DAFTAR PUSTAKA

[1] D. Al-Fraihat, A. M. Alshahrani, M. Alzaidi, A. A. Shaikh, M. Al-Obeidallah, and M. Al-Okaily, "Exploring students' perceptions of the design and use of the moodle learning management system," *Computers in Human Behavior Reports*, vol. 18, p. 100685, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2451958825001009

[2] S. H. P. W. Gamage, J. R. Ayres, M. B. Behrend, and E. J. Smith, "Optimising moodle quizzes for online assessments," *International Journal of STEM Education*, vol. 6, no. 1, p. 27, 2019. [Online]. Available: https://doi.org/10.1186/s40594-019-0181-4

[3] K. Fischer, A. M. Sullivan, A. P. Cohen, R. W. King, B. A. Cockrill, and H. C. Besche, "Using cognitive load theory to evaluate and improve preparatory materials and study time for the flipped classroom," *BMC Medical Education*, vol. 23, no. 1, p. 345, 2023. [Online]. Available: https://doi.org/10.1186/s12909-023-04325-x

[4] M. J. Rudolph, K. K. Daugherty, M. E. Ray, V. P. Shuford, L. Lebovitz, and M. V. DiVall, "Best practices related to examination item construction and post-hoc review," *American Journal of Pharmaceutical Education*, vol. 83, no. 7, 2019. [Online]. Available: https://doi.org/10.5688/ajpe7204

[5] A. Agrawal, M. Suzgun, L. Mackey, and A. Kalai, "Do language models know when they're hallucinating references?" in *Findings of the Association for Computational Linguistics: EACL 2024*, Y. Graham and M. Purver, Eds. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 912–928. [Online]. Available: https://aclanthology.org/2024.findings-eacl.62/

[6] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[7] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 6769–6781. [Online]. Available: https://aclanthology.org/2020.emnlp-main.550/

[8] A. Gan, H. Yu, K. Zhang, Q. Liu, W. Yan, Z. Huang, S. Tong, and G. Hu, "Retrieval augmented generation evaluation in the era of large language models: A comprehensive survey," 2025. [Online]. Available: https://arxiv.org/abs/2504.14891

[9] C. Grévisse, M. A. S. Pavlou, and J. G. Schneider, "Docimological quality analysis of llm-generated multiple choice questions in computer science and medicine," *SN Computer Science*, vol. 5, no. 5, p. 636, 2024. [Online]. Available: https://doi.org/10.1007/s42979-024-02963-6

[10] P. N, R. T, M. Thushara, K. A. Krishna, and P. V, "Retrieval-augmented generation for multiple-choice questions and answers generation," *Procedia Computer Science*, vol. 259, pp. 504–511, 2025, sixth International Conference on Futuristic Trends in Networks and Computing Technologies (FTNCT06), held in Uttarakhand, India. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050925010968

[11] D. Thüs, S. Malone, and R. Brünken, "Exploring generative ai in higher education: a rag system to enhance student engagement with scientific literature," *Frontiers in Psychology*, vol. Volume 15 - 2024, 2024. [Online]. Available: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2024.1474892

[12] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009. [Online]. Available: http://dx.doi.org/10.1561/1500000019

[13] S. Bruch, S. Gai, and A. Ingber, "An analysis of fusion functions for hybrid retrieval," *ACM Trans. Inf. Syst.*, vol. 42, no. 1, Aug. 2023. [Online]. Available: https://doi.org/10.1145/3596512

[14] X. Wang, Z. Wang, X. Gao, F. Zhang, Y. Wu, Z. Xu, T. Shi, Z. Wang, S. Li, Q. Qian, R. Yin, C. Lv, X. Zheng, and X. Huang, "Searching for best practices in retrieval-augmented generation," 2024. [Online]. Available: https://arxiv.org/abs/2407.01219

[15] J. Carbonell and J. Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '98. New York, NY, USA: Association for Computing Machinery, 1998, p. 335–336. [Online]. Available: https://doi.org/10.1145/290941.291025

[16] Y. Yu, W. Ping, Z. Liu, B. Wang, J. You, C. Zhang, M. Shoeybi, and B. Catanzaro, "Rankrag: unifying context ranking with retrieval-augmented generation in llms," in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, ser. NIPS '24. Red Hook, NY, USA: Curran Associates Inc., 2025.

[17] G. Biancini, A. Ferrato, and C. Limongelli, "Multiple-choice question generation using large language models: Methodology and educator insights," in *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, ser. UMAP Adjunct '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 584–590. [Online]. Available: https://doi.org/10.1145/3631700.3665233

[18] C. N. Hang, C. Wei Tan, and P.-D. Yu, "Mcqgen: A large language model-driven mcq generator for personalized learning," *IEEE Access*, vol. 12, pp. 102 261–102 273, 2024.

[19] D. Ru, L. Qiu, X. Hu, T. Zhang, P. Shi, S. Chang, C. Jiayang, C. Wang, S. Sun, H. Li, Z. Zhang, B. Wang, J. Jiang, T. He, Z. Wang, P. Liu, Y. Zhang, and Z. Zhang, "Ragchecker: a fine-grained framework for diagnosing retrieval-augmented generation," in *Proceedings of the 38th International Conference on Neural Information*

Processing Systems, ser. NIPS '24. Red Hook, NY, USA: Curran Associates Inc., 2025.

[20] M. Maryamah, M. M. Irfani, E. B. Tri Raharjo, N. A. Rahmi, M. Ghani, and I. K. Raharjana, "Chatbots in academia: A retrieval-augmented generation approach for improved efficient information access," in *2024 16th International Conference on Knowledge and Smart Technology (KST)*, 2024, pp. 259–264.

[21] S. Es, J. James, L. Espinosa Anke, and S. Schockaert, "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*, N. Aletras and O. De Clercq, Eds. St. Julians, Malta: Association for Computational Linguistics, Mar. 2024, pp. 150–158. [Online]. Available: https://aclanthology.org/2024.eacl-demo.16/

[22] Z. Li, Z. Wang, W. Wang, K. Hung, H. Xie, and F. L. Wang, "Retrieval-augmented generation for educational application: A systematic survey," *Computers and Education: Artificial Intelligence*, vol. 8, p. 100417, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666920X25000578

[23] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[24] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, p. 613–620, Nov. 1975. [Online]. Available: https://doi.org/10.1145/361219.361220

[25] M. Xu, W. Zhou, Y. Babakhin, G. Moreira, R. Ak, R. Osmulski, B. Liu, E. Oldridge, and B. Schifferer, "Omni-embed-nemotron: A unified multimodal retrieval model for text, image, audio, and video," 2025. [Online]. Available: https://arxiv.org/abs/2510.03458

[26] D. Ford, "Introducing contextual retrieval," Anthropic, sep 2024, accessed: Dec. 7, 2025. [Online]. Available: https://www.anthropic.com/engineering/contextual-retrieval

[27] Perplexity, "Architecting and evaluating an ai-first search api," Perplexity Research, sep 2025, accessed: Dec. 7, 2025. [Online]. Available: https://research.perplexity.ai/articles/architecting-and-evaluating-an-ai-first-search-api

[28] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: Bm25 and beyond," *Found. Trends Inf. Retr.*, vol. 3, no. 4, p. 333–389, Apr. 2009. [Online]. Available: https://doi.org/10.1561/1500000019

[29] P. Mandikal and R. Mooney, "Sparse meets dense: A hybrid approach to enhance scientific document retrieval," *ArXiv*, vol. abs/2401.04055, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:266844268

[30] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2021.

[31] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Conference on Empirical Methods in Natural Language Processing*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:52051958

[32] T. Kiss and J. Strunk, "Unsupervised multilingual sentence boundary detection," *Computational Linguistics*, vol. 32, no. 4, pp. 485–525, 12 2006. [Online]. Available: https://doi.org/10.1162/coli.2006.32.4.485

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[34] M. Dougiamas and P. Taylor, "Moodle: Using learning communities to create an open source course management system," in *World Conference on Educational Multimedia, Hypermedia and Telecommunications (EDMEDIA) 2003*, L. Alves, D. Barros, and A. Okada, Eds., Honolulu, Hawaii, USA, 20030623 - 20030628.

[35] Moodle Pty Ltd, "Moodle developer resources," https://moodledev.io/, 2025, accessed: Sep. 16, 2025.

[36] L. W. Anderson and D. R. Krathwohl, *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*, complete edition ed. New York: Longman, 2001.

[37] N. Suwono, J. Chen, T. Hung, T.-H. Huang, I.-B. Liao, Y.-H. Li, L.-W. Ku, and S.-H. Sun, "Location-aware visual question generation with lightweight models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 1415–1432. [Online]. Available: https://aclanthology.org/2023.emnlp-main.88/

[38] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-eval: Nlg evaluation using gpt-4 with better human alignment," in *Conference on Empirical Methods in Natural Language Processing*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257804696

[39] J. Dean and L. A. Barroso, "The tail at scale," *Commun. ACM*, vol. 56, no. 2, p. 74–80, Feb. 2013. [Online]. Available: https://doi.org/10.1145/2408776.2408794

[40] J. Nielsen, *Usability Engineering*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994.

[41] R. Qu, R. Tu, and F. S. Bao, "Is semantic chunking worth the computational cost?" in *Findings of the Association for Computational Linguistics: NAACL 2025*, L. Chiruzzo, A. Ritter, and L. Wang, Eds. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 2155–2177. [Online]. Available: https://aclanthology.org/2025.findings-naacl.114/

[42] S. R. Bhat, M. Rudat, J. Spiekermann, and N. Flores-Herr, "Rethinking chunk size for long-document retrieval: A multi-dataset analysis," 2025. [Online]. Available: https://arxiv.org/abs/2505.21700

[43] LlamaIndex, "Evaluating the ideal chunk size for a rag system using llamaindex," https://www.llamaindex.ai/blog/evaluating-the-ideal-chunk-size-for-a-rag-system-using-llamaindex-6207e5d3fec5, 2024, accessed: Aug. 26, 2025.

[44] Clarifai. (2025) Openai gpt-oss benchmarks: How it compares to glm-4.5, qwen3, deepseek, and kimi k2. Accessed: Aug. 26, 2025. [Online]. Available: https://www.clarifai.com/blog/openai-gpt-oss-benchmarks-how-it-compares-to-glm-4.5-qwen3-deepseek-and-kimi-k2

[45] T. Hardware. (2025) Openai intros two open-weight models that can run on consumer gpus (covers moe & efficiency). Accessed: Aug. 26, 2025. [Online]. Available: https://www.tomshardware.com/tech-industry/artificial-intelligence/openai-intros-two-lightweight-open-model-language-models-that-can-run-on-consumer-gpus-optim

[46] U. of Wisconsin. (2025) Writing good multiple choice test questions. Accessed: Aug. 26, 2025. [Online]. Available: https://wisc.pb.unizin.org/mtle/chapter/writing-good-multiple-choice-test-questions/

[47] J. Nielsen. (1993) Response times: The 3 important limits. Accessed: 2025-08-27. [Online]. Available: https://www.nngroup.com/articles/response-times-3-important-limits/

[48] T. H. (2023) Statistics behind latency metrics: Understanding p90, p95, and p99. Accessed: 2025-08-27. [Online]. Available: https://medium.com/tuanhdotnet/statistics-behind-latency-metrics-understanding-p90-p95-and-p99-dc87420d505d

[49] Zilliz. (2024) Why is tail latency (p95/p99) often more important than average latency for evaluating the performance of a vector search in user-facing applications? Accessed: 2025-08-27. [Online]. Available: https://zilliz.com/ai-faq/why-is-tail-latency-p95p99-often-more-important-than-average-latency-for-evaluating-the-perform