PROPOSITIONS AS STATES: CONTEXT-PRESERVING GRAPH TRAVERSAL FOR MULTI-HOP RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-Augmented Generation (RAG) complements the parametric knowledge of Large Language Models (LLMs) with external knowledge, but multi-hop queries require retrieving interconnected evidence beyond direct semantic similarity. Graph-based methods support this process by connecting information through entities and relations, with recent approaches making graph traversal query-adaptive. However, these advances adapt how the graph is traversed while continuing to use entities as traversal states. When traversal paths from different relational contexts converge on the same entity, the context through which the entity became relevant is not retained as part of the traversal state. To address this limitation, we propose **ProST (Propositions as States)**, a proposition-state graph traversal framework that maintains relevance over propositions while using entities as structural connections between them. By preserving source context within each proposition, ProST conditions subsequent transitions on both the query and the current proposition. Our approach thus jointly considers knowledge representation and graph traversal, allowing preserved context to directly guide query-adaptive traversal. Experiments across multiple QA benchmarks demonstrate consistent improvements in evidence retrieval and downstream question answering over strong retrieval baselines. Further analyses show that both proposition states and query-adaptive transitions contribute to retrieving complete multi-hop evidence chains. Our code is available at Link.

1 INTRODUCTION

Retrieval-Augmented Generation (RAG) augments Large Language Models (LLMs) with external knowledge beyond their parametric knowledge (Lewis et al., 2020; Guu et al., 2020; Borgeaud et al., 2022; Ram et al., 2023; Gao et al., 2023). In multi-hop questions, required evidence is often distributed across documents and connected through intermediate information (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022; 2023). Because some evidence may not be directly similar to the query, effective retrieval requires representing and exploring relationships among distributed evidence.

Knowledge graphs (KGs) provide such relational structure by organizing distributed information as entity–relation triples (Edge et al., 2024; Gutiérrez et al., 2024; Guo et al., 2025; Peng et al., 2024). Personalized PageRank (PPR) has become a widely used approach for retrieving over these structures, distributing relevance from query-relevant seeds across graph connections (Haveliwala, 2002; Gutiérrez et al., 2024; 2025). Recent methods further make graph traversal query-adaptive by using query information to guide graph exploration or adjust transitions (Lau et al., 2026; Zhou et al., 2026; Zhuang et al., 2026; Wu et al., 2026).

However, existing query-adaptive graph traversal methods largely rely on entities as traversal states. In entity-based traversal, the context through which an entity becomes relevant is not retained as part of the traversal state. For example, in Figure 1, Randy Conrads is connected to Oregon State University (OSU) through his bachelor's degree, but this context is not retained once relevance reaches the OSU entity. Consequently, prior context cannot directly guide subsequent traversal, allowing relevance to spread toward query-related distractors associated with bachelor's degrees, such as a dual-degree program or Jensen Huang. This problem can become more pronounced as additional hops introduce more potential distractors.

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.

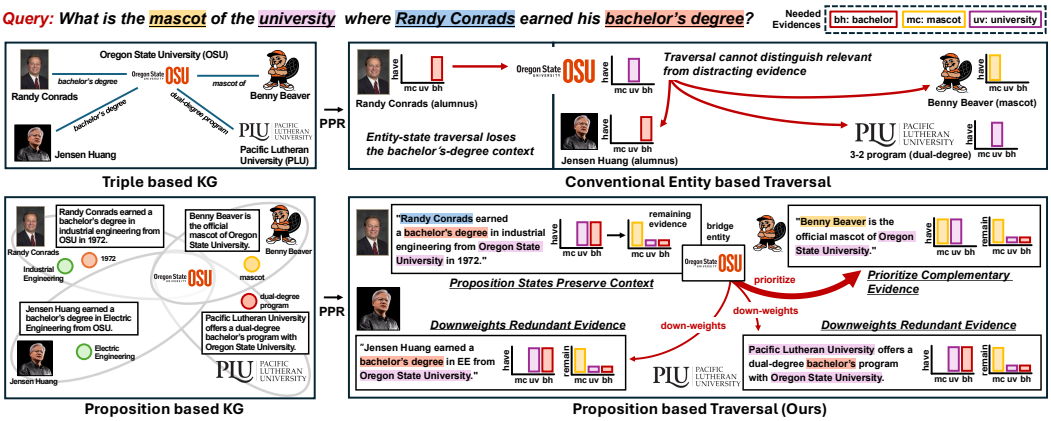

Figure 1: Comparison of entity-based and proposition-based graph traversal.

Effective multi-hop traversal therefore requires a state representation that retains sufficient context to guide subsequent relevance transfer. However, entity–relation triples may lose contextual details, while coarse textual units may include irrelevant information (Chen et al., 2024; Wang & Han, 2025). For example, earning a bachelor's degree and an MBA may both be reduced to the same "graduated" relation, obscuring their distinct contexts. Although richer graph representations can encode such information, doing so increases representational complexity (Wang & Han, 2025). This motivates preserving context within fine-grained units rather than explicitly structuring every contextual detail.

Propositions provide a natural representation for such context-preserving traversal states. As illustrated in Figure 1, a proposition represents an atomic, self-contained fact while preserving contextual details such as relations, attributes, and associated information within a single unit (Min et al., 2023; Chen et al., 2024). Recent studies use propositions as fine-grained retrieval units or exploit their connections for multi-hop retrieval (Chen et al., 2024; Zhang et al., 2025; Wang & Han, 2025), but do not directly use their preserved context to guide subsequent graph traversal.

We therefore propose **proposition-state graph traversal**, which uses propositions as units for both retrieval and graph traversal. By maintaining relevance over propositions, ProST retains source context as part of the traversal state, allowing it to directly guide subsequent transitions. As illustrated in Figure 1, retaining the bachelor's-degree and university context guides traversal toward the remaining mascot evidence while downweighting redundant evidence. Entities provide structural connections between proposition states, while the query and source context jointly guide relevance toward the next proposition. Specifically, the query prioritizes bridge entities that can lead toward relevant evidence, while the source proposition context helps prioritize propositions that complement the information already represented. These transitions expand relevance beyond initially retrieved evidence without additional training or query-time LLM calls.

Our contributions are summarized as follows.

- We identify a limitation of entity-state traversal for multi-hop retrieval, where the context through which an entity becomes relevant is not retained for subsequent traversal.
- We propose **ProST**, a proposition-state graph traversal framework that preserves source context and uses it to guide query-adaptive transitions.
- We demonstrate consistent improvements in multi-hop evidence retrieval and downstream question answering, supported by analyses of proposition states and adaptive transitions.

2 Related Work

Graph-based RAG. RAG systems increasingly organize external knowledge beyond independent text chunks so that retrieval can follow relationships among knowledge units. Hierarchical approaches such as RAPTOR recursively cluster and summarize text, enabling retrieval at different levels of abstraction (Sarthi et al., 2024). Graph-based approaches instead represent relationships explicitly. HippoRAG builds an entity-centric knowledge graph for associative retrieval (Gutiérrez et al., 2024), while LightRAG combines graph-structured knowledge with text retrieval for local and

global queries (Guo et al., 2025). These approaches commonly rely on entity–relation structures, which provide explicit connections but may omit contextual information from the source text (Wang & Han, 2025). The choice of representation therefore determines not only how knowledge is stored, but also what context is available during graph traversal.

Query-adaptive graph traversal. Graph-based RAG can retrieve interconnected evidence by traversing relationships beyond direct query similarity, as exemplified by the PPR-based associative retrieval of HippoRAG (Gutiérrez et al., 2024). Subsequent work has adapted this process to the query in two main ways. Some methods determine *where traversal starts*: HippoRAG 2 filters query-relevant triples and incorporates passage-level signals (Gutiérrez et al., 2025), PropRAG selects seeds through beam search over proposition paths (Wang & Han, 2025), LinearRAG activates query-relevant entities through semantic bridging (Zhuang et al., 2026), and MemGraphRAG uses multi-layer memory to filter knowledge and initialize PPR (Wu et al., 2026). Other methods adapt *how relevance moves*: CatRAG dynamically weights transitions from query-relevant seeds (Lau et al., 2026), while QAFD-RAG weights edges according to query semantics through flow diffusion (Zhou et al., 2026). These methods improve where traversal begins or how relevance is transferred, but do not retain the source context as part of the traversal state. ProST instead maintains relevance over propositions, allowing subsequent transitions to depend on both the query and source proposition context.

Proposition-based retrieval. Propositions express atomic, self-contained facts and provide fine-grained units that retain contextual information from the source text. Dense X Retrieval studies retrieval granularity and demonstrates the effectiveness of propositions as retrieval units (Chen et al., 2024). SiReRAG organizes fine-grained knowledge units through semantic and relational connections (Zhang et al., 2025), while PropRAG explores proposition paths through shared entities to identify multi-hop evidence chains before subsequent graph retrieval (Wang & Han, 2025). These methods demonstrate the value of propositions for knowledge representation and retrieval, but do not maintain propositions as states throughout subsequent graph traversal. In contrast, we use propositions as both retrieval units and traversal states, allowing their preserved context to directly guide subsequent transitions.

3 METHOD

ProST retrieves multi-hop evidence by traversing a proposition graph, keeping relevance on propositions so that the context of each retrieved fact remains available to subsequent transitions. The framework consists of offline graph construction (Section 3.2) and online retrieval (Section 3.3), as illustrated in Figure 2.

3.1 PRELIMINARIES

Proposition representation. Let $\mathcal{D}$, $\mathcal{P}$, and $\mathcal{E}$ denote the sets of passages, propositions, and entities. A proposition $p \in \mathcal{P}$ represents an atomic, self-contained fact (Min et al., 2023; Chen et al., 2024) and is associated with the entities $E(p) \subseteq \mathcal{E}$ that occur in it. Conversely, $P(e)$ denotes the propositions containing an entity e, and $P(d)$ those extracted from a passage $d \in \mathcal{D}$. Propositions serve as traversal states, while shared entities provide structural connections between them.

Personalized PageRank. Personalized PageRank (PPR) distributes relevance over graph states while periodically restarting from a seed distribution (Page et al., 1999; Haveliwala, 2002). Given a transition matrix $\mathbf{T}$, seed distribution $\mathbf{r}$, and damping factor α with restart probability $1-\alpha$, the relevance distribution $\boldsymbol{\pi}^{(t)}$ at iteration t is updated as $\boldsymbol{\pi}^{(t+1)} = (1-\alpha)\mathbf{r} + \alpha\mathbf{T}^{\top}\boldsymbol{\pi}^{(t)}$. Conventional PPR uses fixed transitions, whereas ProST's transitions depend on the query and source proposition.

Problem formulation. Given a query q and a corpus $\mathcal{D}$, we retrieve a set of passages $\mathcal{R}_q \subseteq \mathcal{D}$ that collectively covers the information required to answer q. For multi-hop queries, this evidence can be distributed across passages and connected through intermediate information that is not directly similar to the query. We therefore formulate retrieval as query-dependent traversal over proposition states, where $T(p_{t+1} \mid p_t, q)$ denotes the transition from the current proposition p_t to the next proposition p_{t+1}, conditioned on the query and current proposition context.

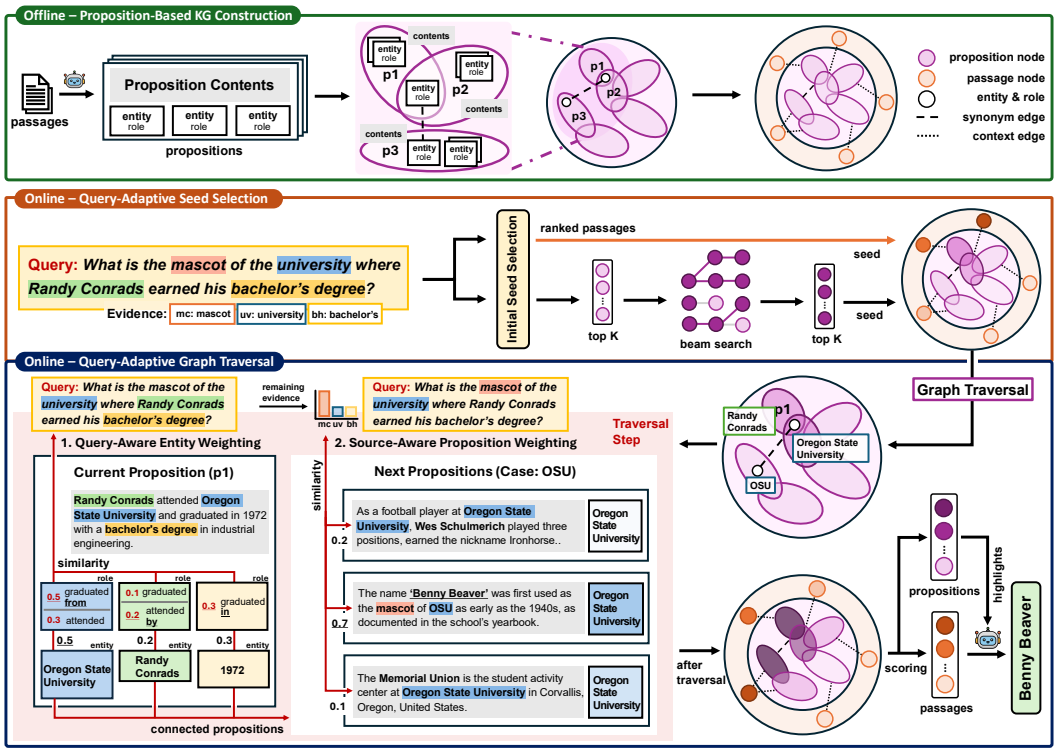

Figure 2: Overview of the ProST retrieval framework.

3.2 Offline: Role-Aware Proposition Graph Construction

Given a corpus $\mathcal{D}$, we construct a graph over propositions $\mathcal{P}$ connected through shared entities $\mathcal{E}$. Beyond extracting propositions and entities, we associate each proposition–entity pair (p, e) with a semantic role $r_{p,e}$ that describes how entity e participates in proposition p.

Proposition extraction. We adopt LLM-based proposition extraction (Wang & Han, 2025), which decomposes each passage into atomic, self-contained propositions and extracts their constituent entities. Each proposition is constructed as a standalone statement that preserves its full context, with the complete extraction prompt provided in Appendix G. Propositions sharing an entity are structurally connected through that entity, while each proposition remains linked to its source passage.[1] This indexing step is the only LLM-based operation in the retrieval pipeline. Extraction examples are provided in Appendix A.2, with proposition quality evaluated in Appendix A.3.

Role-aware representation. Entities can participate differently in the relation expressed by a proposition. For example, given the proposition *person attended university*, the university is assigned *attended*, while the person is assigned its inverse *attended by*. We represent this relational information as a proposition-specific role $r_{p,e}$, derived from its OpenIE triple.[2] Examples of proposition-specific roles and coverage across datasets are in Appendix A.2.

3.3 Online: Query-Adaptive Graph Traversal

Given a query q and the proposition graph, we first identify a focused set of query-relevant seeds and then expand relevance beyond these seeds through proposition-state graph traversal. At each proposition, Query-Aware Entity Weighting prioritizes useful bridge entities, while Source-Aware Proposition Weighting prioritizes connected propositions based on the query and source context. The resulting query-adaptive transitions are used in PPR to expand relevance beyond the paths explored during seed selection.

[1] Synonym edges connect similar entities, while context edges link propositions to their source passages.

[2] Pairs without an available role retain their graph connection but are not assigned a role.

3.3.1 QUERY-ADAPTIVE SEED SELECTION

Multi-hop retrieval requires reasoning over interconnected evidence, and beam search efficiently explores such reasoning paths (Wang & Han, 2025). We therefore use beam search to identify query-relevant proposition paths. However, a fixed set of explicit paths may leave complementary evidence unexplored, so we use their propositions as seeds for query-adaptive graph traversal.

3.3.2 QUERY-ADAPTIVE TRANSITION

To preserve source context, we keep propositions as traversal states and shared entities as bridges between them. A transition from p_t to p_{t+1} is then factorized through a shared entity e as

$$T(p_{t+1} \mid p_t, q) = \sum_{e \in E(p_t) \cap E(p_{t+1})} T_{P \to E}(e \mid p_t, q)\, T_{E \to P}(p_{t+1} \mid p_t, e, q). \tag{1}$$

The first factor prioritizes query-relevant bridge entities, while the second prioritizes propositions based on the query and source context.

Query-Aware Entity Weighting. A proposition may contain multiple entities, but only some provide useful bridges toward the required evidence. We therefore weight each entity $e \in E(p_t)$ according to the relevance of its role to the query, with $s_E(e \mid p_t, q) = \mathrm{sim}(z_q, z_{r_{p_t,e}})$.[3] Here, z_q and $z_{r_{p_t,e}}$ denote the query and role embeddings, respectively, and $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity. The scores are normalized over the entities in p_t as

$$T_{P \to E}(e \mid p_t, q) = \frac{\exp(\tau_E s_E(e \mid p_t, q))}{\sum_{e' \in E(p_t)} \exp(\tau_E s_E(e' \mid p_t, q))}, \tag{2}$$

where τ_E controls the sharpness of the weighting. This prioritizes bridge entities whose roles are most relevant to the query.

Source-Aware Proposition Weighting. A bridge entity may connect to multiple propositions, including those that repeat information already represented by the source proposition. We therefore condition proposition weighting on p_t, using the residual $z_q - z_{p_t}$ to emphasize query information less represented by the source proposition. Each candidate $p' \in P(e)$ is scored as $s_P(p' \mid p_t, e, q) = \mathrm{sim}(z_q - z_{p_t}, z_{p'})$, where z_{p_t} and $z_{p'}$ denote the embeddings of the source and candidate propositions. The scores are normalized over the propositions connected to e as

$$T_{E \to P}(p' \mid p_t, e, q) = \frac{\exp(\tau_P s_P(p' \mid p_t, e, q))}{\sum_{\tilde{p} \in P(e)} \exp(\tau_P s_P(\tilde{p} \mid p_t, e, q))}, \tag{3}$$

where τ_P controls the sharpness of the weighting. Because the weighting depends on p_t, the same entity can lead to different proposition distributions depending on the source context. This prioritizes propositions that complement the information represented by the source proposition.[4]

3.3.3 PROPOSITION-STATE PPR WITH PASSAGE INTEGRATION

We use the query-adaptive transitions in PPR to expand relevance beyond the seed paths.

Passage Integration. Following prior graph retrievers (Gutiérrez et al., 2025; Wang & Han, 2025), we retain passage nodes to incorporate direct passage-level retrieval signals. Each proposition transfers a fraction β of its relevance to its source passage and the remaining $1 - \beta$ through the query-adaptive transitions. Each passage redistributes its relevance uniformly to its propositions.

Joint PPR. We maintain proposition and passage relevance $\boldsymbol{\pi}_P$ and $\boldsymbol{\pi}_D$, with $\boldsymbol{\pi} = [\boldsymbol{\pi}_P; \boldsymbol{\pi}_D]$ denoting their joint distribution. The joint restart distribution $\mathbf{r} = [\mathbf{r}_P; \mathbf{r}_D]$ combines query-adaptive proposition seeds with dense passage retrieval scores. The joint relevance is updated as

$$\boldsymbol{\pi}^{(t+1)} = (1 - \alpha)\mathbf{r} + \alpha \mathbf{T}_q^\top \boldsymbol{\pi}^{(t)}, \tag{4}$$

where $\mathbf{T}_q$ is the joint transition matrix combining query-adaptive proposition transitions with proposition–passage transfer. The resulting $\mathbf{T}_q$ is row-stochastic, and Equation 4 converges to a unique fixed point for $0 < \alpha < 1$, as shown in Appendix E.

[3] Entities without an available role receive $s_E = 0$.

[4] Alternative designs for both transition weightings are analyzed in Appendix B.1.

3.3.4 FINAL RANKING AND EVIDENCE HIGHLIGHTS

Final Ranking. After convergence, passages combine proposition- and passage-level relevance,

$$s(d) = \sum_{p \in P(d)} \pi_P(p) + \pi_D(d), \tag{5}$$

where $\pi_P(p)$ and $\pi_D(d)$ denote proposition- and passage-level relevance, respectively.[5] The top-k passages are returned as the retrieval result, where k denotes the number of retrieved passages.

Evidence Highlights. To surface query-relevant information within retrieved passages, we prepend the N propositions whose embeddings are most similar to the query as *evidence highlights*. Because ProST maintains propositions as traversal states, this requires no further graph traversal and introduces no additional evidence. The prompt instructs the LLM to use the highlights as guidance when processing the retrieved passages, with the full instruction provided in Appendix G.

4 EXPERIMENTS

We conduct experiments to answer the following research questions.

- **RQ1:** How effective is ProST for evidence retrieval and downstream QA?
- **RQ2:** How do proposition states and query-adaptive transitions contribute to retrieval?
- **RQ3:** When does proposition-state traversal provide greater benefits?

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate on three multi-hop QA benchmarks, MuSiQue (Trivedi et al., 2022), 2WikiMultiHopQA (Ho et al., 2020), and HotpotQA (Yang et al., 2018), using 1,000 questions and prior-work corpora (Gutiérrez et al., 2025; Wang & Han, 2025). We also evaluate on the single-hop ChemLit-QA benchmark (Wellawatte et al., 2025) to assess ProST in a scientific domain, where shared entities connect many passages. Corpus and graph statistics are provided in Appendix A.1.

Metrics. For retrieval, we report Recall@$\{2, 5\}$ and Full Chain Retrieval (FCR@5) (Lau et al., 2026), the fraction of questions where all gold passages are retrieved within the top 5. For single-gold ChemLit-QA, we use MRR@5 instead of FCR@5. Additional R-Precision and coverage results are provided in Appendices C.1 and C.3. For downstream QA, we report Exact Match (EM), F1, and LLM-Acc using the LinearRAG judge prompt (Zhuang et al., 2026) with Qwen2.5-32B-Instruct (Yang et al., 2025). All systems use Llama-3.3-70B-Instruct (Grattafiori et al., 2024) with the same QA prompt from Wang & Han (2025). LLM robustness is reported in Appendix D.2.

Baselines. Dense retrievers include Contriever (Izacard et al., 2022), GTR (Ni et al., 2022), GTE-Qwen2-7B (Li et al., 2023), GritLM-7B (Muennighoff et al., 2025), and NV-Embed-v2 (Lee et al., 2025). Structure-aware retrievers include Dense X (Chen et al., 2024), RAPTOR (Sarthi et al., 2024), HippoRAG (Gutiérrez et al., 2024), SiReRAG (Zhang et al., 2025), and LightRAG (Guo et al., 2025). Query-aware retrievers include HippoRAG 2 (Gutiérrez et al., 2025), PropRAG (Wang & Han, 2025), CatRAG (Lau et al., 2026), QAFD-RAG (Zhou et al., 2026), LinearRAG (Zhuang et al., 2026), and MemGraphRAG (Wu et al., 2026). Applicable methods use the same corpora and NV-Embed-v2 encoder. Encoder and training-based comparisons are in Appendices D.1 and C.2.

Implementation. We reuse proposition indexing and its OpenIE triples (Wang & Han, 2025) to derive entity roles, requiring no additional LLM calls or learned parameters. For efficiency, we apply adaptive transition scoring only to high-relevance entities and propositions. The sharpness parameters τ_E and τ_P, damping factor α, passage routing weight β, and focused-region size are selected on each dataset's validation split, with details and sensitivity analysis in Appendix D.3. All methods use their top-5 retrieved passages for downstream QA, with $N{=}3$ evidence highlights for ProST selected on validation.[6] All experiments run on three NVIDIA RTX PRO 6000 Blackwell GPUs (96 GB each), with local models served by vLLM (Kwon et al., 2023).

[5] The contributions of the proposition and passage channels are analyzed in Appendix B.2.

[6] The highlights add 77 tokens on average, about 10% of the retrieved passage text.

Table 1: Retrieval performance (%). Best in **bold**, second best underlined.

		MuSiQue			2Wiki			HotpotQA			ChemLit-QA		
	Method	R@2	R@5	FCR@5	R@2	R@5	FCR@5	R@2	R@5	FCR@5	R@2	R@5	MRR@5
Dense	Contriever	34.8	46.6	11.3	46.5	57.5	27.2	58.4	75.3	54.1	59.1	77.8	55.6
	GTR	37.4	49.1	15.0	60.2	67.9	35.9	59.2	73.9	53.0	72.2	85.6	69.0
	GTE-Qwen2-7B	47.7	63.5	30.6	66.2	74.3	46.5	76.5	90.0	80.2	66.6	81.0	63.1
	GritLM-7B	49.8	65.3	31.5	67.2	76.3	50.3	79.5	92.2	85.4	81.1	91.4	76.6
	NV-Embed-v2	52.6	69.4	37.5	69.2	76.7	50.8	84.0	94.3	89.0	80.4	92.9	77.6
Structure-aware	Dense X	52.5	68.6	37.2	67.5	75.2	48.0	80.0	93.2	86.6	**95.0**	97.8	**91.7**
	RAPTOR	45.3	59.3	28.7	52.9	60.4	36.3	76.3	89.0	79.5	76.2	87.8	72.9
	HippoRAG	42.0	53.8	26.7	73.6	91.5	82.4	62.7	80.0	67.9	30.7	42.0	29.1
	SiReRAG	40.2	52.1	23.5	49.5	57.1	33.2	74.1	85.6	75.4	75.2	82.7	71.3
	LightRAG	33.5	52.5	19.2	41.2	58.6	27.0	47.3	82.2	68.2	63.9	83.2	68.8
Query-aware	HippoRAG 2	55.7	74.4	46.3	76.4	90.8	75.8	83.4	96.2	92.7	80.0	94.7	73.2
	PropRAG	56.6	78.3	53.8	76.2	94.1	83.1	83.5	**97.3**	94.9	88.9	97.1	84.9
	CatRAG	57.8	74.9	47.6	75.0	87.8	69.1	**85.5**	96.2	92.8	80.4	93.7	76.0
	QAFD-RAG	48.8	67.9	40.5	64.3	83.9	64.9	74.2	91.7	85.8	54.7	74.7	53.2
	LinearRAG	39.8	58.9	29.6	63.3	82.6	64.1	55.2	72.2	53.1	74.5	87.2	71.0
	MemGraphRAG	55.2	73.8	45.3	76.6	90.9	76.1	83.0	96.1	92.5	74.7	92.5	70.7
	ProST	**58.8**	**79.0**	**55.6**	**78.0**	**95.8**	**88.4**	84.2	**97.3**	**95.0**	91.7	**98.2**	87.8

Table 2: Downstream QA performance (%). Acc: LLM-Acc. Best in **bold**, second best underlined.

		MuSiQue			2Wiki			HotpotQA			ChemLit-QA		
	Method	EM	F1	Acc	EM	F1	Acc	EM	F1	Acc	EM	F1	Acc
	No retrieval	18.4	27.7	21.9	40.6	47.5	44.5	38.3	49.7	45.0	1.1	27.9	11.7
Dense	Contriever	24.6	31.8	27.4	39.5	43.1	41.4	49.4	60.4	56.0	15.2	52.5	46.0
	GTR	25.7	35.1	30.0	50.9	54.1	52.4	49.5	61.3	56.8	15.8	54.8	49.7
	GTE-Qwen2-7B	32.5	43.0	38.2	55.3	59.9	58.3	58.7	71.5	67.5	16.2	53.4	48.0
	GritLM-7B	34.6	45.4	39.3	57.3	62.5	60.6	60.8	73.6	70.0	17.9	57.3	52.2
	NV-Embed-v2	34.7	45.2	39.5	57.6	62.5	60.5	61.8	74.7	70.3	18.4	58.4	54.9
Structure-aware	Dense X	36.7	47.9	42.7	56.7	61.6	59.1	60.1	73.3	69.3	**19.1**	59.4	54.8
	RAPTOR	21.2	35.7	30.6	42.1	50.3	45.6	54.9	70.8	65.2	16.4	56.4	52.5
	HippoRAG	27.4	36.6	32.1	65.3	72.6	69.1	53.8	65.9	61.6	8.0	40.8	31.0
	SiReRAG	29.9	45.5	38.9	51.7	59.9	55.6	58.5	74.2	69.3	16.7	55.2	51.1
	LightRAG	27.0	37.2	31.2	38.0	42.0	39.9	55.0	67.4	63.5	17.2	55.2	50.7
Query-aware	HippoRAG 2	38.8	50.7	45.4	65.7	72.4	69.9	62.8	75.5	71.4	**19.1**	58.8	54.8
	PropRAG	41.5	52.6	46.9	68.3	75.7	72.2	**63.3**	76.3	71.9	18.6	59.4	54.7
	CatRAG	38.9	50.2	45.1	64.2	70.5	67.9	62.3	75.4	71.1	18.2	58.5	55.1
	QAFD-RAG	36.2	46.5	41.8	59.3	65.7	63.0	60.3	72.6	68.7	13.9	51.4	45.3
	LinearRAG	29.7	39.6	34.6	59.5	66.0	62.8	50.2	62.4	58.1	17.5	56.5	52.9
	MemGraphRAG	37.7	48.9	44.1	65.8	72.6	69.9	62.6	75.5	71.5	17.8	58.3	53.5
	ProST	**45.3**	**56.3**	**50.8**	**68.9**	**76.8**	**73.6**	63.2	**76.4**	**73.0**	18.1	**60.3**	**58.1**

4.2 Main Results

Multi-hop QA. Tables 1 and 2 report retrieval and downstream QA performance. ProST achieves the best FCR@5 across all three multi-hop benchmarks, with larger improvements in FCR@5 than in Recall@5. This indicates that proposition-state traversal is particularly effective at retrieving complete evidence chains rather than individual relevant passages. These retrieval gains also translate into improved downstream QA with the same LLM. Appendix C.4 shows that the gains persist without evidence highlights and compares highlights with additional passages under matched context budgets.

Single-Hop Scientific QA. ChemLit-QA presents a different retrieval setting with single-hop scientific questions. These questions can be answered by directly retrieving fine-grained evidence without connecting information across multiple passages. This setting places less emphasis on graph traversal, and Dense X remains stronger in early-ranking metrics. Nevertheless, ProST achieves the highest Recall@5 and remains the strongest graph-based retriever, while also achieving the highest F1 and LLM-Acc. These results show that ProST remains competitive in scientific retrieval with a training-free design.[7]

[7] Comparisons with training-based retrievers are provided in Appendix C.2.

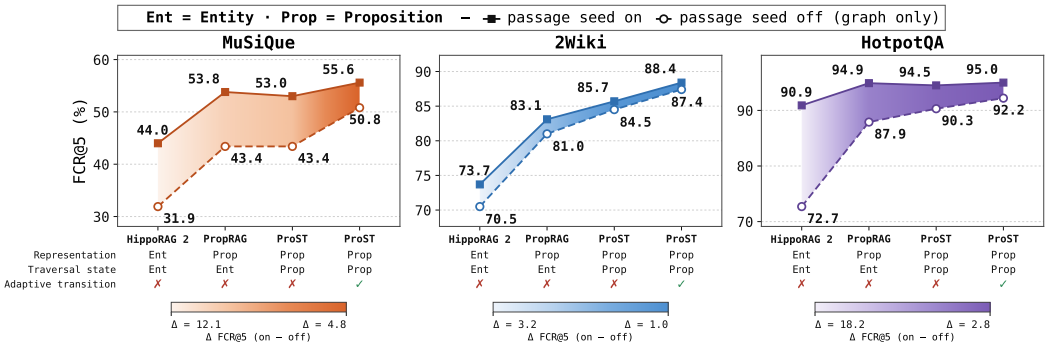

Figure 3: Component ablation on FCR@5, with (solid) and without (dashed) dense passage seeds.

Table 3: Ablation of transition components under different seeds. Best in **bold**.

Seed	Setting	MuSiQue R@5	MuSiQue FCR@5	2Wiki R@5	2Wiki FCR@5	HotpotQA R@5	HotpotQA FCR@5	Avg. R@5	Avg. FCR@5
HippoRAG seed	Full	**74.2**	**45.2**	91.0	77.2	**95.0**	**90.4**	**86.7**	**70.9**
	w/o query-aware entity weighting	73.2	43.5	86.5	67.0	94.7	89.9	84.8	66.8
	w/o source-aware proposition weighting	71.8	42.0	**92.3**	**80.4**	94.3	89.1	86.2	70.5
	w/o both (uniform)	71.0	40.4	87.1	68.7	94.2	88.8	84.1	66.0
HippoRAG 2 seed	Full	**75.5**	**48.9**	94.4	85.8	**95.8**	**92.3**	**88.5**	**75.7**
	w/o query-aware entity weighting	75.1	48.1	93.6	83.4	95.5	91.4	88.0	74.3
	w/o source-aware proposition weighting	73.1	45.8	**94.7**	**86.8**	95.5	91.3	87.8	74.6
	w/o both (uniform)	72.7	45.1	93.8	83.9	95.7	91.8	87.4	73.6
PropRAG seed (ProST)	Full	**79.0**	**55.6**	95.8	88.4	**97.3**	**95.0**	**90.7**	**79.7**
	w/o query-aware entity weighting	79.0	55.4	95.0	85.8	97.1	94.7	90.3	78.6
	w/o source-aware proposition weighting	77.6	52.9	**96.0**	**88.9**	97.0	94.4	90.2	78.7
	w/o both (uniform)	77.7	53.0	95.0	85.7	97.1	94.5	89.9	77.7

4.3 Component Analysis

Proposition representation and states. Figure 3 examines the roles of propositions as knowledge representations and traversal states in ProST, with the HippoRAG 2 LLM filter removed to ensure zero query-time LLM calls. Replacing entity-based representations with propositions substantially improves FCR@5, particularly in the graph-only setting, demonstrating the benefit of preserving context in the graph. Using propositions as traversal states further improves on 2WikiMultiHopQA and HotpotQA. However, on MuSiQue, which involves longer reasoning chains, changing the traversal state alone provides little additional benefit. This suggests that preserving context in the traversal state alone may not be sufficient for more complex multi-hop retrieval.

Context-aware graph traversal. The final step in Figure 3 examines whether the context retained by proposition states can guide graph traversal. Conditioning transitions on the query and source proposition improves FCR@5 across all three benchmarks, with larger gains in the graph-only setting. This shows that preserved context can directly guide subsequent transitions beyond changing the traversal state alone.[8] Moreover, the gap between results with and without dense passage seeds generally narrows as proposition representation, proposition states, and adaptive transitions are introduced. This indicates that context-aware graph traversal reduces reliance on passage retrieval.

Query and source context. Table 3 further decomposes adaptive traversal into Query-Aware Entity Weighting and Source-Aware Proposition Weighting to examine the roles of the query and source context. The former uses the query to prioritize useful bridge entities, while the latter uses the query and source proposition to prioritize candidate propositions connected through each entity. Their relative contributions vary across datasets, with query-aware entity weighting having a larger effect on 2WikiMultiHopQA and source-aware proposition weighting contributing more strongly on MuSiQue. Across HippoRAG, HippoRAG 2, and PropRAG seeds, the full transition achieves the highest average FCR@5, while making both factors uniform produces the largest average degradation. These results show that the two factors capture complementary signals, with the query guiding relevance toward useful bridge entities and the source context helping distinguish complementary propositions from redundant ones.

[8] Appendix B.3 analyzes relevance transfer toward the next gold passage at bridge entities.

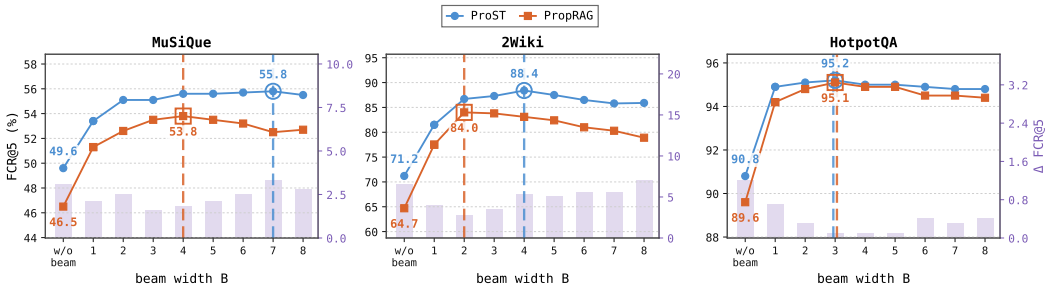

Figure 5: Performance across beam widths at depth 3. Rings mark the best and bars show the gap.

4.4 Performance Analysis

Question structure and hop count. Figure 4 examines how the gains vary with question structure and reasoning depth. On 2WikiMultiHopQA, the largest gain occurs on bridge-comparison questions, which require multi-step traversal through intermediate entities, whereas comparison questions relying on directly retrieved evidence show little difference. MuSiQue shows the same trend more directly, with the gain over PropRAG increasing from 2-hop to 4-hop questions. Together, these results show that the advantage of proposition-state traversal becomes more pronounced as retrieving complete evidence requires longer and more structured traversal.

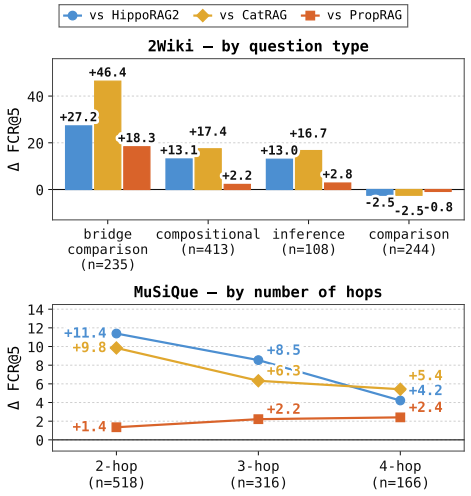

Figure 4: ProST–baseline performance gap by question type and hop count.

Evidence exploration. Figure 5 examines how performance changes as beam search exposes a broader candidate region. Both methods benefit from beam seeding, but ProST maintains its performance over wider beams, whereas PropRAG tends to degrade as the candidate region expands. The gains also persist without beam search, showing that they do not depend on stronger beam-selected seeds. These results suggest that proposition-state traversal can better leverage a broader candidate region for evidence retrieval.

Cost and latency. Table 4 compares offline indexing cost and query-time latency. ProST reuses the PropRAG index, requiring no additional LLM-based preprocessing. It also requires no query-time LLM calls, with proposition-state traversal adding only modest computational overhead over PropRAG. Thus, the retrieval gains require no additional LLM cost. Offline cost measures LLM tokens consumed during indexing, while latency measures retrieval-only wall-clock time with caches disabled.

Table 4: Indexing and query-time costs.

	Offline / 1K psg		Online / query	
Method	In tok	Out tok	Latency (s)	LLM calls
RAPTOR	177K	56K	0.12	0
SiReRAG	1,614K	384K	0.13	0
LightRAG	4,978K	799K	2.25	1
HippoRAG	942K	335K	1.25	1
HippoRAG 2	942K	335K	2.34	1
PropRAG	1,449K	423K	1.30	0
CatRAG	2,212K	430K	14.53	4.8
LinearRAG	0	0	1.27	0
MemGraphRAG	11,306K	1,832K	2.42	1
ProST	1,449K	423K	1.79	0

5 Conclusion

We revisit the traversal state in graph-based retrieval, where entity states do not directly retain the context through which they become relevant. To address this limitation, we propose ProST, a proposition-state graph traversal framework that preserves source context and uses it to guide query-adaptive transitions through bridge entities. Experiments and analyses demonstrate the effectiveness of proposition states and query-adaptive transitions, with the two transition weightings capturing complementary signals. These benefits persist with longer reasoning chains and broader evidence exploration, without additional training or query-time LLM calls.

AI USE STATEMENT

LLMs were used to proofread and polish the writing of this paper and to assist with code development and debugging. All AI-assisted text and code were reviewed and verified by the authors, who take full responsibility for the final manuscript and implementation.

REFERENCES

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162. PMLR, 2022. URL https://arxiv.org/abs/2112.04426.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. Dense X retrieval: What retrieval granularity should we use? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15159–15177, 2024. doi: 10.18653/v1/2024.emnlp-main.845. URL https://aclanthology.org/2024.emnlp-main.845/.

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. URL https://arxiv.org/abs/2404.16130.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023. URL https://arxiv.org/abs/2312.10997.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

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, 2024. URL https://arxiv.org/abs/2405.14831.

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, pp. 21497–21515. PMLR, 2025. URL https://arxiv.org/abs/2502.14802.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 3929–3938. PMLR, 2020. URL https://arxiv.org/abs/2002.08909.

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

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020. URL https://arxiv.org/abs/2011.01060.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022. URL https://arxiv.org/abs/2112.09118.

Jia-Huei Ju, Suzan Verberne, Maarten de Rijke, and Andrew Yates. Controlled retrieval-augmented context evaluation for long-form RAG. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 21102–21121. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1151. URL https://aclanthology.org/2025.findings-emnlp.1151/.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, 2023. URL https://arxiv.org/abs/2309.06180.

Kwun Hang Lau, Fangyuan Zhang, Boyu Ruan, Yingli Zhou, Qintian Guo, Ruiyuan Zhang, and Xiaofang Zhou. Breaking the static graph: Context-aware traversal for graph-based RAG. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 5849–5863. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.290. URL https://aclanthology.org/2026.findings-acl.290/.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-Embed: Improved techniques for training LLMs as generalist embedding models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2405.17428.

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, 2020. URL https://arxiv.org/abs/2005.11401.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023. URL https://arxiv.org/abs/2308.03281.

Linhao Luo, Zicheng Zhao, Gholamreza Haffari, Dinh Phung, Chen Gong, and Shirui Pan. GFM-RAG: Graph foundation model for retrieval augmented generation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2502.01113.

Linhao Luo, Zicheng Zhao, Junnan Liu, Zhangchi Qiu, Junnan Dong, Serge Panev, Chen Gong, Thuy-Trang Vu, Alan Wee-Chung Liew, Gholamreza Haffari, Dinh Phung, and Shirui Pan. G-Reasoner: Foundation models for unified reasoning over graph-structured knowledge. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.24276.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://arxiv.org/abs/2305.14251.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative representational instruction tuning. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2402.09906.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9844–9855, 2022. doi: 10.18653/v1/2022.emnlp-main.669. URL https://aclanthology.org/2022.emnlp-main.669/.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999. URL http://ilpubs.stanford.edu:8090/422/.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. Graph retrieval-augmented generation: A survey. *arXiv preprint arXiv:2408.08921*, 2024. URL https://arxiv.org/abs/2408.08921.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023. URL https://arxiv.org/abs/2302.00083.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2401.18059.

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*, 10:539–554, 2022. URL https://arxiv.org/abs/2108.00573.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10014–10037, 2023. URL https://arxiv.org/abs/2212.10509.

Jingjin Wang and Jiawei Han. PropRAG: Guiding retrieval with beam search over proposition paths. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6212–6227. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.317. URL https://aclanthology.org/2025.emnlp-main.317/.

Geemi P. Wellawatte, Huixuan Guo, Magdalena Lederbauer, Anna Borisova, Matthew Hart, Marta Brucka, and Philippe Schwaller. ChemLit-QA: A human evaluated dataset for chemistry RAG tasks. *Machine Learning: Science and Technology*, 6(2):020601, 2025. doi: 10.1088/2632-2153/adc2d6. URL https://doi.org/10.1088/2632-2153/adc2d6.

Chuanjie Wu, Zhishang Xiang, Yunbo Tang, Zerui Chen, Qinggang Zhang, and Jinsong Su. MemGraphRAG: Memory-based multi-agent system for graph retrieval-augmented generation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, 2026. doi: 10.1145/3770855.3818074. URL https://arxiv.org/abs/2606.00610.

An Yang et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025. URL https://arxiv.org/abs/2412.15115.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018. URL https://arxiv.org/abs/1809.09600.

Nan Zhang, Prafulla Kumar Choubey, Alexander Fabbri, Gabriel Bernadett-Shapiro, Rui Zhang, Prasenjit Mitra, Caiming Xiong, and Chien-Sheng Wu. SiReRAG: Indexing similar and related information for multihop reasoning. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2412.06206.

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

Luyao Zhuang, Shengyuan Chen, Yilin Xiao, Huachi Zhou, Yujing Zhang, Hao Chen, Qinggang Zhang, and Xiao Huang. LinearRAG: Linear graph retrieval augmented generation on large-scale corpora. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.10114.

Appendix

TABLE OF CONTENTS

A Datasets and Proposition Graphs **15**
A.1 Corpus and Graph Statistics 15
A.2 Proposition Extraction and Entity Roles 15
A.3 Proposition Quality 17

B Additional Component Analysis **18**
B.1 Design of the Transition Weightings 18
B.2 Passage and Proposition Channels 18
B.3 Relevance Transfer at a Bridge Entity 19

C Additional Performance Analysis **20**
C.1 Additional Benchmarks and R-Precision 20
C.2 Comparison with Training-Based Retrievers 20
C.3 Coverage-Based Retrieval Evaluation 20
C.4 Evidence Highlights 21

D Sensitivity Analysis **22**
D.1 Encoder 22
D.2 LLM 22
D.3 Hyperparameters 22

E Convergence of Proposition-State PPR **24**

F Case Study **25**

G Prompts **32**

A Datasets and Proposition Graphs

A.1 Corpus and Graph Statistics

Table 5 summarizes the benchmark corpora and their proposition graphs. We use the 1,000-question subsets and corpora from HippoRAG 2 and PropRAG for the multi-hop benchmarks, and all 1,025 questions and 1,464 passages for ChemLit-QA. Across datasets, 35–45% of entities occur in more than one proposition and thus provide connections between proposition states.

Table 5: Corpus and proposition-graph statistics. Prop: proposition; Ent: entity.

Dataset	#Questions	#Passages	#Propositions	#Entities	Prop/passage	Ent/prop
MuSiQue	1,000	11,656	59,605	75,693	5.1	3.5
2Wiki	1,000	6,119	30,114	43,140	4.9	3.5
HotpotQA	1,000	9,811	53,613	74,578	5.5	3.6
ChemLit-QA	1,025	1,464	11,518	15,606	7.9	4.1

A.2 Proposition Extraction and Entity Roles

Propositions and entities. Following PropRAG (Wang & Han, 2025), Llama-3.3-70B-Instruct decomposes each passage into atomic, self-contained propositions and extracts their entities. We use the extracted propositions and entities without additional filtering or post-processing. Figure 6 illustrates how a passage is decomposed into proposition states connected through shared entities.

Entity roles. Entity roles are derived from the OpenIE triples (s, r, o) already produced during indexing, requiring no additional LLM calls. For each proposition–entity pair (p, e), we define $r_{p,e}$ from relations whose arguments occur in p, using r for the object and its inverse for the subject. When multiple roles are available, query-aware entity weighting uses the best-matching role. Links not covered by OpenIE retain their graph connections but receive an entity-weighting score of 0. Examples 1 and 2 show how roles vary with the proposition context, including inverse and missing roles, while Table 6 reports role coverage across datasets.

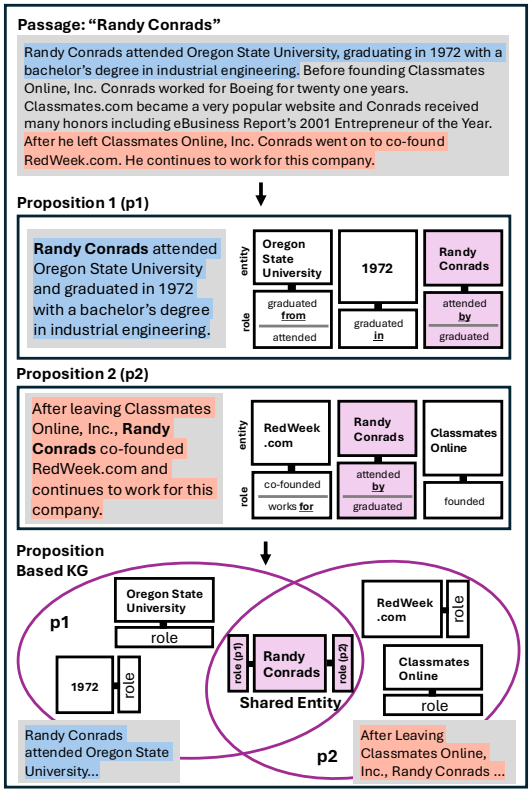

Figure 6: Example of proposition graph construction.

Example 1: Proposition extraction from the passage "Randy Conrads"

Randy Conrads attended Oregon State University, graduating in 1972 with a bachelor's degree in industrial engineering. Before founding Classmates Online, Inc. Conrads worked for Boeing for twenty one years. Classmates.com became a very popular website and Conrads received many honors including eBusiness Report's 2001 Entrepreneur of the Year. After he left Classmates Online, Inc. Conrads went on to co-found RedWeek.com. He continues to work for this company.

	Proposition	**Entity**	**Role** $r_{p,e}$
p_1	Randy Conrads attended Oregon State University and graduated in 1972 with a bachelor's degree in industrial engineering.	1972 Randy Conrads Oregon State University industrial engineering	graduated in attended by, graduated graduated from, attended —
p_2	Randy Conrads worked for Boeing for twenty-one years before founding Classmates Online, Inc.	Randy Conrads Boeing Classmates Online, Inc.	founded by, worked, employed worked for, employed by founded
p_3	Randy Conrads co-founded Classmates Online, Inc. and its website, Classmates.com, became very popular.	Randy Conrads Classmates Online, Inc. Classmates.com	founded by owned by, founded owns
p_4	Randy Conrads received the eBusiness Report's 2001 Entrepreneur of the Year award.	Randy Conrads eBusiness Report 2001 Entrepreneur of the Year	received award received award from — —
p_5	After leaving Classmates Online, Inc., Randy Conrads co-founded RedWeek.com and continues to work for this company.	Randy Conrads Classmates Online, Inc. RedWeek.com	founded by, co-founded by, works founded co-founded, works for

Example 2: Proposition extraction from the passage "Benny Beaver"

Benny Beaver is the official mascot of Oregon State University and winner of the 2011 Capital One Mascot of the Year write-in campaign. The exact date of when the name was first used as the university's mascot is not known, but photographs in the school's yearbook document its use as early as the 1940s.

	Proposition	**Entity**	**Role** $r_{p,e}$
q_1	Benny Beaver is the official mascot of Oregon State University.	Oregon State University Benny Beaver	is mascot of, has mascot by has mascot, is mascot
q_2	Benny Beaver won the 2011 Capital One Mascot of the Year write-in campaign.	2011 Benny Beaver Capital One Mascot of the Year	won in won by, won won
q_3	The name 'Benny Beaver' was first used as the mascot of Oregon State University as early as the 1940s, as documented in the school's yearbook.	Oregon State University Benny Beaver 1940s	is mascot of, has mascot by has mascot, is mascot —

Table 6: Coverage of entity roles derived from OpenIE triples. Prop: proposition.

Dataset	Props with $\geq$1 role	Prop–entity links with a role	Roles / prop	Unique role strings
MuSiQue	82.0%	71.0%	3.8	34,954
2Wiki	84.4%	72.0%	3.6	12,096
HotpotQA	83.6%	72.9%	3.8	28,612
ChemLit-QA	34.9%	23.4%	1.3	3,519

A.3 Proposition Quality

Because the traversal operates on propositions, its states are only as reliable as the extraction. Following Dense X Retrieval (Chen et al., 2024), we therefore check three properties of the extracted propositions—faithfulness to the source, minimality, and stand-alone interpretability—and additionally measure how much of each source sentence survives extraction. For each Wikipedia benchmark, we sample 300 propositions and 300 sentences and judge them with GPT-4o-mini at temperature 0. As shown in Table 7, the error rate is at most 1% for each property, and 85.0–88.7% of source sentences are fully covered, with the remainder mostly partially covered.

Table 7: Proposition quality (%; 300 propositions and sentences per benchmark).

	Error rate ↓			Sentence coverage		
Dataset	Not faithful	Not minimal	Not stand-alone	Full	Partial	None
MuSiQue	0.67	0.67	0.00	88.67	9.67	1.67
2Wiki	0.67	1.00	0.00	85.00	11.67	3.33
HotpotQA	0.67	0.33	0.00	88.67	8.33	3.00

B Additional Component Analysis

B.1 Design of the Transition Weightings

Table 8 compares alternative scoring functions for the two transition weightings, varying one at a time while keeping the other fixed.

Query-aware entity weighting. Directly scoring entity embeddings against the query performs worse than uniform weighting, whereas scoring entity roles against the query consistently performs best. This suggests that the relational role of an entity within the current proposition provides a more useful signal for choosing the next traversal direction than entity identity alone. Using the residual query for entity weighting is less effective, indicating that the full query is more informative at this stage.

Source-aware proposition weighting. Source-blind query–proposition similarity performs worst, while excluding the current proposition already improves retrieval. Conditioning explicitly on the source through the residual $z_q - z_{p_t}$ performs best across all three benchmarks, outperforming both masking and cumulative alternatives. Together, the two analyses support a complementary design in which the full query identifies *which bridge entity to follow*, while the residual query identifies *which proposition to visit next*.

Table 8: Alternative scoring functions for the two weightings (FCR@5).

	Variant	Score	MuSiQue	2Wiki	HotpotQA
Entity weighting	Uniform	1	55.4	85.8	94.7
	Query–entity	$\mathrm{sim}(z_q, z_e)$	53.7	83.7	94.7
	Residual–entity	$\mathrm{sim}(z_q - z_{p_t}, z_e)$	54.2	84.4	94.7
	Residual–role	$\mathrm{sim}(z_q - z_{p_t}, z_{r_{p_t,e}})$	54.7	85.3	94.7
	Query–role (ProST)	$\mathrm{sim}(z_q, z_{r_{p_t,e}})$	**55.6**	**88.4**	**95.0**
Proposition weighting	Source-blind	$\mathrm{sim}(z_q, z_{p'})$	52.8	85.7	93.7
	Self-mask	$\mathrm{sim}(z_q, z_{p'}),\ p' \neq p_t$	54.3	88.0	94.6
	Cumulative	$\mathrm{sim}(z_q, z_{p_t} + z_{p'})$	55.0	87.8	94.2
	Weighted cumulative	$\mathrm{sim}(z_q, 0.3z_{p_t} + 0.7z_{p'})$	53.6	86.5	94.0
	Residual (ProST)	$\mathrm{sim}(z_q - z_{p_t}, z_{p'})$	**55.6**	**88.4**	**95.0**

B.2 Passage and Proposition Channels

Our final passage score combines proposition-level relevance aggregated by passage with passage-state relevance. Table 9 compares the direct sum with either channel alone and with normalized late fusion. The proposition channel accounts for most of the retrieval performance, while adding the passage channel further improves the average result. In contrast, normalizing the two channels before fusion degrades performance, indicating that preserving their relative score magnitudes is important. The routing weight β controls how much relevance passes through passage states and is selected on validation. Performance varies by at most 0.3 points of mean validation FCR@5 over $\beta \in [0.05, 0.4]$ (Figure 10), indicating low sensitivity within this range.

Table 9: Ablation of final passage scoring (FCR@5).

Scoring	Score $s(d)$	MuSiQue	2Wiki	HotpotQA	Avg.
Passage only	$\pi_D(d)$	43.3	63.9	91.9	66.4
Proposition only	$\sum_{p \in P(d)} \pi_P(p)$	54.2	**89.0**	93.4	78.9
Normalized fusion	$0.5\,\tilde{s}_{\mathrm{prop}} + 0.5\,\tilde{s}_{\mathrm{psg}}$	52.4	84.3	94.4	77.0
Direct sum (ProST)	$\sum_{p \in P(d)} \pi_P(p) + \pi_D(d)$	**55.6**	88.4	**95.0**	**79.7**

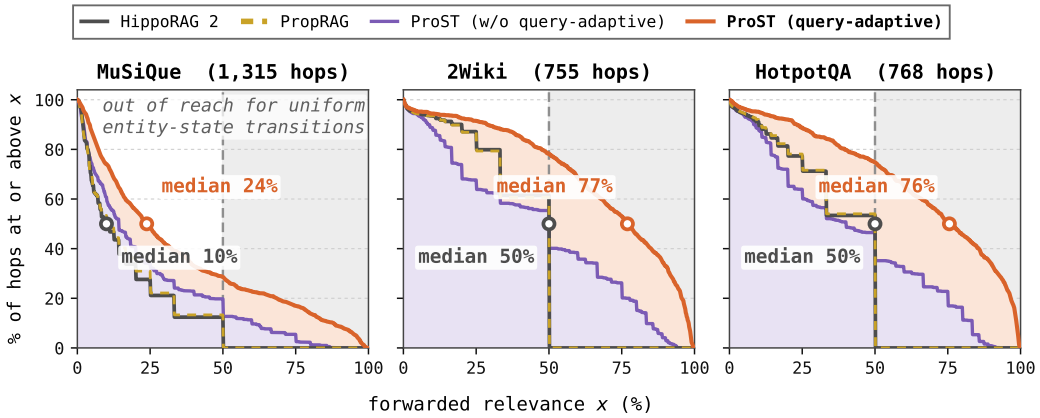

Figure 7: One-step relevance transfer toward the next gold passage.

B.3 Relevance Transfer at a Bridge Entity

To examine how preserved source context guides query-adaptive transitions, we measure how much relevance is transferred across a bridge entity toward the next gold passage.

Measurement. For each pair of consecutive gold passages, we identify their shared bridge entities and measure the fraction of outgoing relevance that reaches the next gold passage in one transition. We compare entity-state traversal with proposition-state variants that progressively incorporate query and source context. For a conservative comparison, entity-state baselines use their best bridge entity and are evaluated only over passage-directed transitions.

Source context enables targeted transfer. Uniform entity-state transitions distribute relevance across all passages connected to a bridge entity, limiting how much can reach the next gold passage. Simply replacing entities with proposition states does not overcome this limitation, and query-only weighting provides little additional benefit. In contrast, ProST uses the source proposition to prioritize complementary evidence, substantially increasing relevance transferred toward the next gold passage. This confirms that the benefit of proposition states comes from retaining source context that can guide subsequent transitions, rather than from changing the state space alone.

Table 10: Median one-step relevance transferred toward the next gold passage (%).

System	State	Transition	MuSiQue	2Wiki	HotpotQA	Avg.
HippoRAG 2	Entity	Uniform	10.0	50.0	50.0	36.7
PropRAG	Entity	Uniform	11.1	50.0	50.0	37.0
ProST	Proposition	Query + source	**23.8**	**76.9**	**75.6**	**58.8**
w/o source-aware weighting	Proposition	Query only	16.8	46.4	43.0	35.4
w/o query-adaptive transition	Proposition	Uniform	14.3	50.0	40.0	34.8

Scope. This diagnostic uses gold annotations to identify bridge entities and target passages, and measures a single transition rather than end-to-end retrieval. It is intended to isolate how source context affects relevance transfer at a bridge entity.

C Additional Performance Analysis

C.1 Additional Benchmarks and R-Precision

R@5 and FCR@5 use a fixed cutoff despite variation in the number of gold passages across questions. We therefore additionally report R-Precision (RP), i.e., recall at R where R is the number of gold passages of each question, and its full-chain variant (full-chain RP), which requires all gold passages to appear within the top R. Table 11 also extends the evaluation to HoVer and MultiHop-RAG. ProST achieves the best RP and full-chain RP on MuSiQue, 2Wiki, and HoVer, and the best RP on MultiHop-RAG. The gains in full-chain RP are particularly clear on MuSiQue and 2Wiki, showing that the improvements in complete-chain retrieval remain robust under a question-specific cutoff.

Table 11: R-Precision (RP) and full-chain RP (%). Best in **bold**, second best underlined.

	MuSiQue		2Wiki		HotpotQA		HoVer		MultiHop-RAG	
Method	RP	Full-chain RP	RP	Full-chain RP	RP	Full-chain RP	RP	Full-chain RP	RP	Full-chain RP
HippoRAG 2	61.6	25.7	83.2	60.3	83.4	68.6	81.2	44.8	58.8	**27.4**
PropRAG	61.9	26.7	84.2	63.4	83.5	69.0	83.1	49.5	59.0	26.5
CatRAG	63.2	28.3	80.4	54.1	**85.5**	**72.5**	82.9	49.3	53.2	19.7
ProST	**64.7**	**30.7**	**87.0**	**69.7**	84.2	70.3	**84.5**	**53.1**	**59.5**	27.2

C.2 Comparison with Training-Based Retrievers

GFM-RAG (Luo et al., 2025) and G-Reasoner (Luo et al., 2026) train graph retrievers on the Wikipedia multi-hop benchmarks. We exclude them from the main comparison because our primary evaluation focuses on training-free retrieval methods, but report their performance under the same evaluation protocol in Table 12. Within their training domain the trained retrievers are competitive, G-Reasoner is the best system on 2Wiki, whereas ProST is best on MuSiQue and HotpotQA without any task-specific training. On the out-of-domain ChemLit-QA benchmark, both trained retrievers degrade sharply, while ProST remains competitive without retraining, demonstrating its transfer to a new domain.

Table 12: Downstream QA performance of training-based retrievers (%). Best in **bold**.

	MuSiQue			2Wiki			HotpotQA			ChemLit-QA		
Method	EM	F1	Acc	EM	F1	Acc	EM	F1	Acc	EM	F1	Acc
GFM-RAG	28.7	37.4	32.6	65.5	72.7	69.2	56.3	68.0	64.4	6.1	37.6	25.1
G-Reasoner	42.5	52.9	47.7	**70.7**	**78.5**	**75.3**	62.8	75.7	71.9	17.4	56.5	52.0
ProST	**45.3**	**56.3**	**50.8**	68.9	76.8	73.6	**63.2**	**76.4**	**73.0**	**18.1**	**60.3**	**58.1**

C.3 Coverage-Based Retrieval Evaluation

Passage-level recall can underestimate retrieval quality when the same evidence appears outside the annotated gold passages. We therefore additionally evaluate information coverage using CRUX (Ju et al., 2025), which measures whether the top-k passages contain the evidence needed for each sub-question. We use the released decomposition for MuSiQue, generate sub-questions from gold supporting facts for 2Wiki and HotpotQA, and use the original question for single-hop ChemLit-QA. An LLM grader (Llama-3.3-70B-Instruct) evaluates each passage–sub-question pair, from which we report *Cov@5* for evidence coverage and α-*nDCG@5* for coverage-aware ranking diversity. As shown in Table 13, ProST achieves the highest or near-highest scores across benchmarks, indicating that it retrieves both sufficient and complementary evidence.

Table 13: CRUX@5: Cov@5 and α-nDCG@5 (%). Best in **bold**, second best underlined.

	Method	MuSiQue		2Wiki		HotpotQA		ChemLit-QA	
		Cov@5	α-nDCG@5	Cov@5	α-nDCG@5	Cov@5	α-nDCG@5	Cov@5	α-nDCG@5
Dense	Contriever	65.8	57.6	68.6	60.4	86.7	78.1	92.0	80.3
	GTR	67.0	61.3	77.5	71.0	86.4	78.0	94.6	88.0
	GTE-Qwen2-7B	79.0	72.6	82.5	77.2	95.5	89.1	94.4	87.8
	GritLM-7B	80.4	74.6	83.5	77.7	96.7	91.0	98.0	93.7
	NV-Embed-v2	82.9	76.9	83.6	78.7	97.3	92.2	99.0	94.9
Structure-aware	Dense X	77.9	70.4	79.1	74.7	94.7	88.5	99.2	96.4
	RAPTOR	75.0	69.3	73.3	68.7	94.8	88.6	97.9	92.5
	HippoRAG	66.7	61.0	94.7	84.8	87.0	79.7	61.4	51.9
	SiReRAG	64.8	61.1	66.9	63.2	91.5	85.6	94.4	89.8
	LightRAG	62.0	54.4	64.7	60.4	85.9	76.6	88.7	78.2
Query-aware	HippoRAG 2	86.2	79.4	94.2	86.3	98.1	92.7	99.5	94.4
	PropRAG	89.0	81.7	96.1	87.5	98.9	**93.4**	99.7	96.3
	CatRAG	82.6	73.1	89.6	82.8	97.4	91.3	96.8	87.8
	QAFD-RAG	75.3	64.8	86.9	76.3	93.1	85.1	82.1	66.7
	LinearRAG	67.5	57.3	85.6	76.4	77.6	70.1	91.4	81.4
	MemGraphRAG	81.9	71.9	92.7	85.0	97.3	90.5	96.4	83.6
	ProST	**89.9**	**82.8**	**97.2**	**88.3**	**99.0**	93.3	**99.8**	**97.0**

C.4 Evidence Highlights

For downstream QA, we prepend the N propositions from the retrieved passages that are most similar to the query as evidence highlights, formatted as shown in Appendix G. These highlights surface information already contained in the retrieved context rather than introducing additional evidence. Figure 8 compares evidence highlights with additional passages under matched context budgets. Even without highlights, ProST matches or exceeds PropRAG, showing that the retrieval gains transfer to downstream QA independently of the highlights. Adding a small number of highlights further improves performance with substantially fewer tokens than additional passages. The gains saturate as N increases, suggesting that a small number of highlights is sufficient to guide the LLM when processing the retrieved passages. We select $N{=}3$ on the validation split and use it for all reported results.

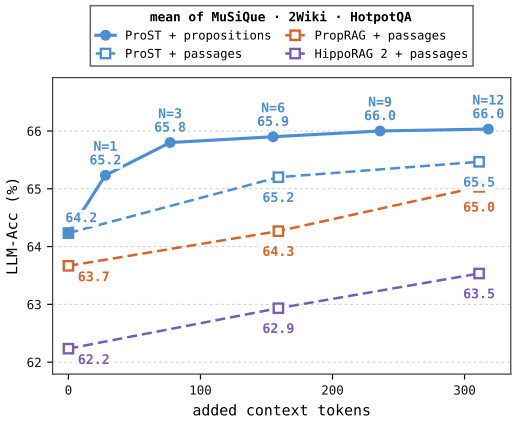

Figure 8: LLM-Acc under matched context budgets.

D SENSITIVITY ANALYSIS

D.1 ENCODER

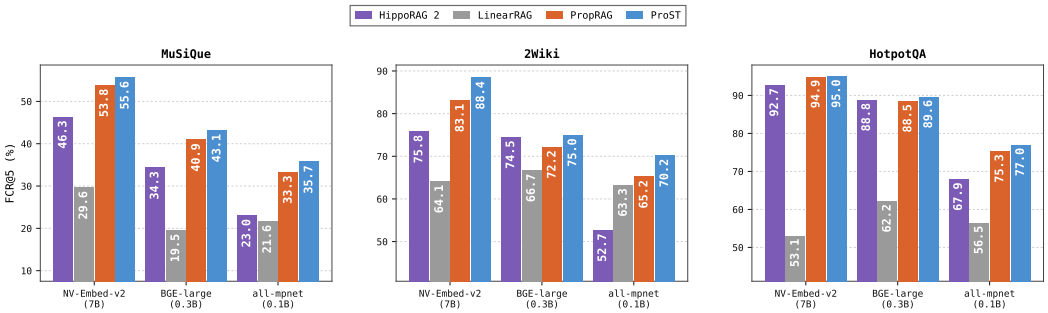

Figure 9: FCR@5 across different shared encoders. All embedding components use the same encoder within each setting, with hyperparameters fixed to those selected for NV-Embed-v2.

Figure 9 evaluates HippoRAG 2, LinearRAG, PropRAG, and ProST with NV-Embed-v2, BGE-large, and all-mpnet-base. All embeddings in each pipeline are replaced by the same encoder while keeping hyperparameters fixed. Although performance decreases with smaller encoders, ProST achieves the best FCR@5 across all encoder–benchmark combinations. The gains over PropRAG persist across encoders, indicating that the adaptive transitions are not specific to NV-Embed-v2.

D.2 LLM

The main experiments use Llama-3.3-70B-Instruct for downstream QA. Table 14 evaluates robustness across LLMs using Llama-3.1-8B-Instruct and GPT-4o-mini, while keeping the retrieved passages, prompt, and judge fixed. We additionally include LinearRAG and MemGraphRAG, whose official implementations use GPT-4o-mini for downstream QA.

For ProST, the number of evidence highlights N is re-selected for each LLM on the validation split using the same rule as in the main experiments. Evidence highlights are selected for both Llama models, whereas the validation rule selects passage-only input for GPT-4o-mini. The effect of highlights under matched context budgets is analyzed in Appendix C.4.

ProST achieves the best average EM, F1, and LLM-Acc across all three LLMs. The gains are most consistent on MuSiQue and 2Wiki, while the differences are smaller on HotpotQA, where retrieval performance is already saturated. LinearRAG and MemGraphRAG also remain below ProST with GPT-4o-mini, showing that the improvements are robust across LLMs.

D.3 HYPERPARAMETERS

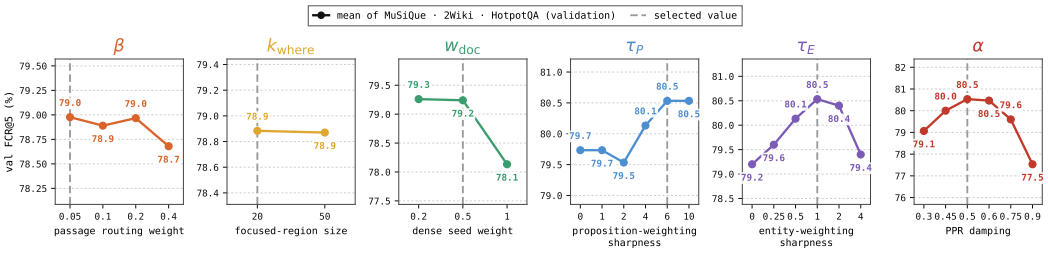

Figure 10: Validation FCR@5 as each hyperparameter varies with the others fixed at their selected values (dashed line). Values are averaged across the three multi-hop benchmarks.

Hyperparameter selection. ProST uses three retrieval-side hyperparameters—the passage routing weight β, dense seed weight w_{doc}, and focused-region size k_{where}—and three traversal-side hyperparameters: the proposition and entity sharpness parameters τ_P and τ_E, and PPR damping factor α. Table 15 reports the values selected on the validation split.

Table 14: Downstream QA across LLMs (%). Acc: LLM-Acc. Best in **bold**, second best underlined within each LLM.

	Method	MuSiQue			2Wiki			HotpotQA			Average		
		EM	F1	Acc	EM	F1	Acc	EM	F1	Acc	EM	F1	Acc
Llama-3.3 70B	HippoRAG 2	38.8	50.7	45.4	65.7	72.4	69.9	62.8	75.5	71.4	55.8	66.2	62.2
	PropRAG	41.5	52.6	46.9	68.3	75.7	72.2	**63.3**	76.3	71.9	57.7	68.2	63.7
	CatRAG	38.9	50.2	45.1	64.2	70.6	67.9	62.3	75.4	71.1	55.1	65.4	61.4
	LinearRAG	29.7	39.6	34.6	59.5	66.0	62.8	50.2	62.4	58.1	46.5	56.0	51.8
	MemGraphRAG	37.7	48.9	44.1	65.8	72.6	69.9	62.6	75.5	71.5	55.4	65.7	61.8
	ProST (N=3)	**45.3**	**56.3**	**50.8**	**68.9**	**76.8**	**73.6**	63.2	**76.4**	**73.0**	**59.1**	**69.8**	**65.8**
Llama-3.1 8B	HippoRAG 2	32.3	44.3	38.0	60.8	68.6	65.6	59.9	72.3	67.8	51.0	61.7	57.1
	PropRAG	35.1	47.0	39.3	60.0	68.8	65.6	**60.8**	73.2	**68.8**	52.0	63.0	57.9
	CatRAG	32.5	44.2	37.8	58.0	66.3	62.6	59.3	71.5	67.3	49.9	60.6	55.9
	LinearRAG	24.7	34.3	28.0	54.3	62.2	59.0	45.4	57.3	52.3	41.5	51.3	46.4
	MemGraphRAG	32.2	44.0	37.3	58.8	66.9	63.3	58.9	71.7	67.3	50.0	60.9	56.0
	ProST (N=3)	**37.2**	**49.9**	**43.2**	**62.6**	**71.5**	**68.2**	60.0	**73.4**	**68.8**	**53.3**	**64.9**	**60.1**
GPT-4o mini	HippoRAG 2	35.3	49.3	42.4	60.1	69.6	66.8	58.8	73.6	69.0	51.4	64.2	59.4
	PropRAG	**39.3**	53.3	47.1	62.5	72.8	69.9	59.5	74.7	**70.1**	53.8	66.9	62.4
	CatRAG	34.5	49.2	42.1	58.3	67.7	64.7	58.9	73.9	70.0	50.6	63.6	58.9
	LinearRAG	28.4	40.9	34.6	54.8	63.8	60.9	45.2	59.4	54.5	42.8	54.7	50.0
	MemGraphRAG	34.8	48.7	43.3	60.4	70.1	66.8	58.1	73.4	68.7	51.1	64.1	59.6
	ProST (N=0)	39.0	**53.8**	**47.3**	**64.2**	**73.9**	**71.2**	**60.5**	**75.1**	70.0	**54.6**	**67.6**	**62.8**

Sensitivity. Figure 10 shows that retrieval-side parameters are relatively insensitive around their selected values, whereas the traversal parameters exhibit clearer optima. In particular, τ_P and τ_E control the sharpness of Source-Aware Proposition Weighting and Query-Aware Entity Weighting, respectively, while excessive damping degrades performance by moving relevance farther from the seeds. Overall, performance remains stable around the selected configuration.

Table 15: Selected hyperparameters. Beam search uses width 4 and depth 3.

Dataset	β	w_{doc}	k_{where}	τ_P	τ_E	α
MuSiQue	0.05	0.5	20	6	1	0.5
2Wiki	0.05	0.2	20	2	1	0.5
HotpotQA	0.05	0.5	20	6	1	0.5
ChemLit-QA	0.2	1.0	20	6	1	0.5

Focused region. Applying the adaptive transitions to every proposition would be wasteful, since most carry negligible relevance. Retrieval therefore first runs a short pass with uniform transitions and keeps the k_{where}=20 propositions with the highest relevance as the region in which the traversal operates. Within this region, the adaptive transitions are applied to the 50 entities that currently carry the most relevance and, for each of them, to the 8 source propositions contributing the most relevance, with the candidate list of each entity capped at 200 propositions. Mass excluded by these truncations follows the uniform weights, preserving proper transition distributions. These limits reduce adaptive scoring while retaining uniform transitions outside the focused region.

E CONVERGENCE OF PROPOSITION-STATE PPR

Valid Transition Matrix. For a fixed query q, ProST defines a PPR process over proposition states with query- and source-dependent transitions. Both transition factors are softmax-normalized: $T_{P\to E}(e \mid p, q)$ over entities $e \in E(p)$, and $T_{E\to P}(p' \mid p, e, q)$ over propositions $p' \in P(e)$. The resulting proposition transition is

$$T(p' \mid p, q) = \sum_{e \in E(p) \cap E(p')} T_{P\to E}(e \mid p, q) T_{E\to P}(p' \mid p, e, q).$$

Since both factors are non-negative and normalized over their respective candidate sets,

$$\sum_{p'} T(p' \mid p, q) = \sum_{e \in E(p)} T_{P\to E}(e \mid p, q) \sum_{p' \in P(e)} T_{E\to P}(p' \mid p, e, q) = 1.$$

Thus, the proposition-state transition is row-stochastic. With passage integration, each proposition routes fractions $1 - \beta$ and β through proposition and passage transitions, respectively, while each passage redistributes its relevance over its propositions. Hence, the joint transition matrix $\mathbf{T}_q$ is also row-stochastic. The focused-region approximation does not affect this property, since mass outside the adaptively scored region follows normalized uniform transitions.

Convergence. Given restart probability $1 - \alpha$ with $0 < \alpha < 1$, the joint relevance distribution is updated as

$$F(\boldsymbol{\pi}) = (1 - \alpha)\mathbf{r} + \alpha \mathbf{T}_q^\top \boldsymbol{\pi}.$$

Because $\mathbf{T}_q$ is row-stochastic, $\mathbf{T}_q^\top$ is non-expansive under the ℓ_1 norm. Therefore, for any $\mathbf{x}$ and $\mathbf{y}$,

$$\|F(\mathbf{x}) - F(\mathbf{y})\|_1 \le \alpha \|\mathbf{x} - \mathbf{y}\|_1.$$

Since $\alpha < 1$, F is a contraction and converges to the unique fixed point

$$\boldsymbol{\pi}^* = (1 - \alpha)(\mathbf{I} - \alpha \mathbf{T}_q^\top)^{-1}\mathbf{r}.$$

where $\mathbf{I}$ is the identity matrix.

Query and Source Adaptivity. Query and source adaptivity do not alter this convergence result. For a fixed query q, $T(p' \mid p, q)$ is a fixed function of the source proposition p and does not depend on the iteration index. Thus, each query induces a fixed joint transition matrix $\mathbf{T}_q$, and the process remains a first-order Markov chain. Source-aware weighting therefore changes the transition matrix being iterated, but not the convergence property of PPR.

F CASE STUDY

The six cases below compare HippoRAG 2, PropRAG, CatRAG, and ProST on six questions, following the presentation of LinearRAG (Zhuang et al., 2026), with one question per hop count on MuSiQue (2-, 3-, and 4-hop) and one per question type on 2Wiki (compositional, inference, and bridge comparison). We omit the 2Wiki comparison type, where all four systems reach FCR@5 of 97.5–100% and the differences lie in the LLM. Each case lists the question and its gold chain, then the top-5 passages of each system (gold passages in **bold** with ✓) and the LLM's prediction. The paragraph below each case reads the outcome and, for ProST, traces the transition that retrieves the missing hop by reporting the source proposition, bridge entity, and source-aware scores of the next-hop candidates computed from the stored proposition embeddings.

Case 1 (MuSiQue, 2-hop): one entity, two contexts (a band and a film character)

Question

"Who played Seether's performer in Willy Wonka and the Chocolate Factory?"
Ground truth: Julie Dawn Cole
Gold chain: [Seether → performer → Veruca Salt]
[Veruca Salt (character) → played by → Julie Dawn Cole]

Baselines

HippoRAG 2
Retrieved context:
1) ✓ **Seether (song)**
2) × Willy Wonka & the Chocolate Factory
3) × Willy Wonka & the Chocolate Factory
4) × Willy Wonka & the Chocolate Factory
5) × Willy Wonka & the Chocolate Factory
Prediction:
× None

PropRAG
Retrieved context:
1) ✓ **Seether (song)**
2) × Willy Wonka & the Chocolate Factory
3) × Willy Wonka & the Chocolate Factory
4) × Willy Wonka & the Chocolate Factory
5) × Willy Wonka & the Chocolate Factory
Prediction:
× None

CatRAG
Retrieved context:
1) ✓ **Seether (song)**
2) × Willy Wonka & the Chocolate Factory
3) × Willy Wonka & the Chocolate Factory
4) × Willy Wonka & the Chocolate Factory
5) × Willy Wonka & the Chocolate Factory
Prediction:
× None

ProST

Retrieved context:
1) × Willy Wonka & the Chocolate Factory
2) ✓ **Seether (song)**
3) ✓ **Julie Dawn Cole**
4) × Willy Wonka & the Chocolate Factory
5) × Willy Wonka & the Chocolate Factory
Prediction:
✓ **Julie Dawn Cole**

All three baselines retrieve the song passage and fill the remaining four slots with the film passage, and the LLM abstains. Our traversal leaves the source proposition through the bridge entity *Veruca Salt*, which occurs in 15 propositions covering both the band and the film character. Among them the source-aware score ranks "Julie Dawn Cole played the role of Veruca Salt in the 1971 film …" first (+0.12, all others at most +0.01), so the passage naming the actress enters the top-5.

Case 2 (MuSiQue, 3-hop): album to band to city to landmark

Question

"What is the name of the castle in the city where the performer of We Sweat Blood formed?"
Ground truth: Casa Loma
Gold chain: [We Sweat Blood → performer → Danko Jones]
[Danko Jones → formed in → Toronto]
[Toronto → castle → Casa Loma]

Baselines

HippoRAG 2
Retrieved context:
1) ✓ **We Sweat Blood**
2) × The Collection (1993 Blood, Sweat & Tears album)
3) × Paula Santiago
4) × The Sitwells
5) × Jesus Egg That Wept
Prediction:
× None

PropRAG
Retrieved context:
1) ✓ **We Sweat Blood**
2) × Arthur's Seat
3) × Royal Edinburgh Military Tattoo
4) × Unexploded
5) ✓ **Danko Jones**
Prediction:
× Edinburgh Castle

CatRAG
Retrieved context:
1) ✓ **We Sweat Blood**
2) × The Collection (1993 Blood, Sweat & Tears album)
3) × Paula Santiago
4) × The Sitwells
5) × Jesus Egg That Wept
Prediction:
× None

ProST

Retrieved context:
1) ✓ **We Sweat Blood**
2) × Arthur's Seat
3) × Royal Edinburgh Military Tattoo
4) ✓ **Danko Jones**
5) ✓ **Casa Loma**
Prediction:
✓ **Casa Loma**

PropRAG reaches the band but then drifts to Edinburgh, whose passages match the word *castle* directly; the other two baselines do not reach the band at all. Conditioning on "Danko Jones is a Canadian rock trio from Toronto, Ontario" increases the relative relevance of propositions about the castle of that city. Among the 57 propositions attached to *Toronto* the score then ranks Casa Loma first, though only by a small margin ($+0.01$, next -0.02).

Case 3 (MuSiQue, 4-hop): two author hops, a filming location, and a population lookup

Question

"There is a state A where Storm of the Century was filmed based on the book by the author who wrote a short story featuring the author of This Is My God. What is the population of state A?"

Ground truth: 1,335,907

Gold chain: [This Is My God → author → Herman Wouk]
["Herman Wouk Is Still Alive" → author → Stephen King]
[Storm of the Century → filmed in → Maine]
[Maine → population → 1,335,907]

Baselines

HippoRAG 2
Retrieved context:
1) ✓ **Storm of the Century**
2) × Dolores Claiborne (film)
3) ✓ **This Is My God**
4) ✓ **Herman Wouk Is Still Alive**
5) × Stockton Springs, Maine
Prediction:
× Not provided in the given text, but approximately 1.362 million

PropRAG
Retrieved context:
1) ✓ **This Is My God**
2) ✓ **Storm of the Century**
3) ✓ **Herman Wouk Is Still Alive**
4) × Dolores Claiborne (film)
5) × Castle Rock (TV series)
Prediction:
× Approximately 1,328,361

CatRAG
Retrieved context:
1) ✓ **Storm of the Century**
2) × Dolores Claiborne (film)
3) ✓ **This Is My God**
4) ✓ **Herman Wouk Is Still Alive**
5) × Stockton Springs, Maine
Prediction:
× Maine

ProST

Retrieved context:
1) ✓ **Storm of the Century**
2) ✓ **This Is My God**
3) ✓ **Herman Wouk Is Still Alive**
4) ✓ **Maine**
5) × Stockton Springs, Maine
Prediction:
✓ **1,335,907**

All three baselines recover the first three hops and fail only on the last one, and the LLM then produces population figures that are close but wrong. From the source proposition that places the filming location in Maine, the score ranks "The state of Maine has . . . a population of approximately 1,335,907" first among the 48 propositions attached to *Maine* (+0.08, next +0.05).

Case 4 (2Wiki, compositional): namesake entities

Question

"When was John of Ibelin, the Old Lord of Beirut's father born?"
Ground truth: 1143
Gold chain: [John of Ibelin, the Old Lord of Beirut → father → Balian of Ibelin]
[Balian of Ibelin → date of birth → 1143]

Baselines

HippoRAG 2
Retrieved context:
1) ✓ **John of Ibelin, the Old Lord of Beirut**
2) × John II of Beirut
3) × John of Ibelin (jurist)
4) × Guy of Ibelin, constable of Cyprus
5) × Alice de la Roche
Prediction:
× Unknown

PropRAG
Retrieved context:
1) × John of Ibelin (jurist)
2) ✓ **John of Ibelin, the Old Lord of Beirut**
3) × John II of Beirut
4) × Guy of Ibelin, constable of Cyprus
5) × Alice de la Roche
Prediction:
× None

CatRAG
Retrieved context:
1) ✓ **John of Ibelin, the Old Lord of Beirut**
2) × John II of Beirut
3) × John of Ibelin (jurist)
4) × Guy of Ibelin, constable of Cyprus
5) × Alice de la Roche
Prediction:
× Not mentioned in the text

ProST

Retrieved context:
1) ✓ **John of Ibelin, the Old Lord of Beirut**
2) × John of Ibelin (jurist)
3) × John II of Beirut
4) × Guy of Ibelin, constable of Cyprus
5) ✓ **Balian of Ibelin**
Prediction:
✓ **1143**

The corpus contains several people named *John of Ibelin*, and the baselines fill their top-5 with these namesakes. Our traversal follows the *father* relation to *Balian of Ibelin*, whose passage is retrieved even though the proposition carrying the answer ranks second (-0.04, behind a proposition about his son's parents at $+0.04$), because the passage accumulates relevance from several of its propositions.

Case 5 (2Wiki, inference): paternal grandfather through two *father* relations

Question

"Who is Eberhard of Limburg's paternal grandfather?"
Ground truth: Johann of Limburg
Gold chain: [Eberhard of Limburg → father → Dietrich III of Limburg]
[Dietrich III of Limburg-Styrum → father → Johann of Limburg]

Baselines

HippoRAG 2
Retrieved context:
1) ✓ **Eberhard of Limburg**
2) × Wilhelm I of Limburg
3) × Adolf of Limburg
4) × Reginald I of Guelders
5) × Otto of Limburg
Prediction:
× The information is not provided

PropRAG
Retrieved context:
1) ✓ **Eberhard of Limburg**
2) × Wilhelm I of Limburg
3) × Adolf of Limburg
4) × Jobst of Limburg
5) × Otto of Limburg
Prediction:
× The information is not provided

CatRAG
Retrieved context:
1) ✓ **Eberhard of Limburg**
2) × Wilhelm I of Limburg
3) × Reginald I of Guelders
4) × Otto of Limburg
5) × Adolf of Limburg
Prediction:
× None

ProST

Retrieved context:
1) ✓ **Eberhard of Limburg**
2) × Wilhelm I of Limburg
3) × Adolf of Limburg
4) ✓ **Dietrich III of Limburg-Styrum**
5) × Jobst of Limburg
Prediction:
✓ **Johann of Limburg**

The father is written as *Dietrich III of Limburg* in one passage and *Dietrich III of Limburg-Styrum* in the other, so the two are different entities in the graph and no direct link exists between them. The baselines therefore return other Limburg entities and the LLM abstains. Our traversal still reaches the father's passage through the coarser shared entity *Limburg*, which shows that proposition states can bridge a surface-form mismatch as long as such an entity exists.

Case 6 (2Wiki, bridge comparison): two director hops and a date comparison

Question

"Which film has the director died later, Lost in the Stratosphere or Blind Man's Eyes?"
Ground truth: Blind Man's Eyes
Gold chain: [Lost in the Stratosphere → director → Melville W. Brown]
[Melville W. Brown → date of death → January 31, 1938]
[Blind Man's Eyes → director → John Ince]
[John Ince → date of death → April 10, 1947]

Baselines

HippoRAG 2
Retrieved context:
1) ✓ **Blind Man's Eyes**
2) ✓ **Lost in the Stratosphere**
3) × William Cagney
4) × Blind Man (film)
5) × Dead Man's Eyes
Prediction:
× Lost in the Stratosphere

PropRAG
Retrieved context:
1) ✓ **Blind Man's Eyes**
2) × Blind Man (film)
3) ✓ **Lost in the Stratosphere**
4) ✓ **Melville W. Brown**
5) × William Cagney
Prediction:
× Lost in the Stratosphere

CatRAG
Retrieved context:
1) ✓ **Lost in the Stratosphere**
2) ✓ **Blind Man's Eyes**
3) × William Cagney
4) × Blind Man (film)
5) × Dead Man's Eyes
Prediction:
× Lost in the Stratosphere

ProST

Retrieved context:
1) ✓ **Blind Man's Eyes**
2) ✓ **Lost in the Stratosphere**
3) × Blind Man (film)
4) ✓ **Melville W. Brown**
5) ✓ **John Ince (actor)**
Prediction:
✓ **Blind Man's Eyes**

The baselines retrieve both films but at most one director, so the LLM compares an incomplete pair and picks the wrong film in all three cases. Our traversal runs the same step from each film: from "X directed the film Y" the score ranks the director's death-date proposition first among that director's propositions (-0.25 against -0.33, and -0.27 against -0.29), so both dates are retrieved rather than guessed.

G PROMPTS

Prompts 1–6 list the prompts used in this work. Indexing uses the proposition-extraction prompt of PropRAG and the two-step OpenIE prompts of HippoRAG 2, consisting of named-entity extraction followed by triple extraction, from which entity roles are derived. All systems use the one-shot QA prompt released with PropRAG. For ProST, evidence highlights are inserted as a "Key facts" block before the passages and are drawn from those same passages, so the LLM receives no additional evidence. LLM-Acc uses the judge prompt of LinearRAG, and the CRUX grader follows the rubric of Ju et al. (2025). Long few-shot demonstrations are abbreviated with "[. . .]".

Prompt 1: Proposition and entity extraction (PropRAG)

```
SYSTEM: Your task is to analyze text passages and break them down into precise, atomic
propositions using a specified list of named entities.  A proposition is a fully
contextualized statement that expresses a single unit of meaning with complete specificity
about the relationships described.

For each proposition:
1.  Extract a complete, standalone statement that preserves the full context
2.  Use ONLY the entities provided in the named_entities list - do not introduce new
entities
3.  Ensure each proposition contains only ONE claim or relationship
4.  Be extremely specific about which entities are involved in each relationship
5.  Maintain clear causal connections between related statements

Respond with a JSON object containing a list of propositions, where each proposition is an
object with:
- "text":  The full proposition text as a complete, contextualized statement
- "entities":  An array of entities from the named_entities list that appear in that
proposition

[Critical guidelines and two worked examples omitted; see the released PropRAG code.]

USER: Passage:  <passage>
Named entities:  <named_entities>
```

Prompt 2: Named-entity extraction (HippoRAG 2 OpenIE, step 1)

```
SYSTEM: Your task is to extract named entities from the given paragraph.
Respond with a JSON list of entities.

USER: Radio City
Radio City is India's first private FM radio station and was started on 3 July 2001.
[...]
ASSISTANT: {"named_entities":  ["Radio City", "India", "3 July 2001", "Hindi", "English",
"May 2008", "PlanetRadiocity.com"]}

USER: <passage>
```

Prompt 3: Triple extraction (HippoRAG 2 OpenIE, step 2; source of entity roles)

```
SYSTEM: Your task is to construct an RDF (Resource Description Framework) graph from the
given passages and named entity lists.
Respond with a JSON list of triples, with each triple representing a relationship in the
RDF graph.

Pay attention to the following requirements:
- Each triple should contain at least one, but preferably two, of the named entities in
the list for each passage.
- Clearly resolve pronouns to their specific names to maintain clarity.

USER: Convert the paragraph into a JSON dict, it has a named entity list and a triple
list.
Paragraph:
<passage>
<named_entity_json>
ASSISTANT (one-shot):  {"triples":  [["Radio City", "located in", "India"], ["Radio City",
"started on", "3 July 2001"], ...]}
```

Prompt 4: QA prompt (PropRAG / HippoRAG 2)

```
SYSTEM: As an advanced reading comprehension assistant, your task is to analyze text
passages and corresponding questions meticulously.  Your response start after "Thought:
", where you will methodically break down the reasoning process, illustrating how you
arrive at conclusions.  Conclude with "Answer:  " to present a concise, definitive
response, devoid of additional elaborations.

USER (one-shot demonstration):  Wikipedia Title:  The Last Horse
The Last Horse (Spanish:  El ultimo caballo) is a 1950 Spanish comedy film directed by
Edgar Neville [...]
[four more demonstration passages]
Question:  When was Neville A. Stanton's employer founded?
Thought:
ASSISTANT: The employer of Neville A. Stanton is University of Southampton.  The
University of Southampton was founded in 1862.  So the answer is:  1862.
Answer:  1862

USER: Key facts:
- <proposition 1>
- <proposition 2>
- <proposition 3>

Wikipedia Title:  <passage 1>
[...]
Wikipedia Title:  <passage 5>

Question:  <question>
Thought:
```

Prompt 5: LLM-Acc judge (LinearRAG)

```
SYSTEM: You are an expert evaluator.

USER: Please evaluate if the generated answer is correct by comparing it with the gold
answer.
Generated answer:  <prediction>
Gold answer:  <gold>

The generated answer should be considered correct if it:
1.  Contains the key information from the gold answer
2.  Is factually accurate and consistent with the gold answer
3.  Does not contain any contradicting information

Respond with ONLY 'correct' or 'incorrect'.
Response:
```

Prompt 6: CRUX grader

```
USER: Instruction:  Determine whether the question can be answered based on the provided
context?  Rate the context with on a scale from 0 to 5 according to the guideline below.
Do not write anything except the rating.
Guideline:
5:  The context is highly relevant, complete, and accurate.
4:  The context is mostly relevant and complete but may have minor gaps or inaccuracies.
3:  The context is partially relevant and complete, with noticeable gaps or inaccuracies.
2:  The context has limited relevance and completeness, with significant gaps or
inaccuracies.
1:  The context is minimally relevant or complete, with substantial shortcomings.
0:  The context is not relevant or complete at all.
Question:  <sub-question>
Context:  <passage>
Rating:
```