

Knowledge Graph-Enhanced LLM Reranking for Complex Information Needs

Cassidy M. Huerta

Department of Computer Science, George Mason University, Fairfax, VA, USA.
cassidywork@gmu.edu

Shona L. Dick

Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
shona.dick69@unr.edu

Vinzenz Klotz

Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
vinzenzklotz77@colostate.edu

Angelique Ditschlerin

Department of Computer Science, Binghamton University, Binghamton, NY, USA.
angelique115@binghamton.edu

Abstract

Complex information needs rarely reduce to isolated keyword matching because they involve temporal constraints, multi-hop relationships, tacit domain assumptions, and conflicting evidentiary standards. Large language models have improved reranking by capturing nuanced query intent and generating relevance judgments beyond lexical overlap, yet they remain vulnerable to hallucination, shallow provenance, and brittle generalization when the underlying evidence graph is absent from the ranking process. This paper examines knowledge graph-enhanced LLM reranking as a socio-technical architecture for complex information needs. It argues that knowledge graphs can supply entity disambiguation, relational context, provenance signals, and constraint validation that improve LLM reranking, but only when graph construction, retrieval orchestration, inference serving, governance, and evaluation are treated as coupled infrastructural concerns. The analysis distinguishes several architectural patterns, including graph-expanded candidate generation, multi-role collaborative reranking, and hybrid symbolic-neural scoring. It then evaluates structural trade-offs involving latency, recall, explainability, update velocity, schema drift, and cost. The paper also considers fairness, robustness, privacy, sustainability, and policy implications, noting that graph-enhanced reranking may amplify representational imbalances if knowledge graphs encode historical inequities or if LLM adjudication obscures contestable judgments. Deployment concerns include graph stores, vector indexes, caching, observability, human oversight, and energy accounting. The conclusion proposes that future systems should be evaluated not only by ranking accuracy but also by contestability, provenance, maintainability, and equitable access across domains.

Keywords

knowledge graphs, large language models, reranking, information retrieval, socio-technical systems, governance, provenance, fairness, sustainability.

1. Introduction

Complex information needs are increasingly common in professional, scientific, legal, medical, and civic settings where users seek not merely documents but defensible evidence assembled from multiple sources. Traditional retrieval pipelines often begin with lexical scoring, such as the probabilistic relevance framework underlying BM25, and then apply neural reranking to improve ordering [4]. Dense retrieval methods have further shifted matching toward learned embeddings that capture semantic similarity between queries and passages [2]. Retrieval-augmented generation has extended this trajectory by coupling language models with external evidence so that outputs can be grounded in retrieved content [1]. Nevertheless, a persistent gap remains between retrieving semantically similar text and satisfying structurally complex information needs that depend on entity relationships, temporal validity, source authority, and contradictory constraints.

Large language model rerankers offer a promising response because they can interpret nuanced instructions, compare passages holistically, and generate rationales for relevance. Early neural reranking work demonstrated that transformer models can substantially improve passage ordering by modeling query-document interactions more richly than bag-of-words or simple embedding similarity [3]. However, LLM reranking introduces its own difficulties. Models may favor fluent passages over authoritative ones, overlook missing links in multi-hop evidence chains, or fabricate justifications that are not grounded in the retrieved corpus. These weaknesses are especially consequential when the information need is not a simple factoid question but a policy analysis, a systematic review, a due diligence investigation, or a clinical decision support request. In such cases, the ranking system must respect constraints that are relational, evidentiary, and often institutional.

Knowledge graphs provide a complementary representational layer. They encode entities, types, attributes, and relationships, thereby making explicit some of the structure that flat text retrieval leaves implicit [7]. Surveys of knowledge graph research show that graphs support representation learning, link prediction, question answering, recommendation, and knowledge-intensive inference across heterogeneous domains [6]. Knowledge graph embedding methods such as translational models illustrate how relational structure can be projected into vector spaces for downstream computation [5]. When combined with LLM reranking, knowledge graphs can help disambiguate entities, expand queries along typed relations, validate constraints, and attach provenance to candidate evidence. The central question is not whether graphs can improve ranking in principle, but how they should be integrated into production-grade architectures without sacrificing robustness, fairness, maintainability, or sustainability.

This paper develops a system-level account of knowledge graph-enhanced LLM reranking for complex information needs. It treats reranking as an infrastructural and governance problem rather than a narrow model optimization task. The discussion proceeds through architectural patterns, structural trade-offs, deployment realities, evaluation challenges, cross-domain comparisons, and policy implications. The argument is that graph-enhanced LLM reranking can be powerful, but only if it is designed as a contestable, observable, and accountable pipeline that connects retrieval, graph curation, model inference, human review, and long-term maintenance.

2. Background and Related Work

Information retrieval has evolved from sparse lexical matching to dense neural representation and then to hybrid systems that combine the strengths of both. BM25 remains influential because it offers strong baselines, interpretability, and computational efficiency [4]. Dense passage retrieval introduced effective dual-encoder architectures that learn a shared embedding space for questions and passages, improving semantic recall in open-domain question answering [2]. Subsequent work on passage reranking with BERT showed that cross-encoder architectures can refine candidate sets by jointly encoding query and passage, although at higher computational cost [3]. Retrieval-augmented generation then connected retrieval with generation, allowing language models to condition on external documents rather than relying solely on parametric memory [1]. These developments provide the foundation for the present focus on reranking, because reranking is where retrieval breadth meets precision, evidentiary filtering, and user-facing ordering.

Knowledge graphs have developed along a parallel track. They provide schema-based representations of entities and relations, often drawing on semantic web standards, ontologies, and graph databases [7]. Research on knowledge graph representation, acquisition, and applications has produced methods for embedding, completion, alignment, and question answering [6]. Translational embedding models demonstrated that relations can be interpreted as transformations in a latent space, enabling link prediction and structured inference [5]. In practice, knowledge graphs are used in search engines, biomedical knowledge bases, financial compliance systems, and enterprise data catalogs. Their value for complex information needs lies in making relational constraints explicit: a user may need evidence about a drug interaction, a corporate ownership chain, a legal precedent, or a scientific claim that depends on multiple experimental findings.

Large language models have changed the landscape of both retrieval and ranking. Few-shot learning showed that sufficiently large models can perform new tasks from prompts without gradient updates [9]. Instruction tuning and reinforcement learning from human feedback further improved the ability of models to follow user intent and align with preference signals [10]. Chain-of-thought prompting demonstrated that models can produce intermediate reasoning steps that improve performance on tasks requiring multi-step inference [8]. These capabilities are relevant to reranking because complex information needs often require comparative judgment, constraint checking, and synthesis. Yet LLMs are not self-sufficient knowledge authorities. They may confabulate details, over-weight stylistic cues, and lack reliable provenance unless the architecture constrains them with retrieved evidence and structured validation.

Existing benchmarks and systems have addressed parts of this problem. FEVER established a large-scale benchmark for fact extraction and verification, emphasizing evidence-based claim checking [11]. KILT provided a unified benchmark for knowledge-intensive language tasks that require access to a common evidence source [12]. Fusion-in-Decoder and related generative retrieval approaches showed how retrieval can be integrated with generation for open-domain question answering [13]. REALM and RETRO explored retrieval-augmented pretraining, demonstrating that models can learn to condition on external corpora during pretraining or inference [14][15]. Self-RAG introduced self-reflection mechanisms so that models can retrieve, generate, and critique their own outputs [16]. A recent multi-role collaborative reranking workflow further illustrates how multiple LLM roles can be coordinated to improve ranking decisions [17]. These contributions collectively motivate a

graph-enhanced reranking architecture that treats structured knowledge as an active participant in candidate expansion, evidence validation, and final ordering.

3. Architecture of Knowledge Graph-Enhanced LLM Reranking

A knowledge graph-enhanced LLM reranking system is best understood as a pipeline of interdependent services rather than a single model. The first stage involves query understanding, where the system identifies entities, intents, temporal constraints, source preferences, and relational requirements. This stage may use named entity recognition, entity linking, intent classification, and decomposition of multi-part questions. The second stage performs candidate retrieval using hybrid methods. Lexical retrieval remains valuable for exact matches, identifiers, and rare terms, while dense retrieval captures semantic similarity and paraphrase. The third stage uses the knowledge graph to expand or constrain the candidate set. For example, a query about a corporate merger may require traversal from an acquired company to subsidiaries, executives, regulatory filings, and jurisdiction-specific legal entities. A biomedical query may require linking a drug to targets, pathways, adverse events, and clinical trials. Graph expansion can improve recall when relevant evidence is not lexically similar to the query, but it can also introduce topical drift if traversal is not constrained by type, relation, time, or confidence.

The fourth stage is evidence validation and graph-based scoring. Here, the reranker does not merely compare query-passage pairs; it examines whether a candidate passage satisfies relational constraints encoded in the graph. A passage may mention the right entity but refer to the wrong time period, jurisdiction, or organizational role. A graph can surface such mismatches by checking typed edges, temporal validity intervals, and provenance metadata. The fifth stage is LLM reranking, where a language model receives the query, candidate passages, graph-derived features, and possibly a structured evidence summary. The model produces a relevance score, a ranking, or a comparative judgment across passages. In multi-role designs, one role may decompose the query, another may retrieve graph facts, another may critique evidence, and another may synthesize the final ranking. The recent collaborative reranking workflow exemplifies this division of labor, using multiple roles to improve robustness and interpretability [17]. The sixth stage is calibration and presentation, where scores are mapped to user-facing confidence signals, citations, and explanations.

This architecture creates several design choices. One choice concerns graph centrality. The graph may be a primary retrieval index, a secondary validation layer, or a feature source for the LLM. Another choice concerns the timing of graph integration. Early fusion uses graph signals during candidate generation, while late fusion uses them only in final reranking. Early fusion can improve recall but may increase latency and bias the candidate pool toward graph-covered entities. Late fusion preserves the diversity of initial retrieval but may miss relevant evidence that lacks graph linkage. A third choice concerns the granularity of graph representation. Entity-level graphs, document-level graphs, and claim-level graphs each support different forms of reasoning. Claim-level graphs are particularly relevant for complex information needs because they can represent propositions, evidence relations, and contradictions. However, they are more expensive to construct and maintain.

The architecture must also address uncertainty. Knowledge graphs are incomplete, noisy, and historically situated. Edges may be missing, outdated, or derived from biased sources. LLM rerankers may be overconfident when graph features are sparse. A robust system therefore needs uncertainty propagation across retrieval, graph lookup, and LLM scoring. It should distinguish between absence of evidence and evidence of absence, and it should expose when

a ranking decision depends on weak graph links. In domains such as medicine, law, and finance, this distinction is not merely technical; it affects liability, safety, and user trust. The architecture should also support counterfactual inspection so that users or auditors can ask which graph facts, passages, or prompt instructions most influenced the ordering.

4. Structural Trade-offs in Graph Integration, Retrieval, and Ranking

The integration of knowledge graphs into LLM reranking introduces structural trade-offs that cannot be resolved by model scaling alone. One trade-off is between coverage and precision. A large, open-domain graph may cover many entities but suffer from schema heterogeneity, stale facts, and noisy relations. A narrow, curated graph may be accurate but fail on emerging entities, niche domains, or cross-domain queries. Complex information needs often cross boundaries, such as a query that links scientific evidence, regulatory policy, and commercial activity. The graph must therefore support federation across multiple knowledge sources, but federation increases entity resolution challenges, latency, and governance complexity. The system must decide whether to resolve entities during ingestion, during query processing, or through human review.

A second trade-off is between transparency and performance. Graph-based features can make reranking more explainable by showing relational paths and evidence chains. However, the most accurate LLM reranking may rely on opaque attention patterns and latent representations. If the system prioritizes benchmark performance, it may down-weight symbolic constraints that are difficult to optimize but important for accountability. Conversely, if it prioritizes transparency, it may sacrifice subtle semantic judgments. A practical compromise is layered transparency, where the final ranking is accompanied by graph-derived rationales, passage citations, and confidence indicators, even if the underlying LLM scoring remains partially opaque. This layered approach aligns with governance requirements for contestability and auditability.

A third trade-off is between update velocity and stability. Knowledge graphs change as new entities, relations, and evidence appear. LLM rerankers may be retrained or prompted with updated graph summaries. Frequent updates improve freshness but can destabilize behavior, introduce regressions, and complicate reproducibility. Slow updates preserve consistency but may produce outdated rankings. Complex information needs often have temporal dimensions, such as legal statutes that change, clinical guidelines that are revised, or financial relationships that shift. The system therefore needs versioned graphs, time-aware edges, and evaluation protocols that test temporal generalization. It also needs rollback mechanisms so that harmful updates can be reversed.

A fourth trade-off is between centralization and federation. A centralized knowledge graph and reranking service can simplify governance, caching, and observability. However, it may create single points of failure, concentrate control, and raise privacy concerns when data crosses organizational boundaries. Federated architectures distribute graph storage and inference, allowing local control and regulatory compliance, but they complicate ranking because the LLM may see only partial graph views. Cross-domain comparisons are instructive. In healthcare, federated graph-enhanced reranking may be necessary to respect patient privacy and institutional data governance. In e-commerce, centralized graphs may be more practical because product catalogs and user interactions are already aggregated. In public-sector applications, hybrid models may be required to balance transparency with jurisdictional autonomy.

A fifth trade-off concerns evaluation. Standard ranking metrics such as precision at k, mean reciprocal rank, and normalized discounted cumulative gain capture only part of system performance. For complex information needs, the system must also be evaluated on evidence coverage, provenance correctness, temporal validity, fairness across user groups, and resistance to manipulation. If the knowledge graph privileges well-represented entities and languages, reranking may systematically disadvantage under-resourced communities or emerging domains. If the LLM is optimized on historical click data, it may reproduce popularity bias and suppress novel but relevant evidence. The architecture should therefore include counterfactual evaluation, subgroup analysis, and adversarial testing as first-class infrastructural functions.

5. Governance, Fairness, Robustness, and Policy Implications

Knowledge graph-enhanced LLM reranking sits at the intersection of data governance, model governance, and information policy. The knowledge graph is not a neutral repository. It reflects decisions about which entities to include, which relations to formalize, which sources to trust, and which languages and cultures to represent. These decisions shape the reranking outcomes. If a graph underrepresents women, minority-owned businesses, indigenous knowledge, or non-English sources, then graph-expanded retrieval may amplify those omissions. Fairness therefore requires attention to graph curation, schema design, and source selection, not only to the LLM prompt or ranking objective. Fairness audits should examine representation across entity types, relation coverage, and evidence quality. They should also consider intersectional harms, because a user may be disadvantaged by multiple overlapping gaps.

Robustness is equally important. Adversaries may attempt to manipulate rankings by injecting false graph facts, creating misleading entities, or poisoning passages that the LLM reranker treats as authoritative. Because graph-enhanced systems often rely on structured relations, an attack on a single high-centrality edge can propagate through many queries. Defenses include provenance checks, edge confidence scores, anomaly detection, source diversity requirements, and human review for high-impact domains. Robustness also involves handling distribution shift. A graph built for one domain may perform poorly when transferred to another. An LLM reranker tuned on one language or culture may misinterpret relational cues in another. The system should therefore expose domain assumptions and support localized adaptation.

Privacy and security present additional challenges. Knowledge graphs can link entities in ways that reveal sensitive relationships, even when individual data points are not obviously personal. LLM reranking may process confidential queries, proprietary documents, or regulated data. The architecture must support access control, encryption, data minimization, and audit logging. Federated learning and confidential computing may help, but they introduce performance overhead and complexity. Policy implications include compliance with data protection regimes, sector-specific regulations, and emerging AI governance frameworks. In healthcare, reranking systems may be classified as clinical decision support and subject to safety validation. In finance, they may fall under fair lending, market conduct, or anti-money laundering rules. In public administration, they may be subject to freedom of information, administrative fairness, and non-discrimination requirements.

Explainability and contestability are central governance values. Users affected by ranking decisions should be able to understand why certain evidence was prioritized and how to challenge errors. Graph-enhanced systems can support this by exposing relational paths, source citations, and confidence estimates. However, explanations must be meaningful rather

than decorative. A list of graph edges is not sufficient if users cannot interpret the schema or assess source reliability. The system should provide layered explanations tailored to different stakeholders, including end users, domain experts, auditors, and regulators. It should also maintain records of model versions, graph versions, prompts, and retrieval configurations so that decisions can be reproduced and reviewed.

6. Deployment, Infrastructure, and Sustainability

Deploying knowledge graph-enhanced LLM reranking requires significant infrastructure. A production system may include a graph database or triple store, a vector index, a document store, a retrieval orchestrator, one or more LLM inference services, a feature store, a cache layer, and an observability stack. Each component has failure modes. The graph store may suffer from slow traversals, inconsistent replicas, or schema drift. The vector index may return stale or biased neighbors. The LLM service may experience latency spikes, rate limits, or version mismatches. The orchestrator may create cascading failures if retries and timeouts are poorly configured. Observability must therefore span retrieval quality, graph query performance, model confidence, user feedback, and end-to-end latency. Without such observability, operators cannot diagnose whether a ranking problem originates in the graph, the retriever, the LLM, or the user interface.

Sustainability is an often-overlooked dimension. LLM inference is computationally expensive, and graph traversal can add further cost. If every query triggers multi-hop graph expansion, multiple LLM calls, and extensive reranking, energy consumption and operational expenses may become prohibitive. The system should use adaptive computation: lightweight retrieval and graph checks for simple queries, and deeper multi-role reranking only when the information need is complex or high stakes. Caching can reduce repeated work, but caches must respect access control and freshness. Model distillation, quantization, and smaller specialized rerankers can reduce cost, though they may reduce nuanced judgment. Sustainability also involves data lifecycle management. Knowledge graphs and document collections grow over time, and retaining everything indefinitely increases storage, energy, and privacy risk. Retention policies should balance evidentiary value against environmental and legal burdens.

From an infrastructure perspective, deployment choices shape equity. Cloud-based services may offer scalability and advanced models, but they can be inaccessible to smaller organizations, public institutions, or regions with limited connectivity. On-premises or edge deployments may improve sovereignty and latency, but they require expertise and capital. A multi-tier architecture may be appropriate, with local graph caches and lightweight rerankers at the edge, and larger LLMs in regional data centers. Interoperability standards for graphs, provenance, and model interfaces can reduce lock-in and support competition. However, standards must be accompanied by governance mechanisms that prevent dominant platforms from defining schemas in ways that disadvantage others.

7. Evaluation, Case Illustrations, and Cross-Domain Comparisons

Evaluation of knowledge graph-enhanced LLM reranking must move beyond static benchmarks. Complex information needs are often dynamic, contested, and context-dependent. A ranking that is correct for one user may be inappropriate for another because of jurisdiction, role, risk tolerance, or evidentiary standard. Evaluation should therefore include scenario-based assessments, expert review, user studies, and longitudinal monitoring. It should test not only whether relevant documents are ranked highly but also whether the

evidence chain is complete, whether sources are authoritative, whether contradictory evidence is surfaced, and whether the system communicates uncertainty. Adversarial evaluation should probe for graph poisoning, prompt injection, entity ambiguity, and temporal confusion. Fairness evaluation should compare outcomes across languages, regions, identities, and domains.

Case illustrations clarify the stakes. In healthcare, a clinician may ask for evidence about a rare adverse event associated with a drug in patients with a specific comorbidity. A graph-enhanced reranker could link the drug to targets, adverse event reports, clinical trials, and contraindications, while the LLM evaluates passages for clinical relevance. The system must be robust to incomplete reporting, distinguish correlation from causation, and comply with privacy rules. In law, a query may concern the precedential status of a case across jurisdictions. A knowledge graph can represent citations, overruling relationships, and jurisdictional hierarchies, while the LLM reranks passages according to relevance and authority. Errors can affect legal advice, so provenance and contestability are essential. In scientific discovery, a researcher may need evidence connecting a gene, pathway, disease, and intervention across publications. Graph expansion can reveal non-obvious links, but the LLM must avoid overstating weak associations. In e-commerce, graph-enhanced reranking may improve product discovery by linking attributes, compatibility, and user intent, but it must avoid manipulative ranking and ensure fair competition among sellers. In public administration, reranking may support benefits eligibility, procurement, or policy analysis, where transparency and non-discrimination are legal requirements.

Cross-domain comparisons reveal recurring architectural tensions. Domains with high safety stakes require stronger provenance, human oversight, and auditability, even at the cost of latency and recall. Domains with rapid change require time-aware graphs and continuous evaluation. Domains with heterogeneous stakeholders require federated governance and contestable explanations. Domains with limited resources require lightweight models, caching, and shared infrastructure. A universal architecture is unlikely to be optimal. Instead, systems should be modular so that graph schemas, retrieval policies, reranking prompts, and governance controls can be adapted to domain-specific needs. The recent graph-based query-focused summarization approach illustrates how graph structure can support global sensemaking, but it also shows that graph construction and community detection choices strongly shape outputs [20]. Similar caution applies to reranking, where graph topology can determine which evidence appears salient.

8. Future Directions

Future research should advance several interconnected areas. First, neurosymbolic architectures could combine differentiable LLM scoring with symbolic graph constraints in ways that are both performant and inspectable. Second, continual learning and graph updating should be studied together, so that rerankers adapt to new evidence without catastrophic forgetting or hidden drift. Third, evaluation should incorporate contestability metrics, provenance quality, and fairness across underrepresented groups. Fourth, infrastructure research should develop energy-aware orchestration, adaptive computation, and privacy-preserving federated graph retrieval. Fifth, governance research should examine how auditors, regulators, and affected communities can participate in the design and oversight of graph-enhanced ranking systems. Surveys of augmented language models and LLM-based information retrieval provide useful foundations, but they also highlight the need for integrated system-level perspectives [18][19].

Another direction concerns multilingual and multicultural knowledge graphs. Complex information needs increasingly cross linguistic and jurisdictional boundaries. A graph that privileges English-language sources may systematically marginalize other knowledge traditions. Multilingual entity linking, cross-lingual relation extraction, and culturally aware schema design are therefore not merely technical enhancements but matters of epistemic justice. Similarly, temporal reasoning deserves deeper attention. Many complex needs involve change over time, and a static graph cannot represent evolving relationships adequately. Time-aware graphs, versioned evidence, and temporal reranking objectives should become standard components. Finally, human-computer interaction research should explore how users interpret graph-enhanced explanations, how they calibrate trust, and how they contest erroneous rankings in practice.

9. Conclusion

Knowledge graph-enhanced LLM reranking offers a promising path for addressing complex information needs that exceed the capacity of lexical matching, dense retrieval, or standalone language model judgment. Knowledge graphs can supply entity structure, relational context, provenance, and constraint validation, while LLMs can provide nuanced semantic comparison and flexible instruction following. The combination, however, is not a simple additive improvement. It creates architectural complexity, governance obligations, and sustainability challenges. This paper has argued that graph-enhanced reranking should be treated as a socio-technical infrastructure that couples retrieval, graph curation, model inference, human oversight, and policy compliance. The most important design choices concern how graph signals are integrated, how uncertainty and provenance are represented, how updates are governed, how fairness is audited, and how costs are managed. Systems that address these concerns are more likely to be robust, contestable, and equitable. Systems that ignore them may achieve impressive benchmark scores while remaining brittle, opaque, and unsustainable. The next generation of information retrieval systems should therefore be evaluated not only by ranking accuracy but also by their ability to support defensible, transparent, and socially accountable access to knowledge.

References

1. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
2. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 6769–6781). Association for Computational Linguistics.
3. Nogueira, R., & Cho, K. (2019). Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.
4. Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333–389.
5. Bordes, A., Usunier, N., Garcia-Durán, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. *Advances in Neural Information Processing Systems*, 26, 2787–2795.

6. Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2022). A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2), 494–514.
7. Hogan, A., Blomqvist, E., Cochez, M., d’Amato, C., de Melo, G., Gutiérrez, C., Kirrane, S., Labra Gayo, J. E., Navigli, R., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge graphs. *ACM Computing Surveys*, 54(4), 1–37.
8. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.
9. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
10. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
11. Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 809–819). Association for Computational Linguistics.
12. Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., De Cao, N., Thorne, J., Jernite, Y., Karpukhin, V., Maillard, J., Plachouras, V., Rocktäschel, T., & Riedel, S. (2021). KILT: A benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2523–2544). Association for Computational Linguistics.
13. Izacard, G., & Grave, E. (2021). 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* (pp. 874–880). Association for Computational Linguistics.
14. Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020). REALM: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning* (pp. 3929–3938). PMLR.
15. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J.-B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Driessche, G., Hennigan, T., Huang, S., Maggiore, L., Jones, C., ... Sifre, L. (2022). Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning* (pp. 2206–2240). PMLR.

16. Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In The Twelfth International Conference on Learning Representations. OpenReview.net.
17. Jin, C., Peng, H., Zhang, A., Chen, N., Zhao, J., Xie, X., ... & Metaxas, D. N. (2025, May). Rankflow: A multi-role collaborative reranking workflow utilizing large language models. In Companion Proceedings of the ACM on Web Conference 2025 (pp. 2484-2493).
18. Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., Grave, E., LeCun, Y., & Scialom, T. (2023). Augmented language models: A survey. Transactions on Machine Learning Research.
19. Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Chen, H., Dou, Z., & Wen, J.-R. (2024). Large language models for information retrieval: A survey. arXiv preprint arXiv:2308.07107.
20. Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., & Larson, J. (2024). From local to global: A graph RAG approach to query-focused summarization. arXiv preprint arXiv:2404.16130.