing-edge hints	.3978	+0.61	.3870	+0.37
Entity-pair BM25 sentences	.4142	+2.25	.3997	+1.64
Random-edge hints	.3864	−0.53	.3854	+0.21
Dual path (QAG)	**.4711**	**+7.94**	**.4387**	**+5.54**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We use subsets of the Medical and Novel domains of GraphRAG-Bench (Xiang et al., 2026), containing 509 and 609 questions, respectively. The main results and cross-system comparisons evaluate answer quality on all questions in these study subsets. Evidence annotations support graph-coverage analysis and GNN supervision. For GNN fitting, questions are partitioned into 60% training, 20% validation, and 20% testing; training-partition evidence triples provide supervision, and validation data support model selection. Appendix C specifies the distinction between the GNN partition and the complete-subset answer evaluation. Unless otherwise specified, the RAG baseline refers to the RAG + rerank configuration.

We report answer correctness (ACC) on a 0–1 scale as the primary measure of answer quality. An ACC-point gain is $100(\mathrm{ACC}_{\mathrm{method}} - \mathrm{ACC}_{\mathrm{base}})$. Link prediction is evaluated with Hits@10 and mean reciprocal rank (MRR); ROUGE-L (Lin, 2004) measures lexical overlap. We report 95% bootstrap confidence intervals for the primary comparison using 10,000 resamples (Appendix C). All numerical comparisons below use the same units for gains.

The downstream comparisons separate three decisions: which candidate relations are predicted, which predictions are selected, and where they enter the pipeline. Structural controls insert all candidates, apply fixed-confidence filtering, or use QAG. Verbal controls add random edges, existing graph edges, or BM25-retrieved source sentences mentioning the predicted entity pairs. We also compare link predictors, ranking functions, and verbalization templates. Appendix C.1 details the roles of these controls, and Appendix E reports the hint-budget sweep.

We train the GNN with Adam (Kingma & Ba, 2015) at learning rate 10^{-3} for up to 200 epochs, with validation-loss early stopping. The complete model has approximately 245K trainable parameters. We use $K = 15$ verbal hints for Medical and $K = 20$ for Novel. Qwen-Plus is the default generator; GPT-4o and Llama-3-70B-Instruct test the combined enhancement on Medical. All cross-system comparisons use Qwen-Plus so that the generator is held fixed across retrieval systems. Appendix B summarizes the implementation settings.

4.2 MAIN RESULTS

Table 2 shows that predicted relations improve answer quality when their downstream use is controlled. On Medical, raw insertion reduces ACC from .3917 to .3394, whereas QAG-filtered insertion raises it to .4264. The gain from selection is therefore larger than the gain from adding edges alone: QAG improves by 8.70 points over raw insertion and by 3.47 points over the unaugmented baseline. On Novel, raw insertion is nearly neutral, while QAG yields a 2.73-point gain. The need for filtering is strongest in Medical, but selective insertion benefits both domains.

Verbal augmentation improves Medical and Novel by 4.95 and 3.05 points. Both exceed the gains from existing-edge hints and random-edge hints. Entity-pair BM25 is a stronger control, but predicted

Table 3: Dual-path augmentation across retrieval systems. Each base score is reproduced with the shared experimental setup. Full single-path results and an overview of system integration appear in Appendix G.

	Medical			Novel		
System	Base	Dual	Gain	Base	Dual	Gain
RAG + rerank	.3917	.4711	+7.94	.3833	.4387	+5.54
HippoRAG2	.5124	.5756	+6.32	.4828	.5247	+4.19
G-reasoner	.5067	.5612	+5.45	.4951	.5319	+3.68
LightRAG	.4893	.5498	+6.05	.4609	.5017	+4.08
MS-GraphRAG	.4571	.5152	+5.81	.4312	.4734	+4.22
Fast-GraphRAG	.4738	.5308	+5.70	.4537	.4951	+4.14

relational hints improve by a further 2.70 and 1.41 points. The advantage over existing and random edges shows that adding relational text alone does not explain the gain. The comparison with entity-pair BM25 further shows that selected predicted connections outperform these source-sentence additions.

Combining the paths gives the highest scores on both domains. Dual-path augmentation exceeds structural-only by 4.47 and 2.81 points, and verbal-only by 2.99 and 2.49 points. The Medical ranking ablation also evaluates a stronger QAG-ranked verbal variant at .4452 ACC; the combined configuration exceeds it by 2.59 points. The combined gain therefore persists when compared with the stronger verbal ranking rule.

4.3 Integration Across Retrieval Systems and Generators

Across the complete study subsets, the benefit extends to the six retrieval-system configurations in Table 3. Medical gains range from 5.45 to 7.94 points, and Novel gains range from 3.68 to 5.54 points. The evaluation uses a shared benchmark graph and Qwen-Plus generation, allowing each adapter to reuse the trained enhancement model. The positive gains across these configurations show that the intervention is useful beyond the reference retrieval pipeline.

The combined enhancement also improves all three evaluated generators on Medical. Qwen-Plus rises from .3917 to .4711, GPT-4o from .4673 to .5327, and Llama-3-70B-Instruct from .4198 to .4892. The gains of 7.94, 6.54, and 6.94 points indicate that the enhancement remains useful across these generation backbones without LLM fine-tuning.

5 Analysis

5.1 Selection Connects Prediction to Downstream Utility

The link-prediction comparison establishes the quality of the candidate model. GAT+QAG reaches .841 Hits@10 and .527 MRR on Medical, and .925 and .691 on Novel. Text-initialized TransE improves over its randomly initialized counterpart, while the text-aware GNNs achieve higher ranking scores. Adding QAG changes GAT Hits@10 by only .005 on Medical and .004 on Novel. Its main role is therefore downstream selection, supported by a strong shared relation representation. Appendix D reports the full comparison with TransE, RotatE, ComplEx, GIN, and GAT.

Figure 2(a) shows why the selection rule matters. The best fixed-confidence cutoffs differ between Medical and Novel: .85 and .80, respectively. QAG improves over these fixed-cutoff configurations by .66 and .49 points while retaining approximately one quarter of candidates. Retaining more edges does not consistently improve answer quality. The useful subset depends on the query as well as the plausibility of each relation.

The same distinction appears in verbal ranking (Figure 2(b)). Ranking by link-prediction score gives a 1.70-point Medical gain; query–endpoint cosine similarity raises it to 3.62; learned query relevance reaches 4.95; and gate-based ranking reaches 5.35. These comparisons separate the ability to predict a plausible connection from the ability to identify one that helps answer a particular question.

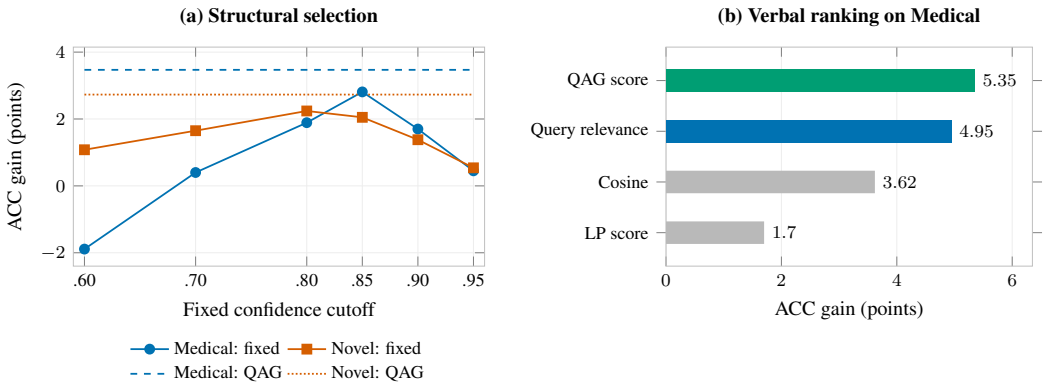


Figure 2: Selection ablations. (a) Fixed-confidence filtering compared with QAG; horizontal reference lines show the QAG configuration, not additional cutoff evaluations. (b) Query-conditioned scores improve verbal ranking on Medical ($K = 15$). The QAG-ranking variant and relevance-ranking variant are evaluated separately.

5.2 The Paths Improve Different Parts of the Pipeline

The retrieval and fixed-context experiments clarify the roles of the two paths. On Medical, structural augmentation raises Recall@5 from .412 to .463. Verbal augmentation leaves retrieval unchanged and raises answer quality with gold passages from .618 to .659. With gold passages fixed, the structural path has no retrieval decision to change. These controls isolate where each intervention acts, while the end-to-end experiment tests their combined effect. Appendix H reports the stage-level results. This stage-level separation is consistent with the complementary roles of the two interventions in the combined configuration.

Hint-adoption annotations provide a second view of generation behavior. Across 1,500 hints for 100 Medical questions, two annotators achieve Cohen's $\kappa = .81$. Correct, relevant hints are adopted in 87.2% of cases, compared with 5.1% for correct but irrelevant hints. Incorrect hints are adopted in 26.4% of cases. The generator responds selectively to the supplied relations, but incorrect predictions can still enter the answer. This motivates selecting hints for relevance and preserving access to source evidence. Appendix H.1 describes the annotation setting and reports the full adoption table.

5.3 Supervision, Prompt Design, and Cost

The enhancement can also be trained from existing graph edges. This self-supervised variant reaches .793 Medical Hits@10 and a 5.87-point dual-path gain, compared with .841 and 7.94 for evidence-supervised training. The comparison shows that annotated evidence improves performance, while the graph itself supplies a useful training signal. The downstream gain remains positive when training relations come from the constructed graph, supporting the use of graph-derived supervision when evidence annotations are unavailable.

Verbalization and hint budget also influence performance. The generic relation template gains 4.95 points on Medical, compared with 3.92 for typed relations and 3.24 for hints containing confidence values. The best tested hint budgets are 15 for Medical and 20 for Novel; larger budgets reduce gains. These results favor concise, selected cues over increasingly detailed or numerous predictions. Appendix E reports the complete sweeps and ROUGE-L results.

The combined layer adds approximately 0.4 seconds per query and increases prompt tokens by 16.0% in our timing setup. GNN inference takes less than 0.1 seconds; structural augmentation accounts for approximately 0.3 seconds. These costs make the trade-off explicit: the model adds a small trainable component and extra context while keeping the generator frozen. Appendix I separates the per-query cost of GNN inference, graph augmentation, and verbalization.

6 Related Work

Retrieval-augmented generation couples document search with answer generation (Lewis et al., 2020; Karpukhin et al., 2020; Izacard & Grave, 2021). REPLUG supports retrieval with a black-box language model, while Self-RAG and corrective retrieval use additional signals to decide how retrieved material should enter generation (Shi et al., 2024; Asai et al., 2024; Yan et al., 2024). Our intervention operates on a structured evidence interface: it predicts missing graph connections and chooses their structural and verbal use.

GraphRAG systems offer several ways to organize that interface. Community summaries support global questions (Edge et al., 2024); HippoRAG and HippoRAG2 use graph-based memory retrieval (Gutiérrez et al., 2024; 2025); and LightRAG combines graph and vector retrieval (Guo et al., 2025). GraphRAG-Bench evaluates how graph construction, retrieval, and generation interact (Xiang et al., 2026). Our evidence-alignment diagnosis focuses on relations required by individual questions and motivates an augmentation stage after graph construction.

Knowledge graph completion scores missing facts through entity and relation representations. TransE, ComplEx, and RotatE provide embedding-based scoring functions (Bordes et al., 2013; Trouillon et al., 2016; Sun et al., 2019). R-GCN and CompGCN incorporate relational structure through message passing (Schlichtkrull et al., 2018; Vashishth et al., 2020), while GIN and GAT provide general neighborhood encoders (Xu et al., 2019; Veličković et al., 2018). We use a text-aware GAT as the shared predictor and evaluate candidate ranking separately from downstream augmentation. The selection heads address a different objective from link plausibility: which predicted connections should influence a given answer pipeline.

Graph-based question answering also uses learned retrieval to provide context to an LLM. GNN-RAG uses GNN retrieval to identify knowledge-graph evidence for LLM reasoning (Mavromatis & Karypis, 2025). G-Retriever retrieves subgraphs from textual graphs (He et al., 2024), while SubgraphRAG scores triples with a lightweight MLP and directional structural features (Li et al., 2025). KAPING and KG-GPT use graph knowledge in language-model prompts (Baek et al., 2023; Kim et al., 2023). Our layer predicts connections absent from the extracted graph and selects their use before retrieval and after retrieval. This placement is central to the method: the same candidate can have different consequences as a graph edge and as a contextual hint.

7 Limitations and Conclusion

Our evaluation uses subsets from two domains of one benchmark and a shared graph-construction setting. Different extraction pipelines may change the balance between missing relations and entity failures. The main and cross-system answer evaluations cover the complete study subsets, including the GNN training portion; generalization to independent questions remains to be evaluated. Replacing annotated training relations with graph-derived supervision reduces the Medical gain from 7.94 to 5.87 points, exposing the cost of removing evidence supervision. Finally, the observed adoption of incorrect hints identifies source grounding and provenance as requirements for deploying predicted relations.

We identify missing question-relevant relations as a target for improving GraphRAG and develop a unified GNN layer to recover and use those connections. Query-adaptive structural selection improves evidence retrieval, while ranked verbal hints provide additional relational context to the generator. Their combination gives the strongest answer scores in our comparisons and improves the evaluated retrieval systems and LLM backbones. The resulting design principle is to connect relation prediction with decisions about its downstream use: an augmentation layer must determine which connection is useful, for which question, and at which stage.

Reproducibility Statement

The graph-coverage definition and selection rules appear in Sections 2–3. Appendix B summarizes architecture and optimization, Appendix C outlines evaluation, and the remaining appendices provide detailed ablations and system configurations. The main tables specify the evaluated configurations

and baselines. The appendices include all selection sweeps, hint budgets, full cross-system results, and generator comparisons needed to trace the reported gains.

ETHICS STATEMENT

Predicted relations can introduce unsupported information into generated answers. We measure answer-level adoption of incorrect hints and distinguish predictions from source-grounded evidence. The Medical subset is used as a research benchmark; the system is not intended for clinical decision-making. Deployments should preserve evidence provenance and account for errors in both graph construction and augmentation.

AI USE STATEMENT

Generative AI tools assisted with Chinese-to-English translation, manuscript drafting and editing, interpretation and presentation of experimental results, literature search and reference formatting, and preparation of figures and tables. The experimental results originate from the authors' experiments. The authors take responsibility for the technical content, citations, figures, and final manuscript.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *International Conference on Learning Representations*, pp. 9112–9141, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/25f7be9694d7b32d5cc670927b8091e1-Paper-Conference.pdf.

Jinheon Baek, Alham Fikri Aji, and Amir Saffari. Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering. In *Proceedings of the First Workshop on Matching From Unstructured and Structured Data (MATCHING 2023)*, pp. 70–98. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.matching-1.7. URL https://aclanthology.org/2023.matching-1.7/.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf.

Circlemind. Fast GraphRAG. Software repository, n.d. URL https://github.com/circlemind-ai/fast-graphrag. Accessed September 26, 2026.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv preprint arXiv:2404.16130*, 2024. doi: 10.48550/arXiv.2404.16130. URL https://arxiv.org/abs/2404.16130.

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*, pp. 10746–10761. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.568. URL https://aclanthology.org/2025.findings-emnlp.568/.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 59532–59569. Curran Associates, Inc., 2024. doi: 10.52202/079017-1902. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/6ddc001d07ca4f319af96a3024f6dbd1-Paper-Conference.pdf.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to Memory: Non-Parametric Continual Learning for Large Language Models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21497–21515. PMLR, 2025. URL https://proceedings.mlr.press/v267/gutierrez25a.html.

Taher H. Haveliwala. Topic-sensitive PageRank. In *Proceedings of the 11th international conference on World Wide Web*, WWW02, pp. 517–526. ACM, 2002. doi: 10.1145/511446.511513. URL https://doi.org/10.1145/511446.511513.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V. Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering. In *Advances in Neural Information Processing Systems*, volume 37, pp. 132876–132907. Curran Associates, Inc., 2024. doi: 10.52202/079017-4224. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/efaf1c9726648c8ba363a5c927440529-Paper-Conference.pdf.

Gautier Izacard and Edouard Grave. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 874–880. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.74. URL https://aclanthology.org/2021.eacl-main.74/.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Jiho Kim, Yeonsu Kwon, Yohan Jo, and Edward Choi. KG-GPT: A General Framework for Reasoning on Knowledge Graphs Using Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9410–9421. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.631. URL https://aclanthology.org/2023.findings-emnlp.631/.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Mufei Li, Siqi Miao, and Pan Li. Simple is Effective: The Roles of Graphs and Large Language Models in Knowledge-Graph-Based Retrieval-Augmented Generation. In *International Conference on Learning Representations*, pp. 6061–6089, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/11e1900e680f5fe1893a8e27362dbe2c-Paper-Conference.pdf.

Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, 2004. URL https://aclanthology.org/W04-1013/.

Linhao Luo, Zicheng Zhao, Junnan Liu, Zhangchi Qiu, Junnan Dong, Serge Panev, Chen Gong, Thuy-Trang Vu, Gholamreza Haffari, Dinh Phung, Alan Liew, and Shirui Pan. G-reasoner: Foundation Models for Unified Reasoning over Graph-structured Knowledge. In *International Conference on Learning Representations*, pp. 63769–63793, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/67e2da8af2476c7441c881fc253042a6-Paper-Conference.pdf.

Costas Mavromatis and George Karypis. GNN-RAG: Graph Neural Retrieval for Efficient Large Language Model Reasoning on Knowledge Graphs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16682–16699. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.856. URL https://aclanthology.org/2025.findings-acl.856/.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling Relational Data with Graph Convolutional Networks. In *The Semantic Web*, pp. 593–607. Springer International Publishing, 2018. doi: 10.1007/978-3-319-93417-4_38. URL https://doi.org/10.1007/978-3-319-93417-4_38.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. REPLUG: Retrieval-Augmented Black-Box Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8371–8384. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.463. URL https://aclanthology.org/2024.naacl-long.463/.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HkgEQnRqYQ.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Eric Gaussier, and Guillaume Bouchard. Complex Embeddings for Simple Link Prediction. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2071–2080. PMLR, 2016. URL https://proceedings.mlr.press/v48/trouillon16.html.

Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. Composition-based Multi-Relational Graph Convolutional Networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BylA_C4tPr.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

Zhishang Xiang, Chuanjie Wu, Qinggang Zhang, Shengyuan Chen, Zijin Hong, Xiao Huang, and Jinsong Su. When to use Graphs in RAG: A Comprehensive Analysis for Graph Retrieval-Augmented Generation. In *International Conference on Learning Representations*, pp. 66145–66178, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/6c9e01d6cefbbf4cdd265032550e767f-Paper-Conference.pdf.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How Powerful are Graph Neural Networks? In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective Retrieval Augmented Generation. *arXiv preprint arXiv:2401.15884*, 2024. doi: 10.48550/arXiv.2401.15884. URL https://arxiv.org/abs/2401.15884.

A Evidence Alignment and Graph Diagnostics

We align question-specific evidence triples with the entities and relations in the constructed graph. A full match preserves both endpoints and their relation; a partial match preserves the endpoints but omits the relation. Entity failures cover cases in which an endpoint is absent or cannot be correctly aligned. This distinction identifies whether a relation predictor operating over existing entities can address the missing evidence.

The Medical diagnosis contains 1,367 evidence occurrences: 198 full matches, 833 partial matches, and 336 in the remaining category. Full and partial matches account for 14.48% and 60.94%, respectively, rounded to 14.5% and approximately 61% in the main text. The Novel graph has 43.2% coverage and 38.4% partial evidence. Its 9.5% entity-collision rate captures an additional source of error: identical names can refer to different entities across documents.

We also assess 400 sampled graph edges per domain. Edge precision is 85.8% on Medical and 82.3% on Novel. Edge precision and evidence coverage answer different questions: the former measures the correctness of extracted relations, while the latter measures whether relations required for the benchmark questions survive graph construction. A graph can therefore have high edge precision and still omit much of the evidence needed for answering.

B Architecture, Training, and Inference

B.1 Shared Encoder and Training Objectives

The GNN initializes entities with the text embeddings associated with graph construction. Two GAT layers combine these features with graph neighborhoods. Four heads share the resulting representations: link prediction scores candidate connections, confidence scoring estimates their quality, query relevance ranks their relevance to the question, and QAG determines query-dependent structural selection. For QAG, the edge and query representations are concatenated before the MLP and sigmoid output.

Training evidence triples provide positive supervision for relation prediction. We jointly optimize the four heads with the weights in Eq. 4, using Adam and early stopping on validation loss. The answer-generating LLM remains frozen. Table 4 collects the model and optimization settings.

Table 4: Enhancement-layer implementation. Parameter counts refer to the GNN and its prediction heads.

Setting	Value
Encoder	Two-layer GAT
Hidden dimension / attention heads	128 / 4
Dropout	0.3
QAG head	$256 \to 64 \to 1$ MLP, sigmoid
Base / full model parameters	Approximately 229K / 245K
Loss weights: LP / confidence / relevance / QAG	$1 : 1 : 1 : 0.5$
Optimizer / learning rate	Adam / 10^{-3}
Training epochs	Up to 200, validation-loss early stopping
Training platform / time	CPU / less than two minutes
Medical / Novel verbal hint budget	15 / 20 edges per query
Default answer generator	Qwen-Plus, frozen

B.2 Structural and Verbal Configurations

The structural path selects predicted edges with QAG before PPR-based evidence retrieval. The gate cutoff is selected on validation data. The verbal path ranks candidate relations and appends a bounded set of hints to the generation prompt. Query-relevance ranking gives the verbal-only comparison, while QAG ranking is evaluated as an alternative and used in the combined QAG configuration. Each hint uses the template "A is related to B." Medical uses $K = 15$ and Novel uses $K = 20$. The fixed-confidence structural baseline uses cutoff .85 on Medical and .80 on Novel. Table 5 distinguishes the inputs changed by each intervention.

Table 5: Configuration definitions. The final generator is frozen in every row.

Configuration	Retrieval graph	Generation context
Baseline	Original G	Question and retrieved passages
Structural only	G_q selected by QAG	Question and passages retrieved from G_q
Verbal only	Original G	Baseline context plus top-K hints
Dual path	G_q selected by QAG	Augmented-retrieval context plus top-K hints

The paths share the learned relation model but need not select identical edge sets. Structural additions affect graph transitions and retrieved passages. Verbal hints enter after passage retrieval and are interpreted by the generator together with the retrieved evidence. The combined configuration makes both changes while keeping the generator frozen.

C Evaluation Protocol and Comparisons

The study uses subsets of 509 Medical questions and 609 Novel questions from GraphRAG-Bench. These counts refer to the study subsets. The main answer-quality results and cross-system comparisons are computed over all 509 and 609 questions, respectively. The 60% training, 20% validation, and 20% testing partition governs GNN fitting: training evidence triples supply relation supervision, and validation data support model selection. Downstream answer evaluation spans the complete study subsets, including questions in the GNN training partition. The same trained enhancement model is reused across retrieval systems.

ACC measures answer correctness on a 0–1 scale. We express an ACC-point gain as 100 times the difference between the augmented and baseline scores. The distinction from lexical overlap is illustrated in Table 11: answers with different wording can receive the same correctness score. ROUGE-L measures overlap with the reference answer. Link-prediction performance is measured by Hits@10 and mean reciprocal rank. We use 10,000 bootstrap resamples for the 95% confidence intervals in Table 6.

Table 6: Dual-path gains relative to the RAG baseline, with 95% bootstrap confidence intervals. Gains and interval endpoints are ACC points.

Domain	Gain	95% confidence interval
Medical	$+7.94$	$[+5.11, +10.77]$
Novel	$+5.54$	$[+2.88, +8.20]$

C.1 Controls for Relation Selection and Context

The structural comparison tests three ways of using the predictor: insert all predicted edges, filter by a fixed confidence cutoff, or select by QAG. The cutoff sweep evaluates how filtering affects both the retained candidate fraction and downstream answer quality. Its purpose is to distinguish the benefit of predicting additional connections from the benefit of selecting which predictions enter graph propagation.

The verbal comparison tests whether improvements depend on predicted relations. Random-edge hints control for adding relational text to the prompt. Existing-edge hints sample relations already present in the original graph. Entity-pair BM25 retrieves corpus sentences containing the endpoints of predicted edges, giving a source-text alternative to the compact relational hint. These controls compare the content and presentation of the additional context; the separate hint-budget sweep measures sensitivity to its quantity.

D Link Prediction and Structural Selection

We compare TransE (Bordes et al., 2013), RotatE (Sun et al., 2019), ComplEx (Trouillon et al., 2016), GIN (Xu et al., 2019), and GAT (Veličković et al., 2018). The GNN models use entity text features. TransE with text initialization provides an additional comparison that introduces textual information without GNN neighborhood aggregation.

Table 7: Link-prediction performance. Parameter counts are those of our evaluated configurations.

	Medical		Novel		
Model	Hits@10	MRR	Hits@10	MRR	Parameters
TransE	.714	.381	.802	.523	205K
TransE + text initialization	.743	.412	.831	.557	205K
RotatE	.752	.426	.843	.582	410K
ComplEx	.738	.408	.826	.561	410K
GIN	.823	.500	.912	.668	214K
GAT	.836	.519	.921	.682	229K
GAT + QAG	**.841**	**.527**	**.925**	**.691**	245K

GAT+QAG attains the strongest link-ranking scores in both domains, but its improvement over GAT is small relative to the effect of downstream selection. The confidence head achieves AUC in the range .93–.95, and the query-relevance head achieves nDCG@10 in the range .74–.80. These head-level metrics concern prediction quality and ranking, while Table 8 measures whether selected predictions improve the final answers.

Table 8: Structural selection sweep. Gains are ACC points; retention is the average fraction of candidate edges retained.

Selection	Medical gain	Novel gain	Retention
All predicted edges	-5.23	$+0.12$	100%
Fixed cutoff .60	-1.89	$+1.08$	68%
Fixed cutoff .70	$+0.40$	$+1.65$	47%
Fixed cutoff .80	$+1.89$	$+2.24$	31%
Fixed cutoff .85	$+2.81$	$+2.05$	22%
Fixed cutoff .90	$+1.70$	$+1.38$	14%
Fixed cutoff .95	$+0.45$	$+0.54$	6%
QAG	**+3.47**	**+2.73**	$\approx$25%

In Medical, increasing the fixed cutoff first improves the gain and then reduces it as more candidates are discarded. The best fixed cutoff differs between domains. QAG improves over the best fixed-cutoff configurations while retaining approximately one quarter of candidates, linking the selection decision to the question rather than relying only on relation confidence.

E Verbalization, Hint Budget, and Answer Form

E.1 Ranking and Template Choice

Table 9 compares query-conditioned ranking with link-prediction and cosine-similarity scores. The query-relevance and QAG heads both improve on the two alternatives. The generic template also exceeds typed relations and confidence-bearing hints. Predicted relation-type accuracy is approximately 67%, so including the predicted type introduces another potential error into the prompt. The generic template communicates an endpoint connection while leaving its interpretation to the generator and retrieved context.

E.2 Number of Hints

The hint-budget sweep in Table 10 identifies a useful intermediate range. Medical peaks at 15 hints and Novel at 20; larger hint sets reduce gains on both domains. Thus, the value of verbal augmentation

Table 9: Verbal ranking and template ablations on Medical. All gains use the .3917 baseline. The ranking variants are distinct configurations.

Configuration	ACC	Gain
LP-score ranking	.4087	+1.70
Query–endpoint cosine ranking	.4279	+3.62
Query-relevance head	.4412	+4.95
QAG-score ranking	**.4452**	**+5.35**
Generic template: "A is related to B"	.4412	+4.95
Relation-typed template: "A treats B"	.4309	+3.92
Confidence values in the hint	.4241	+3.24

comes from selecting a limited set of useful relations, rather than exposing the generator to every prediction.

Table 10: Number of verbal hints per query. Values are ACC-point gains.

K	Medical	Novel
5	+2.71	+1.32
10	+4.03	+2.41
15	**+4.95**	+2.79
20	+4.78	**+3.05**
30	+3.94	+2.83
50	+2.47	+1.71

E.3 Lexical Overlap and Answer Expansion

Verbal augmentation changes Medical ROUGE-L from .1842 to .1786 and Novel from .2088 to .1904, while improving ACC. Dual-path augmentation reaches .1897 ROUGE-L on Medical. The two metrics therefore capture different effects of augmentation: answer quality can improve even when the generated wording becomes less similar to the reference.

Table 11 illustrates this behavior with Medical example Q-127. The baseline closely paraphrases the reference. The verbal-path answer adds a mechanistic explanation and terminology related to glucose metabolism. Both variants receive ACC 1 in this example, but the expanded answer has lower ROUGE-L. This example connects the aggregate metric difference to a concrete change in answer content and length.

Table 11: Answer comparison for Medical example Q-127. Scores are per-example values.

Variant	Answer text	ACC	ROUGE-L
Reference	Metformin is used for type 2 diabetes treatment.	—	—
Baseline	Metformin is commonly prescribed for managing type 2 diabetes.	1	.72
Verbal path	Metformin treats type 2 diabetes by reducing hepatic glucose production, as it is related to glucose metabolism pathways.	1	.41

F Novel Entity Disambiguation

Novel contains namesakes across documents. We attach document prefixes to entity identifiers and restrict candidate connections to the same document. We then add anchor expansion and an edge-reward training signal with coefficient $\lambda = .5$. Table 12 reports these stages cumulatively.

Unqualified verbal hints reduce ACC below the baseline. Document-qualified identifiers and the same-document constraint reverse that loss; the subsequent expansion and reward stages further improve the score. The final dual-path configuration combines the resulting verbal hints with structural

Table 12: Cumulative Novel configurations. The staged additions address entity identity and relation use; each row includes the preceding adjustments.

Configuration	ACC	Gain
RAG + rerank	.3833	—
Verbal hints without document prefixes	.3518	−3.15
Document prefixes and same-document constraints	.3891	+0.58
Anchor expansion	.4021	+1.88
Edge reward, $\lambda = .5$	.4138	+3.05
Dual-path augmentation	**.4387**	**+5.54**

augmentation. This ablation makes endpoint identity part of the augmentation procedure, rather than assuming that a shared surface name denotes the same entity throughout the corpus.

G System Integration and Generator Comparison

We evaluate six retrieval-system configurations under a shared benchmark graph and a common Qwen-Plus generator. The GraphRAG configurations are based on their official implementations. We train the enhancement model once and reuse it across systems. Structural augmentation precedes each system's retrieval stage; verbalized edges are appended to its generation prompt. Base scores are reproduced under this shared setup. Table 13 gives the full comparison and Figure 3 summarizes the dual-path gains.

Table 13: Full cross-system results. S: structural path; V: verbal path; D: dual path.

	Medical				Novel			
System	Base	S	V	D	Base	S	V	D
RAG + rerank	.3917	.4264	.4412	.4711	.3833	.4106	.4138	.4387
HippoRAG2	.5124	.5387	.5534	.5756	.4828	.5014	.5083	.5247
G-reasoner	.5067	.5298	.5421	.5612	.4951	.5108	.5162	.5319
LightRAG	.4893	.5154	.5297	.5498	.4609	.4791	.4852	.5017
MS-GraphRAG	.4571	.4817	.4963	.5152	.4312	.4503	.4567	.4734
Fast-GraphRAG	.4738	.4981	.5109	.5308	.4537	.4716	.4778	.4951

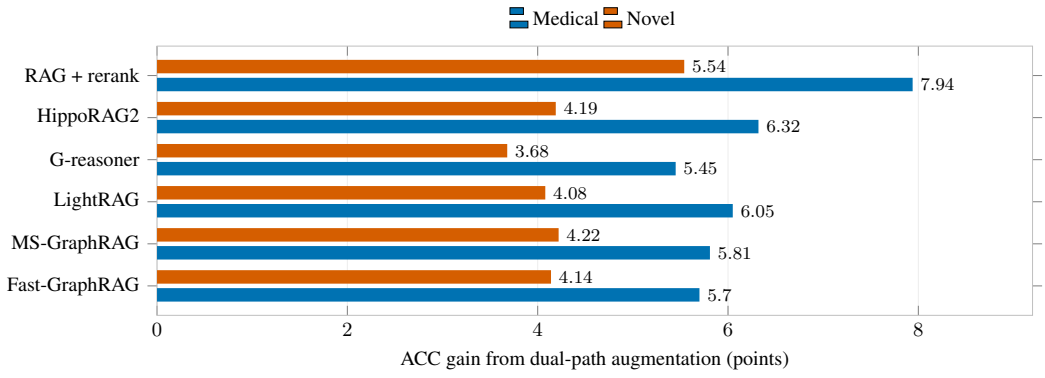


Figure 3: Dual-path gains across the six evaluated retrieval-system configurations. Graph and generator settings are shared; each gain is relative to the corresponding reproduced base score.

The generator comparison holds the domain fixed to Medical and evaluates the combined enhancement with Qwen-Plus, GPT-4o, and Llama-3-70B-Instruct. GPT-4o attains the highest final ACC among these backbones. All three improve after augmentation, supporting the use of the enhancement with different frozen generators.

Table 14: Generator comparison on Medical. Each row evaluates the combined enhancement.

Generator	Base ACC	Dual ACC	Gain
Qwen-Plus	.3917	.4711	+7.94
GPT-4o	.4673	.5327	+6.54
Llama-3-70B-Instruct	.4198	.4892	+6.94

H Stage-Level Effects and Hint Adoption

The controls in Table 15 separate evidence retrieval from generation with a fixed evidence set. Structural augmentation changes retrieved passages, so its contribution disappears when gold passages are supplied directly. Verbal augmentation leaves retrieval unchanged but can improve generation from those fixed passages. The end-to-end comparison measures the effect of combining the two interventions.

Table 15: Retrieval and generation controls on Medical. Gold-passage generation fixes retrieved evidence across the compared configurations.

Metric	Baseline	Structural	Verbal	Dual
Retrieval Recall@5	.412	.463	.412	.463
Generation ACC with gold passages	.618	.618	.659	.659
End-to-end ACC	.3917	.4264	.4412	.4711

H.1 Answer-Level Hint Adoption

We randomly sample 100 Medical test questions and annotate their 1,500 verbalized edges. Two annotators independently judge whether each relation is adopted in the final answer, obtaining Cohen's $\kappa = .81$. Table 16 separates correct and relevant relations, incorrect predictions, and correct but irrelevant relations.

Table 16: Answer-level adoption of verbalized relations. The measurement concerns generated answers, rather than internal attention weights.

Hint category	Adopted	Not adopted
Correct and relevant	87.2%	12.8%
Incorrect	26.4%	73.6%
Correct but irrelevant	5.1%	94.9%

Correct and relevant relations are adopted much more frequently than the other categories. Most incorrect predictions are ignored, but their 26.4% adoption rate shows that generation does not reliably remove all prediction errors. These annotations measure answer-level use of hints; they do not measure the model's internal attention weights.

I Supervision and Efficiency

Training on existing graph edges provides a self-supervised alternative to annotated evidence. Table 17 reports the Medical comparison. Graph-derived supervision lowers both link-ranking performance and the downstream gain, but dual-path augmentation still improves on the unaugmented baseline. This comparison supports the practical value of the graph's own relational signal while retaining the advantage of evidence supervision in the evaluated setting.

Table 17: Supervision comparison on Medical. Downstream values are dual-path ACC-point gains.

Training supervision	Hits@10	Dual gain
Annotated evidence	.841	+7.94
Existing graph edges	.793	+5.87

Table 18: Additional inference cost in our timing setup. Token changes are relative to the baseline prompt.

Component	Additional time per query	Prompt-token increase
GNN inference on CPU	< 0.1 s	—
QAG scoring	< 1 ms	—
Structural augmentation	≈ 0.3 s	4.2%
Verbalization	< 1 ms	11.8%
Dual path, total	≈ 0.4 s	16.0%

The total additional latency is approximately 0.4 seconds per query, with a 16.0% increase in prompt tokens. CPU GNN inference takes less than 0.1 seconds, while graph augmentation accounts for approximately 0.3 seconds. Gate scoring and verbalization each take less than 1 millisecond. These component times are parts of the total pipeline cost rather than separate overheads to add again to the dual-path total. The measurements characterize our implementation; actual latency depends on the graph size, serving environment, and generator.