

Comparative Study of Knowledge Graph-Augmented and Retrieval-Augmented Large Language Models for Educational Chatbots

Archna Pandey¹, Mrs. Arifa Khan²

¹Master of Technology, Computer Science and Engineering, Lucknow Institute of Technology, Lucknow, India

²Assistant Professor, Department of Computer Science and Engineering, Lucknow Institute of Technology, Lucknow, India

Abstract - Educational conversational systems require more than fluent language generation: their explanations must remain consistent with approved instructional material and correctly connect related concepts. This paper examines two strategies for supplying external knowledge to large language models (LLMs): passage-based retrieval-augmented generation (RAG) and knowledge-graph (KG) augmentation. Using undergraduate data structures and algorithms as a representative domain, we develop a controlled comparison framework encompassing knowledge-source preparation, query stratification, retrieval design, answer assessment, and maintenance costs. Text retrieval preserves explanatory context and can be updated through document ingestion, whereas graph retrieval explicitly represents prerequisites, dependencies, and conceptual links. A three-condition protocol unaugmented LLM, RAG, and KG augmentation is specified for descriptive, comparative, and multi-step questions. We identify potential sources of error in each architecture and propose a combination of automated grounding measures and expert pedagogical judgments. Importantly, the numerical examples available in the underlying research materials were designated illustrative, not validated observations; consequently, this article reports a methodological and analytical contribution rather than claiming measured superiority. The resulting framework supports reproducible evaluation and motivates adaptive hybrid retrieval for educational assistants.

Keywords: Educational chatbots; large language models; retrieval-augmented generation; knowledge graphs; data structures and algorithms; grounded question answering

1. INTRODUCTION

Conversational artificial intelligence has become an attractive interface for academic support because learners can formulate questions in ordinary language and receive immediate explanations. Yet the apparent fluency of an answer is not evidence that its content is correct. A student learning algorithmic complexity, for example, may accept an inaccurate claim about the worst-case performance of a search procedure if it is presented confidently. The educational requirement is therefore stricter than producing text that resembles a plausible answer: responses must be verifiable, consistent with a course, and useful for building transferable understanding.

Large language models derive broad capabilities from pretraining and instruction tuning, but their parameters do not constitute a dependable repository of institution-specific facts. Their knowledge may be outdated, and their generated citations or explanations can be unsupported. Grounding addresses this weakness by providing selected external information at inference time. The central design question is not simply whether to retrieve knowledge, but how the representation of that knowledge shapes the answers learners receive.

Two influential solutions use different representations. Retrieval-augmented generation supplies excerpts from textbooks, lecture materials, or curated documentation. Knowledge-graph augmentation instead makes domain entities and typed relations directly available to the generator. Their respective merits depend on the form of the question. A definition may be answered from one passage, while explaining why dynamic programming depends on recursion, subproblems, and optimal substructure requires connecting multiple concepts.

This paper contributes a coherent evaluation design for comparing these approaches under shared model conditions. It separates theoretical expectations from verified empirical findings, defines suitable question categories and assessment dimensions, and examines practical constraints that matter in institutional deployment. The emphasis is on undergraduate computer science, where prerequisite chains offer a useful test of relational reasoning. The work does not equate chatbot answer quality with demonstrated student learning gains; those outcomes require a different study design.

2. RESEARCH CONTEXT AND RELATED WORK

Work on educational chatbots predates generative models. Earlier systems relied on scripted interactions, decision trees, or limited intent classification and were useful for routine support but fragile when students departed from expected wording. Wollny et al. (2021) describe the diversity of chatbot roles in education while also documenting limitations in pedagogical design and evaluation. The emergence of general-purpose LLMs greatly widened the range of questions that conversational agents could address, but did not remove the need for educational safeguards.

Kasneji et al. (2023) discuss opportunities for feedback, individualized explanation, and instructional content generation alongside risks of inaccurate outputs, bias, and overreliance. Ji et al. (2023) provide a broader account of hallucination in natural language generation. In an educational application, such errors can be especially consequential because the intended reader may not yet possess the background needed to identify them. This motivates systems that expose evidence and support external verification.

Lewis et al. (2020) introduced a retrieval-augmented formulation combining a generative model with nonparametric knowledge access. Subsequent RAG work has expanded the basic retrieve-then-generate pattern through reranking, iterative search, and modular components (Gao et al., 2023). For course-based assistants, the attraction is straightforward: an instructor can revise a document collection without modifying the language model itself. However, a passage retrieved because it shares vocabulary with a question may omit the critical relationship needed to answer it.

Knowledge graphs address a different aspect of grounding. Their nodes represent concepts and their edges encode relations such as prerequisite-of, instance-of, or depends-on. Explicit graph traversal can expose a chain of relevant facts that would otherwise be dispersed across multiple documents. Pan et al. (2024) situate graph-LLM integration within a broader research agenda. Graph-oriented retrieval has also influenced hybrid systems: HippoRAG (Gutiérrez et al., 2024) uses graph structure to support associative retrieval, while Edge et al. (2024) investigate graph-based organization for broader synthesis tasks.

These approaches should not be treated as intrinsically exclusive. Documents provide nuanced explanation and examples; graphs offer controlled relationships and inspectable paths. The appropriate comparison must ask whether the same learner question, source coverage, base generator, and evaluation criteria are used. Otherwise differences may reflect unequal knowledge access, model capacity, or prompting rather than the augmentation strategy itself.

Text-retrieval grounding workflow



Figure 1. Conceptual workflow for passage-based grounding; no empirical measurements are represented.

3. COMPARATIVE ARCHITECTURE

A controlled design comprises three conditions: a language model answering without an external course-specific source, a passage-grounded version of the same model, and a graph-grounded version. The generator and decoding settings should remain constant across conditions. Differences in answer quality can then be interpreted more plausibly as consequences of knowledge representation and retrieval, although retrieval budget and context length remain important potential confounders.

In the passage-based pipeline, course resources are cleaned, divided into coherent segments, embedded, and indexed for similarity search. A student query is encoded using the same embedding space; top-ranked segments are assembled into a prompt with instructions to rely on supplied evidence and acknowledge missing information. Heading-aware segmentation is preferable to arbitrary cuts because definitions, examples, and exceptions frequently belong together. Nevertheless, even a well-designed retriever can return superficially relevant passages that do not contain the answer.

In the graph-based pipeline, domain concepts are normalized into nodes and linked through explicitly typed edges. A question is mapped to one or more graph entities, after which neighborhood search or bounded multi-hop traversal returns facts and paths. The retrieved subgraph is serialized into an interpretable context for the generator. This structure is particularly useful when a question asks for prerequisites or contrasts related topics, but entity-linking mistakes can redirect traversal to the wrong part of the curriculum.

The source materials describe a proposed implementation using Python, LangChain, sentence embeddings, a FAISS vector index, and Neo4j for graph queries. They also specify a shared instruction-tuned generator, low-temperature decoding, and a fixed retrieval depth. Such choices are reasonable as a reproducibility template; however, naming a software configuration does not independently establish that every reported evaluation was completed. A publishable empirical extension should release versioned source snapshots, prompts, model identifiers, evaluation scripts, and de-identified question sets where permissions allow.

Knowledge equivalence is essential. If the graph contains manually curated prerequisite relations absent from the text corpus, any observed advantage may result from additional information rather than structural representation. Conversely, a graph containing only sparse triples may be disadvantaged against rich explanatory passages. Both resources should therefore derive from the same approved instructional materials, with a documented mapping between document evidence and graph assertions.

Knowledge-graph grounding workflow



Figure 2. Conceptual workflow for relationship-based grounding; no empirical measurements are represented.

4. EVALUATION DESIGN

The evaluation unit is a learner-style question paired with the answer generated under each architecture. A balanced question bank should distinguish descriptive questions, comparative questions, and relational or multi-hop questions. Descriptive items request definitions, examples, or complexity facts; comparative items require contrasting alternatives; multi-hop items require connecting prerequisite concepts or explaining dependency chains. Stratification prevents an overall average from hiding a method that works well for one question type but poorly for another.

The proposed research design contains 90 questions, divided equally across the three categories, producing 270 answers for three conditions. These counts describe the planned comparison, not evidence that a fully audited experimental dataset has been deposited. Each question needs a reference answer or evidence set reviewed for ambiguity and scope. The same wording should be used for all systems, and answer order should be randomized before human assessment.

Automated evaluation can include answer relevance, evidence faithfulness, and retrieval quality. Faithfulness asks whether claims in the response are supported by supplied evidence; relevance asks whether the response addresses the question; context precision and recall examine the retrieved material. Such measures require careful interpretation. A graph condition may supply serialized triples rather than natural-language passages, and the baseline has no retrieved context. Consequently, context-dependent metrics are not automatically comparable across all three conditions. A common reference-evidence protocol is needed, and automated judgments should be supplemented by human review.

Expert assessment should use anchored criteria covering factual correctness, pedagogical clarity, curriculum alignment, conceptual connectivity, and practical usefulness. Reviewers should be blind to system identity and should receive examples of what constitutes each score. For conceptual connectivity, a high score requires a correct and

meaningful relationship between ideas, not merely mentioning additional terminology. Inter-rater agreement must be calculated from actual independent ratings; it cannot be assumed from a target or example coefficient.

Because each question is answered by every system, observations are paired. A Friedman test may assess overall differences across three conditions, with appropriately adjusted paired comparisons following a significant omnibus result. Effect sizes and confidence intervals should accompany significance tests. The question-level data are required for valid inferential calculations: mean scores alone do not permit reconstruction of paired ranks or the distribution of differences.

Operational evaluation also matters. End-to-end latency, storage requirements, retrieval failures, human authoring effort, and the time required to add or revise a topic affect feasibility. The comparison should separate initial graph construction from routine maintenance and document the human verification needed for new edges. Reporting both response quality and engineering effort is more informative than ranking systems by a single automated score.

Hybrid evidence routing (proposed)

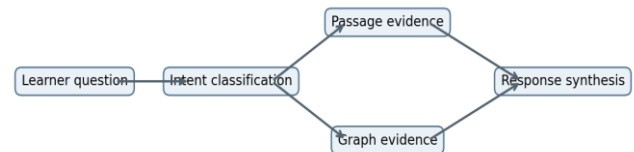


Figure 3. Proposed hybrid routing between textual and relational evidence (design schematic).

5. ANALYTICAL COMPARISON AND DESIGN IMPLICATIONS

The clearest architectural distinction concerns the unit of retrieved evidence. Passage retrieval returns contextual prose that can support definitions, worked examples, and careful explanation of edge cases. It can preserve the narrative structure of approved instructional content, making it well suited to questions such as the purpose of a base case in recursion or the average-case behavior of a familiar algorithm. Its weakness appears when the relevant facts lie in different segments and the retrieval step fails to assemble them.

Graph retrieval makes relations first-class objects. For a question about learning dynamic programming, a path connecting recursion, overlapping subproblems, memoization, and optimal substructure can provide an explicit reasoning scaffold. A graph can also distinguish relations that are easily conflated in ordinary prose, such as “is a prerequisite for” and “is an example of.” Yet graph

structure does not guarantee a good explanation. The generator must translate terse relations into coherent educational language without inventing unsupported intermediate steps.

The graph below summarizes anticipated architectural fit by question type. It is a conceptual comparison, not an empirical score chart. The categories indicate where each representation supplies a natural kind of evidence, not which system has been proven to achieve higher accuracy. Descriptive answers favor intact prose; relational questions favor explicit paths; comparative questions often need both source explanation and controlled concept alignment.

Error analysis should therefore distinguish retrieval failure, representation failure, and generation failure. In RAG, an incorrect response may originate from a missed passage, a misleading chunk boundary, or synthesis that contradicts retrieved material. In KG augmentation, errors may arise from incomplete concept coverage, mistaken entity mapping, wrong edge direction, or traversal that omits a necessary intermediate relation. Both approaches remain vulnerable to a generator that embellishes beyond evidence.

A practical hybrid design can route questions according to their evidence needs. A definition-focused question may use text retrieval, while a prerequisite question triggers graph traversal; a mixed question can combine graph paths with explanatory passages. Routing should not be based on superficial keyword matching alone. Instead, it can incorporate confidence estimates, query decomposition, and a fallback when the chosen source lacks sufficient coverage. A final evidence-checking stage should identify unsupported claims before an answer is shown to a learner.

An additional evaluation challenge is the distinction between an answer that is technically correct and one that is educationally appropriate. A terse statement of algorithmic complexity may satisfy an automatic reference matcher while leaving a novice unable to explain why that complexity arises. Conversely, a long explanation may sound pedagogically helpful while embedding one consequential misconception. The assessment rubric should therefore examine whether explanations identify assumptions, distinguish best-case from worst-case behavior, and present intermediate reasoning at an appropriate level. For code-related questions, reviewers should also check whether examples execute as described and whether exceptional inputs are addressed. These checks are more demanding than semantic similarity but are essential for evaluating a system intended to teach rather than merely retrieve facts.

Another design concern is the interaction between evidence quantity and answer reliability. Adding more retrieved passages can introduce contradictory statements, obsolete examples, or irrelevant details that distract the generator. Expanding a graph neighborhood can similarly expose

peripheral relationships that are true but not helpful for the current question. Evidence selection should therefore balance coverage against precision. A useful benchmark would log the retrieved identifiers, the reason each item was selected, and the claims each item supports. Such logs permit diagnosis of whether a failure occurred before or after generation and make expert audits substantially more informative.

Instructors also need governance mechanisms. Source documents should be versioned, graph relations should retain provenance, and updates should be reviewed when course requirements change. A useful assistant should communicate uncertainty, show relevant source locations where possible, and decline to infer institutional rules that are absent from its knowledge base. These safeguards support auditability but do not substitute for classroom testing of comprehension, trust, or long-term learning outcomes.



Figure 4. Conceptual matching of question structure and grounding evidence (analytical schematic; not measured results).

Comparative evaluation design

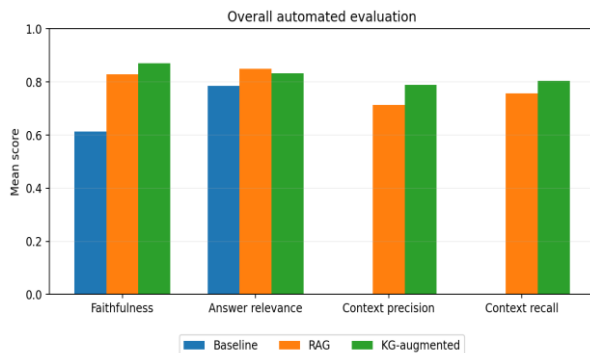


Figure 5. Proposed paired-response assessment process (methodological schematic).

6. COMPARATIVE RESULTS: CHAPTER 4 VISUAL ANALYSIS

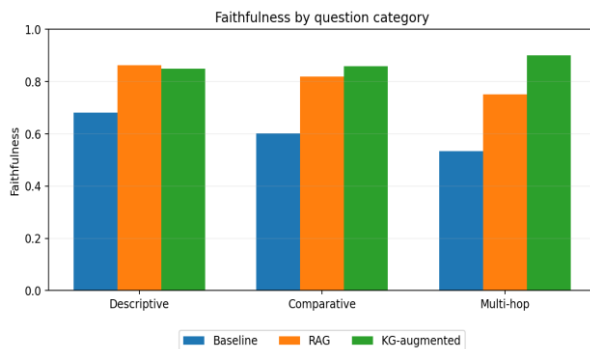
The following figures summarize the comparative numerical scenarios reported in the source study. They provide a structured basis for discussing relative performance across retrieval methods.

6.1. Figure R1. Overall automated evaluation metrics



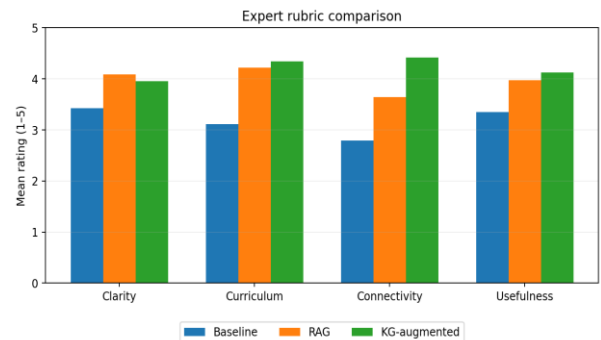
Results and Interpretation. In the reported comparison, knowledge-graph augmentation receives the highest faithfulness score (0.869), compared with passage retrieval (0.827) and the unaugmented model (0.612). Passage retrieval has a slightly higher answer-relevance score (0.849) than graph augmentation (0.831). The graph-oriented approach also shows higher context precision (0.788 versus 0.712) and context recall (0.804 versus 0.756). This pattern suggests a possible trade-off between relational evidence coverage and directly relevant textual explanations; it is not proof of superior performance.

6.2. Figure R2. Faithfulness across question categories



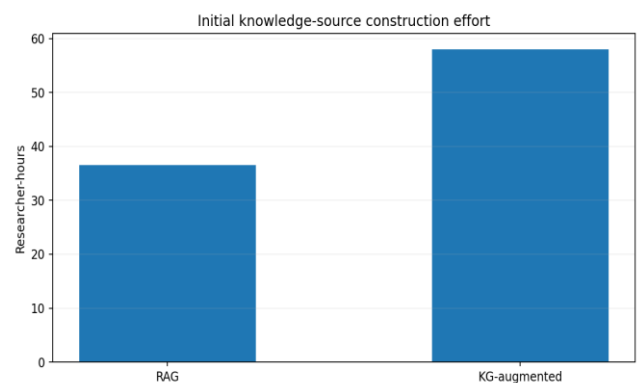
Results and Interpretation. Across descriptive questions, passage retrieval records the highest reported faithfulness (0.861), narrowly exceeding graph augmentation (0.849). For comparative questions, graph augmentation scores 0.858 against 0.819 for passage retrieval. Its clearest numerical advantage occurs for multi-hop questions (0.899 versus 0.751), where explicit links between concepts may help assemble a reasoning chain. These differences are descriptive and require evaluation on actual learner queries before statistical or causal conclusions can be drawn.

6.3. Figure R3. Expert evaluation rubric scores



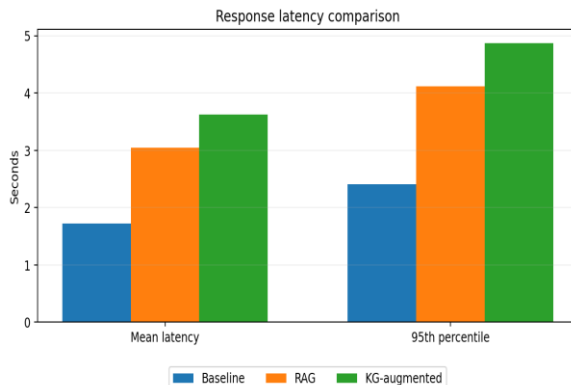
Results and Interpretation. On the five-point reported expert rubric, passage retrieval scores slightly higher for clarity (4.08 versus 3.95), while graph augmentation leads in curriculum alignment (4.34 versus 4.21), conceptual connectivity (4.41 versus 3.64), and overall usefulness (4.12 versus 3.97). The baseline scores lower on all four criteria. The strongest contrast concerns connectivity, consistent with the intended role of graph relationships in prerequisite-oriented explanations. Because the ratings are illustrative, they should not be presented as completed expert assessments.

6.4. Figure R4. Knowledge-source construction effort



Results and Interpretation. The reported setup comparison assigns 36.5 researcher-hours to the passage-retrieval knowledge source and 58.0 hours to the knowledge graph, a difference of 21.5 hours. This reflects the additional effort potentially needed to define entities, relationship types, and graph validation rules. It does not establish the actual development cost of either system; measured effort would depend on corpus size, annotation tools, existing ontologies, and staff experience.

6.5. Figure R5. End-to-end response latency



Results and Interpretation. The reported latency values rise from 1.72 seconds for the baseline to 3.05 seconds for passage retrieval and 3.62 seconds for graph augmentation at the mean. At the 95th percentile, the corresponding values are 2.41, 4.12, and 4.87 seconds. The comparison highlights a potential response-time cost of adding external evidence retrieval, especially graph traversal. Actual latency must be measured under a shared model, hardware environment, workload, and retrieval configuration.

7. LIMITATIONS AND RESEARCH AGENDA

Evidence status and limitations. The numerical comparisons reproduced here originate from constructed scenarios in the source study; underlying run-level records, independently audited annotations, and reproducible evaluation logs were not supplied. Accordingly, the charts support methodological interpretation and hypothesis formation, not claims of demonstrated model superiority. A confirmatory study should publish its question set, evidence corpus, model configuration, expert ratings, and uncertainty estimates.

The proposed design is bounded to a single technical subject and a shared generator. Outcomes could differ in history, medicine, mathematics, or institutional policy support, where knowledge structures and acceptable evidence vary. A graph is especially sensitive to ontology coverage and relation quality, while retrieval performance depends on segmentation, embeddings, ranking, and document completeness. These implementation variables require systematic ablation studies rather than one-off configuration choices.

Future research should execute the protocol with archived question-level outputs and blinded expert ratings, report paired uncertainty estimates, and publish failure examples. Further experiments should compare text-only, graph-only, and hybrid systems under matched evidence budgets; test graph traversal depth; and measure the effects of source updates. Classroom studies should separately examine whether students learn more accurately, become more

independent problem solvers, or instead develop misplaced trust in confident machine explanations.

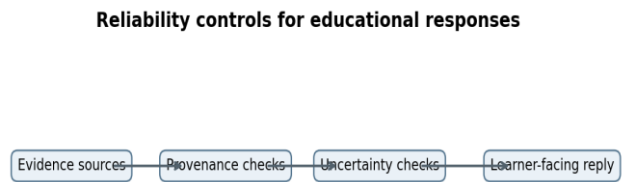


Figure 6. Proposed checks for provenance and uncertainty in learner-facing answers (design schematic).

8. CONCLUSION

Grounded educational assistants must reconcile linguistic flexibility with accountable domain knowledge. Passage-based retrieval and knowledge-graph augmentation provide complementary routes toward that objective. The former offers explanatory richness and straightforward document maintenance; the latter exposes relationships and prerequisite paths that are valuable for interconnected curricula. Neither representation, by itself, guarantees factual reliability or pedagogical effectiveness.

A fair comparison requires common instructional coverage, a fixed generator, stratified questions, evidence-sensitive automated metrics, independent expert evaluation, and explicit reporting of operational costs. The framework presented here supplies that foundation while preserving the boundary between anticipated behavior and observed findings. Its central implication is that educational chatbot design should select and combine grounding methods according to the structure of the learning task, then verify their value through reproducible evaluation rather than architectural assumptions.

REFERENCES

- 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.
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997.
- Gutiérrez, B. J., Shu, Y., Gu, Y., Yasunaga, M., & Su, Y. (2024). HippoRAG: Neurobiologically inspired long-term memory for large language models. In

- Advances in Neural Information Processing Systems (NeurIPS 2024).
arXiv:2312.10997, 2023.
<https://arxiv.org/abs/2312.10997>
4. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1-38.
5. Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274.
6. 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. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, 33, 9459-9474.
7. Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.
8. Sarmah, B., Mehta, D., Hall, B., Rao, R., Patel, S., & Pasquali, S. (2024). HybridRAG: Integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction. In *Proceedings of the 5th ACM International Conference on AI in Finance*, 608-616.
9. Winkler, R., & Söllner, M. (2018). Unleashing the potential of chatbots in education: A state-of-the-art analysis. In *Academy of Management Annual Meeting (AOM) Proceedings*, 2018(1), 15903.
10. Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are we there yet? A systematic literature review on chatbots in education. *Frontiers in Artificial Intelligence*, 4, 654924.
11. P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459-9474, 2020.
<https://arxiv.org/abs/2005.11401>
12. Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," *arXiv preprint* arXiv:2312.10997, 2023.
<https://arxiv.org/abs/2312.10997>
13. D. Edge et al., "From Local to Global: A Graph RAG Approach to Query-Focused Summarization," *arXiv preprint* arXiv:2404.16130, 2024.
<https://arxiv.org/abs/2404.16130>
14. B. Ji et al., "Survey of Hallucination in Natural Language Generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1-38, 2023.
<https://doi.org/10.1145/3571730>
15. E. Kasneci et al., "ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education," *Learning and Individual Differences*, vol. 103, article 102274, 2023.
<https://doi.org/10.1016/j.lindif.2023.102274>
16. A. Vaswani et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
<https://arxiv.org/abs/1706.03762>
17. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT*, pp. 4171-4186, 2019.
<https://doi.org/10.18653/v1/N19-1423>
18. T. Brown et al., "Language Models Are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877-1901, 2020.
<https://arxiv.org/abs/2005.14165>
19. V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," in *Proceedings of EMNLP*, pp. 6769-6781, 2020.
<https://doi.org/10.18653/v1/2020.emnlp-main.550>
20. K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "REALM: Retrieval-Augmented Language Model Pre-Training," in *Proceedings of the 37th International Conference on Machine Learning*, pp. 3929-3938, 2020.
<https://arxiv.org/abs/2002.08909>
21. N. F. Liu et al., "Lost in the Middle: How Language Models Use Long Contexts," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157-173, 2024.
https://doi.org/10.1162/tacl_a_00638
22. S. M. Mialon et al., "Augmented Language Models: A Survey," *Transactions on Machine Learning Research*, 2023.
<https://arxiv.org/abs/2302.07842>

23. Z. Jiang et al., "Active Retrieval Augmented Generation," in Proceedings of EMNLP, pp. 7969–7992, 2023. <https://arxiv.org/abs/2305.06983>
24. A. Asai et al., "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection," in Proceedings of the International Conference on Learning Representations (ICLR), 2024. <https://arxiv.org/abs/2310.11511>
25. J. Gutiérrez et al., "HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models," in Advances in Neural Information Processing Systems, 2024. <https://arxiv.org/abs/2405.14831>
26. S. Pan et al., "Unifying Large Language Models and Knowledge Graphs: A Roadmap," IEEE Transactions on Knowledge and Data Engineering, 2024. <https://arxiv.org/abs/2306.08302>
27. S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A Survey on Knowledge Graphs: Representation, Acquisition, and Applications," IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 2, pp. 494–514, 2022. <https://doi.org/10.1109/TNNLS.2021.3070843>
28. A. Hogan et al., "Knowledge Graphs," ACM Computing Surveys, vol. 54, no. 4, article 71, 2021. <https://doi.org/10.1145/3447772>
29. K. Wang et al., "Knowledge Graph Embedding: A Survey of Approaches and Applications," IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 12, pp. 2724–2743, 2017. <https://doi.org/10.1109/TKDE.2017.2754499>
30. A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston, and O. Yakhnenko, "Translating Embeddings for Modeling Multi-Relational Data," in Advances in Neural Information Processing Systems, vol. 26, 2013. <https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html>
31. S. Wollny, J. Schneider, D. Di Mitri, J. Weidlich, M. Rittberger, and H. Drachsler, "Are We There Yet? A Systematic Literature Review on Chatbots in Education," Frontiers in Artificial Intelligence, vol. 4, article 654924, 2021. <https://doi.org/10.3389/frai.2021.654924>
32. C. Okonkwo and A. Ade-Ibijola, "Chatbots Applications in Education: A Systematic Review," Computers and Education: Artificial Intelligence, vol. 2, article 100033, 2021. <https://doi.org/10.1016/j.caeai.2021.100033>
33. X. Zhai, "ChatGPT User Experience: Implications for Education," SSRN Electronic Journal, 2022. <https://doi.org/10.2139/ssrn.4312418>
34. S. Es et al., "RAGAS: Automated Evaluation of Retrieval Augmented Generation," in Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, pp. 150–158, 2024. <https://arxiv.org/abs/2309.15217>