Comparing RAG and GraphRAG for Page-Level Retrieval Question Answering on a Math Textbook

Eason Chen
Carnegie Mellon University
eason.tw.chen@gmail.com

Chuangji Li
Carnegie Mellon University
chuangjl@andrew.cmu.edu

Eric Li
Carnegie Mellon University
shizhuol@andrew.cmu.edu

Zimo Xiao
Carnegie Mellon University
zimox@andrew.cmu.edu

Jionghao Lin
The University of Hong Kong
jionghao@hku.hk

Ken Koedinger
Carnegie Mellon University
koedinger@cmu.edu

ABSTRACT

Large language models (LLMs) show promise as educational aids but often lack alignment with specific course materials. We investigate Retrieval-Augmented Generation (RAG) and GraphRAG for page-level question answering on an undergraduate mathematics textbook. Using a curated dataset of 477 question-answer pairs, each tied to a specific textbook page, we compare five embedding-based RAG models, a BM25 baseline, and GraphRAG across two metrics: retrieval accuracy (whether the correct page is retrieved) and answer quality (F1 score). Our results show that embedding-based RAG outperforms GraphRAG for page-level retrieval, with `voyage-3-large` achieving 99.4% accuracy at top-10 (bootstrap 95% CI for top-1: [.644, .728]). BM25 proves a strong baseline, outperforming several embedding models. Error analysis reveals that 63.3% of top-1 failures retrieve same-chapter content, suggesting pedagogical relevance even in failure cases. GraphRAG retrieves excessive context (∼47K tokens vs. ∼3.7K for RAG), reducing generation quality. We further replicate key experiments using an open-source local LLM (Qwen3.5-35B-A3B), finding that RAG benefits are proportionally larger for weaker models (+39% vs. +16% relative F1 improvement), an important result for cost-sensitive educational deployments. These findings inform the design of AI tutoring systems that reference specific textbook pages.

Keywords

Retrieval-Augmented Generation, GraphRAG, Mathematics Education, Question Answering, AI Tutoring

1. INTRODUCTION

During self-paced study, students frequently need to locate specific textbook pages, for example to revisit a definition they have forgotten, review a proof they found confusing, or find the relevant section for a homework problem. Large language models (LLMs) have shown notable capabilities in domains such as mathematics [1, 27], yet they often lack alignment with specific course content and may produce hallucinated sources [15, 4]. For instance, prior work has shown that LLMs tend to overestimate question quality without precise course alignment [21] and struggle with complex mathematical problems in digital learning games [24]. Retrieval-Augmented Generation (RAG) addresses these limitations by combining information retrieval with LLM generation, grounding responses in domain-specific documents [10, 31]. Recent work has demonstrated RAG's effectiveness in building LLM-based tutors across various educational settings [12, 20, 9, 16, 33, 18], though challenges remain around groundedness, user trust, and retrieval transparency [6, 17].

GraphRAG [8] extends standard RAG by constructing knowledge graphs from the corpus, capturing entity relationships and enabling more structured retrieval [5]. While research has shown GraphRAG can enhance question-answering performance in some settings [13, 19], it may also introduce excessive or irrelevant content in fine-grained retrieval tasks [28, 32]. In educational contexts, it is critical that AI tutors not only provide correct answers but also precisely reference the pages students need [7, 29], enabling verification and deeper learning [6, 17].

We address the following research question: *"To what extent can an AI-based retrieval system identify the correct textbook page for a question derived from that specific page?"* We compare embedding-based RAG, a BM25 sparse-retrieval baseline, and GraphRAG on a custom dataset of 477 question-answer pairs from an undergraduate mathematics textbook [23], evaluating retrieval accuracy, generated answer quality (F1 score), and error patterns. Our contributions include: (1) a systematic comparison of seven retrieval approaches (five embedding models, BM25, and GraphRAG) for page-level educational QA; (2) an error analysis showing that most retrieval failures surface pedagogically relevant content from nearby pages; (3) evidence that BM25 remains a competitive baseline, outperforming several neural embedding models; and (4) practical recommendations for deploying RAG-based AI tutors.

2. METHODS

2.1 Dataset

Eason Chen, Chuangji Li, Shizhuo Li, Zimo Xiao, Jionghao Lin, and Ken Koedinger. Comparing RAG and GraphRAG for Page-Level Retrieval Question Answering on a Math Textbook. In Anthony Botelho, Maria Mercedes T. Rodrigo, Adish Singla, Hiroaki Ogata, Hyojeong So, and Young Hoan Cho (eds.) Proceedings of the 19th International Conference on Educational Data Mining, Seoul, Republic of Korea, June, 2026, pp. 616–621. International Educational Data Mining Society (2026).

https://doi.org/10.5281/zenodo.21039806

We used the undergraduate textbook *An Infinite Descent into Pure Mathematics* [23] as our corpus. The textbook contains 628 pages covering topics in pure mathematics. Each page was converted from PDF to LaTeX-based Markdown via GPT Vision OCR [26, 11]. We generated one question-answer pair per page (628 total) using `gpt-4o-mini`, prompting it to produce questions representative of what students might ask during self-study, for example requesting clarification of a definition, asking how to apply a theorem, or seeking the proof of a result. Two authors, who completed undergraduate coursework using this textbook, then manually reviewed all pairs, filtering out front/back matter (reducing to 528) and eliminating poorly formed or educationally irrelevant questions, yielding a final set of 477 curated pairs that reflect the types of page-specific queries students encounter when studying mathematics independently. Each pair is labeled with its source page number, enabling precise evaluation of page-level retrieval.

2.2 RAG Pipeline

Figure 1 illustrates our pipeline, consisting of three stages: (1) *Indexing*, where pages are embedded as vectors (RAG), indexed as term-frequency vectors (BM25), or encoded as relational entities (GraphRAG); (2) *Retrieval*, where the top-k most relevant pages or entities are identified; and (3) *Generation*, where the query, prompt, and retrieved content are fed to an LLM for answer generation.

2.3 Models and Baselines

We selected five embedding models from the December 2024 Massive Text Embedding Benchmark (MTEB) leaderboard [22, 14]: `voyage-3-large` and `nv-embed-v2` were chosen for their high rankings; `gte-large` is commonly used as a fine-tuning base; `text-embedding-3-large` was included for its widespread adoption; and `multilingual-e5` for its multilingual capabilities. We also include a **BM25** sparse-retrieval baseline using Okapi BM25 with default parameters, to contextualize neural embedding gains against a classical term-matching approach that requires no training or fine-tuning. For each model, we tested retrieval at $k \in \{1, 3, 5, 10\}$.

For GraphRAG [8], we adapted the framework to expose `document_ids` for each entity and text unit, enabling page-level traceability. Specifically, we modified various GraphRAG components to store and output fields such as `document_ids` and `entity_ids`, allowing our system to link each retrieved item back to a corresponding page number. We tested GraphRAG with both `gpt-4o-mini` and `o3-mini` as the underlying LLM. Note that since GraphRAG retrieves entire entities and community summaries rather than ranked pages, we cannot directly control the number of pages returned; accuracy is evaluated based on whether the correct page's entities appear in the retrieved data.

2.4 Evaluation Metrics

We evaluate with two metrics: (1) **Retrieval Accuracy**, whether the correct source page appears in the retrieved set, averaged across all 477 queries; and (2) **F1 Score**, word-overlap F1 between the generated answer (produced by `gpt-4o-mini` given retrieved context) and the ground-truth answer. The F1 score captures the balance between precision (fraction of generated words appearing in the reference) and recall (fraction of reference words appearing in the generated output). We report bootstrap 95% confidence intervals (10,000 resamples) for key comparisons to assess statistical significance. We also conducted an error analysis of retrieval failures, categorizing top-1 misses by page distance and chapter membership to understand whether failures are near-misses or far-misses. Additionally, we tested LLM-based re-ranking of retrieved pages using `gpt-4o-mini`, where the model reorders retrieved pages by estimated relevance to the query.

Table 1: Retrieval accuracy for embedding-based RAG models, BM25, and GraphRAG. GraphRAG retrieves entities rather than ranked pages, so only overall accuracy is reported.

Model	Top-1	Top-3	Top-5	Top-10
voyage-3-large	**.686**	**.910**	**.958**	**.994**
nvidia/nv-embed-v2	.585	.843	.912	.964
BM25	.579	.780	.855	.929
OpenAI emb-3-large	.549	.811	.893	.933
gte-large	.461	.711	.799	.881
multilingual-e5	.457	.702	.795	.870
GraphRAG (o3-mini)		.914		
GraphRAG (4o-mini)		.845		

3. RESULTS

3.1 Retrieval Accuracy

Table 1 presents retrieval accuracy results. Among all models, `voyage-3-large` achieves the highest accuracy across all top-k settings, reaching 99.4% at top-10. Bootstrap 95% CIs for top-1 accuracy confirm that `voyage-3-large` ([.644, .728]) significantly outperforms `nv-embed-v2` ([.541, .629]), as these intervals do not overlap at top-1 or top-3.

BM25 proves a strong sparse-retrieval baseline, achieving top-1 accuracy of .579, outperforming both `gte-large` (.461) and `multilingual-e5` (.457) at all k values despite using no learned representations. This highlights the importance of including classical baselines in retrieval evaluations. Despite strong MTEB leaderboard rankings, `multilingual-e5` underperformed on this domain-specific task, underscoring that benchmark performance does not always transfer to specialized mathematical domains.

GraphRAG achieves 84.5% with `gpt-4o-mini` and 91.4% with `o3-mini`. All five embedding models surpass GraphRAG's accuracy at top-5 or above. The gap between the two GraphRAG configurations suggests that the underlying LLM quality substantially affects graph-based retrieval performance.

3.2 Generated Answer Quality

Table 2 shows F1 scores for answer generation. The no-retrieval baseline (`gpt-4o-mini` alone) achieves 0.475. All RAG configurations improve upon this, with F1 scores ranging from 0.514 to 0.552, confirming that retrieval grounding consistently benefits domain-specific QA. GraphRAG achieves F1 of $\sim$0.524, comparable to but slightly below the best embedding-based results.

Interestingly, `nv-embed-v2` achieves the highest top-1 F1

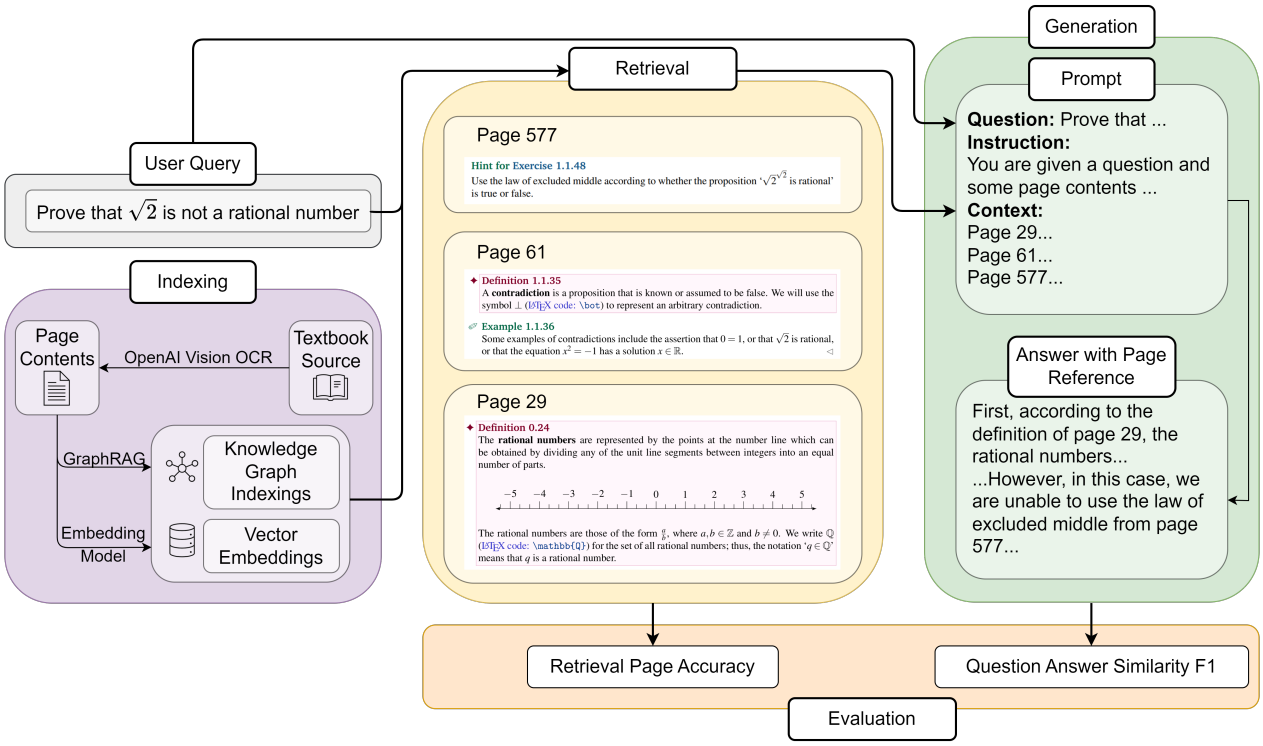

Figure 1: Our RAG pipeline: Indexing, Retrieval, and Generation stages with two evaluation metrics.

Table 2: F1 scores for generated answers. Baseline is `gpt-4o-mini` without retrieval (F1 = 0.475).

Model	Top-1	Top-3	Top-5	Top-10
voyage-3-large	.523	.543	.544	.547
nvidia/nv-embed-v2	**.537**	.542	.541	.539
OpenAI emb-3-large	.531	.549	.533	.543
gte-large	.526	.533	.550	**.552**
multilingual-e5	.514	.534	.541	.535
GraphRAG (o3-mini)		.524		
GraphRAG (4o-mini)		.525		
No retrieval		.475		

(.537) despite lower retrieval accuracy than `voyage-3-large`, suggesting retrieval accuracy and generation quality do not always correlate, because near-miss retrievals still provide useful context. Increasing k generally does not improve F1, as additional pages introduce noise. GraphRAG's ∼47K tokens of context per query (vs. ∼3.7K for top-5 RAG) overwhelms the generator, suggesting that *retrieval precision* matters more than *recall* for generation quality.

3.3 Re-Ranking

LLM-based re-ranking with `gpt-4o-mini` yielded mixed results. Re-ranking top-5 pages degraded `voyage-3-large`'s top-1 accuracy from .686 to .593, while re-ranking top-10 caused further degradation across all models. Critically, we observed hallucinated page numbers in re-ranking outputs: the LLM occasionally referenced pages not in the retrieved set. These findings suggest that LLM-based re-ranking is unreliable for page-level retrieval, particularly with larger candidate sets.

3.4 Error Analysis

We examined the 150 top-1 failures from `voyage-3-large`, categorizing each by page distance and chapter membership. Of these, 46.0% retrieved an adjacent page (distance ≤ 2), 17.3% retrieved a same-chapter but non-adjacent page, and 35.3% were far misses (different chapter). Overall, 63.3% of failures retrieved same-chapter content, with a median page distance of just 3 pages. This suggests that even "incorrect" retrievals often surface pedagogically relevant material. For example, a student asking about a proof may be directed to the adjacent page containing the theorem statement, content that is still valuable for building understanding and that an instructor might recommend reviewing first. The high proportion of near-misses reflects the textbook's pedagogical structure, where definitions, theorems, and proofs are deliberately organized across consecutive pages to support incremental learning. The 35.3% far misses, where shared terminology across chapters causes cross-topic confusion, represent the genuinely problematic failures that could misdirect students.

3.5 Open-Source LLM Replication

To assess whether our findings generalize beyond commercial APIs, we replicated key experiments using Qwen3.5-35B-A3B [30], an open-source 36B Mixture-of-Experts model running locally on 2× NVIDIA TITAN RTX GPUs with Q4_K_M quantization via Ollama. For retrieval, we tested `nomic-embed-text` [25], an open-source embedding model, also served locally.

Table 3: F1 scores: commercial API vs. open-source local LLM.

Configuration	gpt-4o-mini	Qwen3.5-35B
No retrieval	0.475	0.276
Best RAG + generation	0.552	0.384
Relative improvement	+16%	+39%

Table 3 compares generation quality (F1) between the commercial `gpt-4o-mini` and the local Qwen3.5 setup. Without retrieval, Qwen3.5 achieves substantially lower F1 (0.276 vs. 0.475). However, adding RAG with `nomic-embed-text` retrieval improves Qwen3.5's F1 to 0.384, a **39% relative improvement**, compared to only 16% for `gpt-4o-mini` with its best embedding model. This suggests that *weaker models benefit disproportionately from retrieval grounding*, as they rely more heavily on external context to compensate for limited parametric knowledge.

For retrieval accuracy, `nomic-embed-text` achieved top-1 accuracy of 45.1%, top-3 of 71.3%, and top-10 of 88.7%, below the best commercial embeddings but still serviceable. Notably, BM25 (61.2% top-1) again outperforms this neural embedding model, reinforcing BM25's strength as a zero-cost baseline.

We also attempted GraphRAG v3.0.6 indexing with Qwen3.5, successfully extracting 4,528 entities, 10,828 relationships, and 1,107 communities. However, query-time generation failed due to incompatibilities between GraphRAG's OpenAI-compatible API expectations and Qwen3.5's thinking-mode architecture. This demonstrates that while local LLM indexing for GraphRAG is feasible, end-to-end integration with open-source models remains challenging.

4. DISCUSSION AND CONCLUSION

Our findings yield several key insights for designing RAG-based educational AI systems.

RAG consistently improves over baseline LLMs. All RAG configurations produced higher F1 scores than the no-retrieval baseline (0.475 vs. 0.514 to 0.552), confirming that grounding LLM responses in textbook content improves answer quality for domain-specific question answering. This improvement is meaningful in educational settings where students benefit from answers grounded in their assigned course materials rather than the LLM's general pre-training data, which may contain incorrect or outdated mathematical content [3, 2].

Embedding-based RAG outperforms GraphRAG for page-level retrieval. GraphRAG retrieves excessive context (∼47K tokens vs. ∼3.7K for top-5 RAG), leading to lower precision and comparable or lower F1 [28, 32]. This reflects a *task-method mismatch*: GraphRAG was designed for global summarization and multi-hop reasoning, not page-level lookup. For educational applications where students need a specific page reference, the simpler embedding-based approach is more appropriate. However, GraphRAG may excel at cross-concept questions (e.g., "How do induction and well-ordering relate across chapters?") where its entity-based structure could surface connections that page-level retrieval would miss.

BM25 is a competitive baseline. BM25 outperforms two of five embedding models at all k values, partly due to high lexical overlap (72% unigram) between questions and source pages. This underscores the importance of including classical baselines, as neural embeddings do not always yield proportional gains, particularly when questions and documents share substantial vocabulary.

Most retrieval failures are near-misses with pedagogical value. Our error analysis reveals that 63.3% of top-1 failures retrieve same-chapter content (median distance: 3 pages). From an educational perspective, this means a student directed to the "wrong" page is still reading closely related material in most cases, analogous to how a tutor might say "look at the previous page first" before addressing the specific question. The top-1 accuracy of 68.6% may appear insufficient, but the near-miss pattern means the system rarely sends students to entirely irrelevant content.

Re-ranking introduces more risk than benefit. LLM-based re-ranking degraded performance for top-performing retrievers and introduced hallucinated page references, suggesting general-purpose LLMs struggle to distinguish between mathematically similar pages. Deployments requiring re-ranking should explore fine-tuned cross-encoder models or constrained decoding.

Open-source LLMs are viable but benefit more from RAG. RAG provides proportionally larger gains for weaker models (+39% vs. +16% relative F1 improvement), suggesting that retrieval grounding partially compensates for reduced model capability. This is significant for educational settings where commercial API costs are prohibitive, and aligns with EDM 2026's "Across Borders" theme, because institutions can deploy useful textbook-grounded QA systems on local hardware rather than relying on commercial APIs. Our results show that a locally hosted open-source model combined with BM25 retrieval (requiring no GPU for the retrieval component) can deliver meaningful QA quality at near-zero marginal cost.

Practical recommendations. (1) Use top-5 RAG retrieval as default (accuracy .958, ∼3.7K tokens); (2) present top-3 candidate pages to students (accuracy >91%); (3) avoid LLM re-ranking due to hallucination risk; (4) consider adjacent-page windowing to capture theorem-proof pairs spanning page boundaries.

Limitations and Future Work. While questions were curated by course-familiar authors to reflect realistic student queries, their high lexical overlap with source pages (72% unigram) may inflate retrieval accuracy relative to more diverse student phrasings; future work should supplement with authentic questions from discussion forums or office hours. Word-overlap F1 is a weak proxy for math QA, as mathematically equivalent expressions can score poorly; LLM-as-judge or symbolic equivalence checking would better capture correctness. Extending to multiple STEM textbooks, investigating hybrid BM25+dense retrieval, and conducting classroom studies measuring whether page references improve comprehension and student trust are important next steps. Notably, `multilingual-e5`'s underperformance despite strong general benchmarks raises questions about deploying these systems for non-English textbooks, an important consideration for

educational equity across linguistic borders.

5. REFERENCES

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[2] E. Chen, S. Judicke, K. Beigh, X. Tang, I. Wang, N. Yuan, Z. Xiao, C. Li, S. Li, R. Luttmer, S. Singh, and K. R. Koedinger. Chat-based support alone may not be enough: Comparing conversational and embedded llm feedback for mathematical proof learning. *arXiv preprint arXiv:2602.18807*, 2026.

[3] E. Chen, S. Judicke, K. Beigh, X. Tang, Z. Xiao, C. Li, S. Li, R. Luttmer, S. Singh, and K. R. Koedinger. Generative ai alone may not be enough: Evaluating ai support for learning mathematical proof. *arXiv preprint arXiv:2509.16778*, 2025.

[4] E. Chen, D. Wang, L. Xu, C. Cao, X. Fang, and J. Lin. A systematic review on prompt engineering in large language models for k-12 stem education. *arXiv preprint arXiv:2410.11123*, 2024.

[5] X. Chen, S. Jia, and Y. Xiang. A review: Knowledge reasoning over knowledge graph. *Expert systems with applications*, 141:112948, 2020.

[6] S. Chiesurin, D. Dimakopoulos, M. A. Sobrevilla Cabezudo, A. Eshghi, I. Papaioannou, V. Rieser, and I. Konstas. The dangers of trusting stochastic parrots: Faithfulness and trust in open-domain conversational question answering. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 947–959, Toronto, Canada, July 2023. Association for Computational Linguistics.

[7] W. Dai, Y.-S. Tsai, J. Lin, A. Aldino, H. Jin, T. Li, D. Gašević, and G. Chen. Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence*, 7:100299, 2024.

[8] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson. From local to global: A graph rag approach to query-focused summarization, 2024.

[9] T. Feng, S. Liu, and D. Ghosal. Courseassist: Pedagogically appropriate ai tutor for computer science education, 2024.

[10] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang. Retrieval-augmented generation for large language models: A survey, 2024.

[11] A. Ghiriti, W. Göderle, and R. Kern. Exploring the capabilities of gpt4-vision as ocr engine. In *International Conference on Theory and Practice of Digital Libraries*, pages 3–12. Springer, 2024.

[12] Z. F. Han, J. Lin, A. Gurung, D. R. Thomas, E. Chen, C. Borchers, S. Gupta, and K. R. Koedinger. Improving assessment of tutoring practices using retrieval-augmented generation. *arXiv preprint arXiv:2402.14594*, 2024.

[13] X. He, Y. Tian, Y. Sun, N. Chawla, T. Laurent, Y. LeCun, X. Bresson, and B. Hooi. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. *Advances in Neural Information Processing Systems*, 37:132876–132907, 2024.

[14] Hugging Face. MTEB leaderboard. `https://huggingface.co/spaces/mteb/leaderboard`, 2025. [Accessed 20-12-2024].

[15] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

[16] G. Lang and T. Gürpınar. AI-Powered Learning Support: A study of retrieval-augmented generation (RAG) chatbot effectiveness in an online course. *Information Systems Education Journal*, 23(2):4–13, 2025.

[17] Z. Levonian, C. Li, W. Zhu, A. Gade, O. Henkel, M.-E. Postle, and W. Xing. Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference, 2023.

[18] H. Li, T. Xu, C. Zhang, E. Chen, J. Liang, X. Fan, H. Li, J. Tang, and Q. Wen. Bringing generative ai to adaptive learning in education. *arXiv preprint arXiv:2402.14601*, 2024.

[19] J. Lin, S. Mai, B. Bu, M. He, and X. Wang. Research on the application of stem practical teaching based on rag knowledge graph and large models. In *7th International Conference on Educational Technology Management*, pages 520–527, 2024.

[20] C. Mitra, M. Miroyan, R. Jain, V. Kumud, G. Ranade, and N. Norouzi. Retllm-e: Retrieval-prompt strategy for question-answering on student discussion forums. In *Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)*, 2024.

[21] S. Moore, H. A. Nguyen, N. Bier, T. Domadia, and J. Stamper. Assessing the quality of student-generated short answer questions using gpt-3. In *European conference on technology enhanced learning*, pages 243–257. Springer, 2022.

[22] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022.

[23] C. Newstead. *An Infinite Descent into Pure Mathematics*. 2024. Last updated on Wednesday 29th May 2024.

[24] H. A. Nguyen, H. Stec, X. Hou, S. Di, and B. M. McLaren. Evaluating chatgpt's decimal skills and feedback generation in a digital learning game. In *European conference on technology enhanced learning*, pages 278–293. Springer, 2023.

[25] Z. Nussbaum, J. X. Morris, B. Duderstadt, and A. Mulyar. Nomic embed: Training a reproducible long context text embedder. *arXiv preprint arXiv:2402.01613*, 2024.

[26] OpenAI. Openai vision model documentation. `https://platform.openai.com/docs/guides/vision`, 2024. [Accessed 20-02-2025].

[27] e. a. OpenAI. Gpt-4 technical report, 2024.

[28] B. Peng, Y. Zhu, Y. Liu, X. Bo, H. Shi, C. Hong, Y. Yan, and Y. Li. Graph retrieval-augmented generation: A survey. *ACM Transactions on Information Systems*, 2025.

[29] A. Scarlatos, D. Smith, S. Woodhead, and A. Lan. Improving the validity of automatically generated feedback via reinforcement learning. In *International Conference on Artificial Intelligence in Education*, pages 280–294. Springer, 2024.

[30] Q. Team. Qwen3.5 technical report, 2026. Accessed: 2026-03-28.

[31] S. Wang, T. Xu, H. Li, C. Zhang, J. Liang, J. Tang, P. S. Yu, and Q. Wen. Large language models for education: A survey and outlook, 2024.

[32] Q. Zhang, S. Chen, Y. Bei, Z. Yuan, H. Zhou, Z. Hong, J. Dong, H. Chen, Y. Chang, and X. Huang. A survey of graph retrieval-augmented generation for customized large language models. *arXiv preprint arXiv:2501.13958*, 2025.

[33] C. Q. Zhao, J. Cao, E. Chen, K. R. Koedinger, and J. Lin. Slideitright: Using ai to find relevant slides and provide feedback for open-ended questions. In *International Conference on Artificial Intelligence in Education*, pages 378–392, 2025.