UNIVERSITAS
GADJAH MADA
**Analisis Komparasi Arsitektur Retrieval-Augmented Generation (RAG) Berbasis Model Llama 3.1 Pada
Implementasi Sistem Chatbot Pembantu Pengambilan Keputusan Investasi**
Ryvalino Dhanu Ekaputra, Dr.Eng. Ir. Ganjar Alfian, S.T., M.Eng.
**Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/**

# DAFTAR PUSTAKA

[1] J. Chen *et al.*, "FinTextQA: A Dataset for Long-form Financial Question Answering," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 6025–6047. doi: 10.18653/v1/2024.acl-long.328.

[2] J. Liu and S. Fu, "Financial big data management and intelligence based on computer intelligent algorithm," *Sci. Rep.*, vol. 14, no. 1, p. 9395, Apr. 2024, doi: 10.1038/s41598-024-59244-8.

[3] Y. Ye, Y. Wang, and X. Yang, "Bank loan information and information asymmetry in the stock market: evidence from China," *Financ. Innov.*, vol. 8, no. 1, p. 62, Dec. 2022, doi: 10.1186/s40854-022-00367-0.

[4] L. Wang *et al.*, "A survey on large language model based autonomous agents," *Front. Comput. Sci.*, vol. 18, no. 6, p. 186345, Dec. 2024, doi: 10.1007/s11704-024-40231-1.

[5] Z. Ji *et al.*, "Survey of Hallucination in Natural Language Generation," *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–38, Dec. 2023, doi: 10.1145/3571730.

[6] L. Huang *et al.*, "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," *ACM Trans. Inf. Syst.*, vol. 43, no. 2, pp. 1–55, Mar. 2025, doi: 10.1145/3703155.

[7] J. Gallifant *et al.*, "Peer review of GPT-4 technical report and systems card," *PLOS Digit. Health*, vol. 3, no. 1, p. e0000417, Jan. 2024, doi: 10.1371/journal.pdig.0000417.

[8] W. Fan *et al.*, "A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Barcelona Spain: ACM, Aug. 2024, pp. 6491–6501. doi: 10.1145/3637528.3671470.

[9] P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks".

[10] B. Peng *et al.*, "Graph Retrieval-Augmented Generation: A Survey," Sep. 10, 2024, *arXiv*: arXiv:2408.08921. doi: 10.48550/arXiv.2408.08921.

[11] X. Wang *et al.*, "FINSAGE: A Multi-aspect RAG System for Financial Filings Question Answering," in *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, Seoul Republic of Korea: ACM, Nov. 2025, pp. 6144–6152. doi: 10.1145/3746252.3761587.

UNIVERSITAS
GADJAH MADA
**Analisis Komparasi Arsitektur Retrieval-Augmented Generation (RAG) Berbasis Model Llama 3.1 Pada
Implementasi Sistem Chatbot Pembantu Pengambilan Keputusan Investasi**
Ryvalino Dhanu Ekaputra, Dr.Eng. Ir. Ganjar Alfian, S.T., M.Eng.
**Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/**

[12] M. Mridul Provakar and E. K. Hashi, "Exploring the Effectiveness of Large Language Models in Financial Question Answering," in *Proceedings of the 3rd International Conference on Computing Advancements*, Dhaka Bangladesh: ACM, Oct. 2024, pp. 498–505. doi: 10.1145/3723178.3723244.

[13] P. Anderson, M. V. Janardhanan, J. He, W. Cheng, and C. Flanagan, "Greenback Bears and Fiscal Hawks: Finance is a Jungle and Text Embeddings Must Adapt," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Miami, Florida, US: Association for Computational Linguistics, 2024, pp. 362–370. doi: 10.18653/v1/2024.emnlp-industry.26.

[14] J. Wang, W. Ding, and X. Zhu, "Financial Analysis: Intelligent Financial Data Analysis System Based on LLM-RAG," Mar. 20, 2025, *Computer Science and Mathematics*. doi: 10.20944/preprints202503.1532.v1.

[15] I. Iaroshev, R. Pillai, L. Vaglietti, and T. Hanne, "Evaluating Retrieval-Augmented Generation Models for Financial Report Question and Answering," *Appl. Sci.*, vol. 14, no. 20, p. 9318, Oct. 2024, doi: 10.3390/app14209318.

[16] A. Mokashi, B. Puthuparambil, C. Daniel, and T. Hanne, "Analysis of Large Language Models for Company Annual Reports Based on Retrieval-Augmented Generation," *Information*, vol. 16, no. 9, p. 786, Sep. 2025, doi: 10.3390/info16090786.

[17] L. Ngartera, S. Nadarajah, and R. Koina, "Bayesian RAG: uncertainty-aware retrieval for reliable financial question answering," *Front. Artif. Intell.*, vol. 8, p. 1668172, Jan. 2026, doi: 10.3389/frai.2025.1668172.

[18] B. Sarmah, D. Mehta, B. Hall, R. Rao, S. Patel, and S. Pasquali, "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*, Brooklyn NY USA: ACM, Nov. 2024, pp. 608–616. doi: 10.1145/3677052.3698671.

[19] P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks".

[20] X. Zhu, "Intelligent decision support systems for improving financial forecasting and market trend analysis," *Expert Syst. Appl.*, vol. 297, p. 129462, Feb. 2026, doi: 10.1016/j.eswa.2025.129462.

[21] Z. Liu, K. Zhang, and H. Zhang, "A new era of financial services: How AI enhances investment efficiency," *Int. Stud. Econ.*, vol. 19, no. 4, pp. 578–588, Dec. 2024, doi: 10.1002/ise3.97.

UNIVERSITAS
GADJAH MADA
**Analisis Komparasi Arsitektur Retrieval-Augmented Generation (RAG) Berbasis Model Llama 3.1 Pada
Implementasi Sistem Chatbot Pembantu Pengambilan Keputusan Investasi**
Ryvalino Dhanu Ekaputra, Dr.Eng. Ir. Ganjar Alfian, S.T., M.Eng.
**Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/**

[22] K. Du, Y. Zhao, R. Mao, F. Xing, and E. Cambria, "Natural language processing in finance: A survey," *Inf. Fusion*, vol. 115, p. 102755, Mar. 2025, doi: 10.1016/j.inffus.2024.102755.

[23] S. Klampfl, M. Granitzer, K. Jack, and R. Kern, "Unsupervised document structure analysis of digital scientific articles," *Int. J. Digit. Libr.*, vol. 14, no. 3–4, pp. 83–99, Aug. 2014, doi: 10.1007/s00799-014-0115-1.

[24] D. Balsiger, H.-R. Dimmler, S. Egger-Horstmann, and T. Hanne, "Assessing Large Language Models Used for Extracting Table Information from Annual Financial Reports," *Computers*, vol. 13, no. 10, p. 257, Oct. 2024, doi: 10.3390/computers13100257.

[25] Y. Li, S. Wang, H. Ding, and H. Chen, "Large Language Models in Finance: A Survey," in *4th ACM International Conference on AI in Finance*, Brooklyn NY USA: ACM, Nov. 2023, pp. 374–382. doi: 10.1145/3604237.3626869.

[26] T. E. Oliphant, "Python for Scientific Computing," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 10–20, 2007, doi: 10.1109/MCSE.2007.58.

[27] X. Guo, H. Okamura, and T. Dohi, "Improving Test Suite Generation Quality Through Machine Learning-Driven Boundary Value Analysis," 2024, *SSRN*. doi: 10.2139/ssrn.5027291.

[28] M. Shao, A. Basit, R. Karri, and M. Shafique, "Survey of Different Large Language Model Architectures: Trends, Benchmarks, and Challenges," *IEEE Access*, vol. 12, pp. 188664–188706, 2024, doi: 10.1109/ACCESS.2024.3482107.

[29] M. Raza, Z. Jahangir, M. B. Riaz, M. J. Saeed, and M. A. Sattar, "Industrial applications of large language models," *Sci. Rep.*, vol. 15, no. 1, p. 13755, Apr. 2025, doi: 10.1038/s41598-025-98483-1.

[30] T. B. Brown *et al.*, "Language Models are Few-Shot Learners".

[31] A. Grattafiori *et al.*, "The Llama 3 Herd of Models," Nov. 23, 2024, *arXiv*: arXiv:2407.21783. doi: 10.48550/arXiv.2407.21783.

[32] L. Ajallouda, M. H. Saissi, and A. Zellou, "Embedding Models: A Comprehensive Review with Task-Oriented Assessment," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 10, 2025.

[33] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation," in *Findings of the Association for Computational Linguistics ACL 2024*, Bangkok, Thailand and virtual meeting:

UNIVERSITAS
GADJAH MADA
**Analisis Komparasi Arsitektur Retrieval-Augmented Generation (RAG) Berbasis Model Llama 3.1 Pada
Implementasi Sistem Chatbot Pembantu Pengambilan Keputusan Investasi**
Ryvalino Dhanu Ekaputra, Dr.Eng. Ir. Ganjar Alfian, S.T., M.Eng.
**Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/**

Association for Computational Linguistics, 2024, pp. 2318–2335. doi: 10.18653/v1/2024.findings-acl.137.

[34] Octen Team, "Octen Series: Optimizing Embedding Models to #1 on RTEB Leaderboard." [Online]. Available: https://octen-team.github.io/octen_blog/posts/octen-rteb-first-place/

[35] M. Klesel and H. F. Wittmann, "Retrieval-Augmented Generation (RAG)," *Bus. Inf. Syst. Eng.*, vol. 67, no. 4, pp. 551–561, Aug. 2025, doi: 10.1007/s12599-025-00945-3.

[36] X. Wang *et al.*, "Searching for Best Practices in Retrieval-Augmented Generation," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 17716–17736. doi: 10.18653/v1/2024.emnlp-main.981.

[37] R. Qu, R. Tu, and F. Bao, "Is Semantic Chunking Worth the Computational Cost?," Oct. 16, 2024, *arXiv*: arXiv:2410.13070. doi: 10.48550/arXiv.2410.13070.

[38] 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.

[39] A. Hogan *et al.*, "Knowledge Graphs," *ACM Comput. Surv.*, vol. 54, no. 4, pp. 1–37, May 2022, doi: 10.1145/3447772.

[40] J. Guia, V. Gonçalves Soares, and J. Bernardino, "Graph Databases: Neo4j Analysis:," in *Proceedings of the 19th International Conference on Enterprise Information Systems*, Porto, Portugal: SCITEPRESS - Science and Technology Publications, 2017, pp. 351–356. doi: 10.5220/0006356003510356.

[41] 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)*, Hong Kong, China: Association for Computational Linguistics, 2019, pp. 3980–3990. doi: 10.18653/v1/D19-1410.

[42] Y.-Y. Liu *et al.*, "A comprehensive taxonomy of prompt engineering techniques for large language models," *Front. Comput. Sci.*, vol. 20, no. 3, p. 2003601, Mar. 2026, doi: 10.1007/s11704-025-50058-z.

[43] N. Esmi, A. Shahbahrami, Y. Nabati, B. Rezaei, G. Gaydadjiev, and P. De Jonge, "Stress detection through prompt engineering with a general-purpose

UNIVERSITAS
GADJAH MADA
**Analisis Komparasi Arsitektur Retrieval-Augmented Generation (RAG) Berbasis Model Llama 3.1 Pada
Implementasi Sistem Chatbot Pembantu Pengambilan Keputusan Investasi**
Ryvalino Dhanu Ekaputra, Dr.Eng. Ir. Ganjar Alfian, S.T., M.Eng.
**Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/**

LLM," *Acta Psychol. (Amst.)*, vol. 260, p. 105462, Oct. 2025, doi: 10.1016/j.actpsy.2025.105462.

[44] Vasilios Mavroudis, "LangChain v0.3," Nov. 2024, [Online]. Available: https://hal.science/hal-04817573v1

[45] P. Islam, A. Kannappan, D. Kiela, R. Qian, N. Scherrer, and B. Vidgen, "FinanceBench: A New Benchmark for Financial Question Answering," Nov. 20, 2023, *arXiv*: arXiv:2311.11944. doi: 10.48550/arXiv.2311.11944.

[46] M. K. Yusof, M. Man, W. M. A. F. Wan Hamzah, S. Safei, M. Tasani Haryani, and N. A. A. A. Shah, "A Secure Model Based on Symmetric Encryption for REST API in Data Integration," in *2023 27th International Computer Science and Engineering Conference (ICSEC)*, Samui Island, Thailand: IEEE, Sep. 2023, pp. 402–406. doi: 10.1109/ICSEC59635.2023.10329720.

[47] J. Chen, "Model Algorithm Research based on Python Fast API," *Front. Sci. Eng.*, vol. 3, no. 9, pp. 7–10, Sep. 2023, doi: 10.54691/fse.v3i9.5591.

[48] A. P. Putra and D. P. S. Putri, "Scientific Paper Recommendation System: Application of Sentence Transformers and Cosine Similarity Using arXiv Data," vol. 9, no. 4.

[49] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTSCORE: EVALUATING TEXT GENERATION WITH," 2020.

[50] S. Lee *et al.*, "A Survey on Evaluation Metrics for Machine Translation," *Mathematics*, vol. 11, no. 4, p. 1006, Feb. 2023, doi: 10.3390/math11041006.

[51] N. Kimseng, D. A. Kurnia, I. Vuthy, R. W. Arifin, and D. Setiyadi, "UI/UX Development Using Figma based on Inclusive Design," *JINAV J. Inf. Vis.*, vol. 4, no. 2, pp. 227–234, Jan. 2024, doi: 10.35877/454RI.jinav2257.

[52] L. Baresi, G. Quattrocchi, and N. Rasi, "A qualitative and quantitative analysis of container engines," *J. Syst. Softw.*, vol. 210, p. 111965, Apr. 2024, doi: 10.1016/j.jss.2024.111965.

[53] "Black box software testing techniques :a literature review," *Passer J. Basic Appl. Sci.*, vol. 7, no. 2, Aug. 2025, doi: 10.24271/psr.2025.508216.1963.

[54] E. C. D. Santos, P. Vilain, and D. H. Longo, "A systematic literature review to support the selection of user acceptance testing techniques," in *Proceedings of the 40th International Conference on Software Engineering: Companion Proceeedings*, Gothenburg Sweden: ACM, May 2018, pp. 418–419. doi: 10.1145/3183440.3195036.

UNIVERSITAS
GADJAH MADA
**Analisis Komparasi Arsitektur Retrieval-Augmented Generation (RAG) Berbasis Model Llama 3.1 Pada
Implementasi Sistem Chatbot Pembantu Pengambilan Keputusan Investasi**
Ryvalino Dhanu Ekaputra, Dr.Eng. Ir. Ganjar Alfian, S.T., M.Eng.
**Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/**

[55] Aliyah Aliyah, Nahrun Hartono, and Asrul Azhari Muin, “Penggunaan User Acceptance Testing (UAT) Pada Pengujian Sistem Informasi Pengelolaan Keuangan Dan Inventaris Barang,” *Switch J. Sains Dan Teknol. Inf.*, vol. 3, no. 1, pp. 84–100, Dec. 2024, doi: 10.62951/switch.v3i1.330.

[56] H. Patil, V. Mahandule, and A. Gunjal, “Python in the Evolution of AI: A Comparative Study of Emerging Technologies,” *SSRN Electron. J.*, 2025, doi: 10.2139/ssrn.5075929.

[57] A. Tsagaris, “Mechatronic system usability evaluation,” *J. Technol. Sci. Educ.*, vol. 13, no. 1, p. 65, Jan. 2023, doi: 10.3926/jotse.1349.

[58] Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, F. Huang, and J. Zhou, "Qwen3 Embedding: Advancing text embedding and reranking through foundation models," arXiv:2506.05176, Jun. 2025.

[59] C. Kiss, M. Nagy, and P. Szilágyi, "Max–Min semantic chunking of documents for RAG application," *Discover Computing*, vol. 28, p. 117, 2025. doi: 10.1007/s10791-025-09638-7.