DAFTAR PUSTAKA

[1] D. Reinsel, J. Gantz, and J. Rydning, "Data age 2025: The evolution of data to life-critical. Don't focus on big data; focus on the data that's big," *IDC White Pap.*, no. s 3, 2017.

[2] M. Sanderson and W. B. Croft, "The History of Information *Retrieval* Research," *Proc. IEEE*, vol. 100, no. Special Centennial Issue, pp. 1444–1451, 2012, doi: 10.1109/JPROC.2012.2189916.

[3] J. Guo *et al.*, "A Deep Look into neural ranking models for information *retrieval*," *Inf. Process. Manag.*, vol. 57, no. 6, p. 102067, Nov. 2020, doi: 10.1016/j.ipm.2019.102067.

[4] G. Salton and C. Buckley, "Term-weighting approaches in automatic *text retrieval*," *Inf. Process. Manag.*, vol. 24, no. 5, pp. 513–523, Jan. 1988, doi: 10.1016/0306-4573(88)90021-0.

[5] S. Robertson and H. Zaragoza, *The Probabilistic Relevance Framework: BM25 and Beyond*. Now Publishers Inc, 2009.

[6] B. Miutra and N. Craswell, "An Introduction to Neural Information *Retrieval*," *Found. Trends Inf. Retr.*, vol. 13, no. 1, pp. 1–126, Dec. 2018, doi: 10.1561/1500000061.

[7] F. Li *et al.*, "CoT-RAG: Integrating Chain of Thought and *Retrieval*-Augmented Generation to Enhance Reasoning in Large Language Models," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds., Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 3119–3171. doi: 10.18653/v1/2025.findings-emnlp.168.

[8] R. Yang *et al.*, "*Retrieval*-augmented generation for generative artificial intelligence in health care," *Npj Health Syst.*, vol. 2, no. 1, p. 2, Jan. 2025, doi: 10.1038/s44401-024-00004-1.

[9] A. Mallen, A. Asai, V. Zhong, R. Das, D. Khashabi, and H. Hajishirzi, "When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 9802–9822. doi: 10.18653/v1/2023.acl-long.546.

[10] P. Lewis *et al.*, "*Retrieval*-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 9459–9474. Accessed: Apr. 24, 2026. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[11] C. Sharma, "*Retrieval*-Augmented Generation: A Comprehensive Survey of Architectures, Enhancements, and Robustness Frontiers," May 28, 2025, *arXiv*: arXiv:2506.00054. doi: 10.48550/arXiv.2506.00054.

[12] P. Sarthi, S. Abdullah, A. Tuli, S. Khanna, A. Goldie, and C. D. Manning, "RAPTOR: Recursive Abstractive Processing for Tree-Organized *Retrieval*,"

in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?*id*=GN921JHCRw

[13] S. Xu, L. Pang, H. Shen, X. Cheng, and T.-S. Chua, "Search-in-the-Chain: Interactively Enhancing Large Language Models with Search for Knowledge-intensive Tasks," in *Proceedings of the ACM Web Conference 2024*, in WWW '24. New York, NY, USA: Association for Computing Machinery, 2024, pp. 1362–1373. doi: 10.1145/3589334.3645363.

[14] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "*Retrieval* Augmentation Reduces Hallucination in Conversation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, and S. W. Yih, Eds., Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3784–3803. doi: 10.18653/v1/2021.findings-emnlp.320.

[15] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms, 3rd Edition*. MIT Press, 2009. [Online]. Available: http://mitpress.mit.edu/books/introduction-algorithms

[16] T. Kočiský *et al.*, "The NarrativeQA Reading Comprehension Challenge," *Trans. Assoc. Comput. Linguist.*, vol. 6, pp. 317–328, Jan. 2018, doi: 10.1162/tacl_a_00023.

[17] P. Dasigi, K. Lo, I. Beltagy, A. Cohan, N. A. Smith, and M. Gardner, "A *Dataset* of Information-Seeking Questions and Answers Anchored in Research Papers," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds., Online: Association for Computational Linguistics, Jun. 2021, pp. 4599–4610. doi: 10.18653/v1/2021.naacl-main.365.

[18] R. Y. Pang *et al.*, "QuALITY: Question Answering with Long Input Texts, Yes!," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds., Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 5336–5358. doi: 10.18653/v1/2022.naacl-main.391.

[19] K. Goel and M. Chandak, "HIRO: Hierarchical Information *Retrieval* Optimization," Sep. 04, 2024, *arXiv*: arXiv:2406.09979. doi: 10.48550/arXiv.2406.09979.

[20] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "MTEB: Massive Text Embedding Benchmark," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos and I. Augenstein, Eds., Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 2014–2037. doi: 10.18653/v1/2023.eacl-main.148.

[21] H. Su *et al.*, "BRIGHT: A Realistic and Challenging Benchmark for Reasoning-Intensive *Retrieval*," Mar. 26, 2025, *arXiv*: arXiv:2407.12883. doi: 10.48550/arXiv.2407.12883.

[22] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, "*Interleaving Retrieval* with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 10014–10037. doi: 10.18653/v1/2023.acl-long.557.

[23] Z. Jiang *et al.*, "Active *Retrieval* Augmented Generation," 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. 7969–7992. doi: 10.18653/v1/2023.emnlp-main.495.

[24] M. Arslan, H. Ghanem, S. Munawar, and C. Cruz, "A Survey on RAG with LLMs," *Procedia Comput. Sci.*, vol. 246, pp. 3781–3790, Jan. 2024, doi: 10.1016/j.procs.2024.09.178.

[25] H. Liu *et al.*, "HopRAG: Multi-Hop Reasoning for Logic-Aware *Retrieval*-Augmented Generation," in *Findings of the Association for Computational Linguistics: ACL 2025*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds., Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 1897–1913. doi: 10.18653/v1/2025.findings-acl.97.

[26] J. H. Clark *et al.*, "TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages," *Trans. Assoc. Comput. Linguist.*, vol. 8, pp. 454–470, 2020, doi: 10.1162/tacl_a_00317.

[27] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge: Cambridge university press, 2008.

[28] C. Zhai and S. Massung, *Text data management and analysis: a practical introduction to information retrieval and text mining*. Morgan & Claypool, 2016.

[29] V. Karpukhin *et al.*, "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. doi: 10.18653/v1/2020.emnlp-main.550.

[30] J. Johnson, M. Douze, and H. Jégou, "Billion-Scale Similarity Search with GPUs," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, Jul. 2021, doi: 10.1109/TBDATA.2019.2921572.

[31] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Apr. 25, 2026. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[32] R. Bommasani *et al.*, "On the Opportunities and Risks of Foundation Models," Jul. 12, 2022, *arXiv*: arXiv:2108.07258. doi: 10.48550/arXiv.2108.07258.

[33] Z. Ji *et al.*, "Survey of Hallucination in Natural Language Generation," *ACM Comput Surv*, vol. 55, no. 12, p. 248:1-248:38, Mar. 2023, doi: 10.1145/3571730.

[34] S. Zhang *et al.*, “Instruction Tuning for Large Language Models: A Survey,” *ACM Comput Surv*, vol. 58, no. 7, Jan. 2026, doi: 10.1145/3777411.

[35] J. Lin *et al.*, “AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration,” *Proc. Mach. Learn. Syst.*, vol. 6, pp. 87–100, May 2024.

[36] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. doi: 10.18653/v1/D19-1410.

[37] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, “Multilingual E5 Text Embeddings: A Technical Report,” Feb. 08, 2024, *arXiv*: arXiv:2402.05672. doi: 10.48550/arXiv.2402.05672.

[38] “What is RAG? - *Retrieval*-Augmented Generation AI Explained - AWS,” Amazon Web Services, Inc. Accessed: Apr. 25, 2026. [Online]. Available: https://aws.amazon.com/what-is/*retrieval*-augmented-generation/

[39] J. Han and M. Kamber, *Data mining: concepts and techniques*, 2. ed., [Nachdr.]. in The Morgan Kaufmann series in data management systems. Amsterdam Heidelberg: Elsevier, Morgan Kaufmann, 20.

[40] M. DPatil and S. S. Sane, “Dimension Reduction: A Review,” *Int. J. Comput. Appl.*, vol. 92, no. 16, pp. 23–29, Apr. 2014, doi: 10.5120/16094-5390.

[41] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” Sep. 18, 2020, *arXiv*: arXiv:1802.03426. doi: 10.48550/arXiv.1802.03426.

[42] C. M. Weber, D. Ray, A. A. Valverde, J. A. Clark, and K. S. Sharma, “Gaussian mixture model clustering algorithms for the analysis of high-precision mass measurements,” *Nucl. Instrum. Methods Phys. Res. Sect. Accel. Spectrometers Detect. Assoc. Equip.*, vol. 1027, p. 166299, Mar. 2022, doi: 10.1016/j.nima.2021.166299.

[43] L. Jiao, T. Denœux, Z. Liu, and Q. Pan, “EGMM: An evidential version of the Gaussian mixture model for clustering,” *Appl. Soft Comput.*, vol. 129, p. 109619, Nov. 2022, doi: 10.1016/j.asoc.2022.109619.

[44] L. Ma *et al.*, “A Comprehensive Survey on Vector Database: Storage and *Retrieval* Technique, Challenge,” Mar. 24, 2026, *arXiv*: arXiv:2310.11703. doi: 10.48550/arXiv.2310.11703.

[45] Y. A. Malkov and D. A. Yashunin, “Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs,” *IEEE Trans Pattern Anal Mach Intell*, vol. 42, no. 4, pp. 824–836, Apr. 2020, doi: 10.1109/TPAMI.2018.2889473.

[46] S. Ockerman *et al.*, “Exploring Distributed Vector Databases Performance on HPC Platforms: A Study with Qdrant,” in *Proceedings of the SC ’25 Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*, in SC Workshops ’25. New York, NY,

USA: Association for Computing Machinery, Nov. 2025, pp. 575–581. doi: 10.1145/3731599.3767404.

[47] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P. Isabelle, E. Charniak, and D. Lin, Eds., Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. doi: 10.3115/1073083.1073135.

[48] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. Accessed: Apr. 28, 2026. [Online]. Available: https://aclanthology.org/W04-1013/

[49] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, Eds., Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72. Accessed: Apr. 28, 2026. [Online]. Available: https://aclanthology.org/W05-0909/

[50] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," presented at the Eighth International Conference on Learning Representations, Apr. 2020. Accessed: Apr. 28, 2026. [Online]. Available: https://iclr.cc/virtual_2020/poster_SkeHuCVFDr.html

[51] S. S. M. Wara, A. F. Adziima, M. Nasrudin, and A. R. Pratama, "Evaluasi Kinerja Uji Normalitas pada Ragam Distribusi dan Ukuran Sampel," *J. Difer.*, vol. 7, no. 2, pp. 172–183, Nov. 2025, doi: 10.35508/jd.v7i2.24042.

[52] E. H. Cui, Y. Li, and Z. Liu, "The Kolmogorov-Smirnov Statistic Revisited," Feb. 27, 2025, *arXiv*: arXiv:2503.11673. doi: 10.48550/arXiv.2503.11673.

[53] I. C. Anaene Oyeka and G. U. Ebuh, "Modified Wilcoxon Signed-Rank Test," *Open J. Stat.*, vol. 02, no. 02, pp. 172–176, 2012, doi: 10.4236/ojs.2012.22019.

[54] O. A. Adedokun and W. D. Burgess, "Analysis of Paired Dichotomous Data: A Gentle Introduction to the McNemar Test in SPSS," *J. Multidiscip. Eval.*, vol. 8, no. 17, pp. 125–131, Jan. 2012, doi: 10.56645/jmde.v8i17.336.

Halaman ini sengaja dikosongkan