UNIVERSITAS
GADJAH MADA
**Implementasi dan Analisis Kinerja Layanan Chatbot Berbasis Large Language Model dengan Arsitektur
Retrieval Augmented Generation (Studi Kasus Layanan UGM Anjem)**
Anugrah Dwiki Ardana, Ir. Yuris Mulya Saputra, S.T., M.Sc., Ph.D., IPM.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

DAFTAR PUSTAKA

[1] Universitas Gadjah Mada, "About UGM," [Online]. Available: https://ugm.ac.id/en/about

[2] Universitas Gadjah Mada, "Transportasi Mahasiswa," [Online]. Available: https://ugm.ac.id/id/3550-transportasi-mahasiswa/

[3] T. Andriyani, "UGM pinjamkan sepeda bagi mahasiswa untuk mobilitas ke kampus," Universitas Gadjah Mada, 7 Agustus 2024. [Online]. Available: https://ugm.ac.id/id/berita/ugm-pinjamkan-sepeda-bagi-mahasiswa-untuk-mobilitas-ke-kampus/

[4] Y. Gao *et al.*, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint* arXiv:2312.10997, 2023. [Online]. Available: https://arxiv.org/abs/2312.10997

[5] S. U. Singh and A. S. Namin, "A survey on *chatbot*s and large language models: Testing and evaluation techniques," *Nat. Lang. Process. J.*, vol. 10, p. 100128, Mar. 2025, doi: 10.1016/j.nlp.2025.100128.

[6] B. Beyer, C. Jones, J. Petoff, and N. R. Murphy, Eds., *Site Reliability Engineering: How Google Runs Production Systems*. Sebastopol, CA: O'Reilly Media, 2016. [Online]. Available: https://sre.google/sre-book/table-of-contents/

[7] M. D. Elradi, "Prometheus & Grafana: A metrics-focused monitoring stack," *J. Comput. Allied Intell. (JCAI)*, vol. 1, 2025, doi: 10.69996/jcai.2025015.

[8] Prometheus Authors, "Metric types," [Online]. Available: https://prometheus.io/docs/concepts/metric_types/

[9] Prometheus Authors, "Histograms and summaries," [Online]. Available: https://prometheus.io/docs/practices/histograms/

[10] Prometheus Authors, "Querying basics," [Online]. Available: https://prometheus.io/docs/prometheus/latest/querying/basics/

[11] Grafana Labs, "About metrics and telemetry," [Online]. Available: https://grafana.com/docs/grafana/latest/explore/simplified-exploration/metrics/about-metrics

UNIVERSITAS GADJAH MADA
Implementasi dan Analisis Kinerja Layanan Chatbot Berbasis Large Language Model dengan Arsitektur Retrieval Augmented Generation (Studi Kasus Layanan UGM Anjem)
Anugrah Dwiki Ardana, Ir. Yuris Mulya Saputra, S.T., M.Sc., Ph.D., IPM.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

[12] S. Kusumawardhani, M. F. Rafi, and E. Marpanaji, "Experimental study of high availability cloud learning management system and monitoring system based on Grafana, Prometheus, and Telegram," *J. Adv. Res. Des.*, vol. 136, no. 1, pp. 66–78, 2024, doi: 10.37934/ard.136.1.6678.

[13] M. A. Hasbi, R. Imanda, and M. F. Fauzan, "Implementasi *chatbot* berbasis large language model untuk pencarian skripsi mahasiswa terintegrasi dengan WhatsApp," *Arcitech: J. Comput. Sci. Artif. Intell.*, 2025. [Online]. Available: https://journal.iaincurup.ac.id/index.php/arcitech/article/view/13974

[14] M. Khasanova Zafar Kizi and Y. Suh, "Design and performance evaluation of LLM-based RAG *pipeline*s for *chatbot* services in international student admissions," *Electronics*, vol. 14, no. 15, p. 3095, Aug. 2025, doi: 10.3390/electronics14153095.

[15] M. L. Husain, Y. Wibisono, and A. Anisyah, "Development of an academic services *chatbot* based on retrieval-augmented generation (RAG)," *Brilliance: Res. Artif. Intell.*, vol. 5, no. 2, pp. 727–735, 2025.

[16] A. Brown, M. Roman, and B. Devereux, "A systematic literature review of retrieval-augmented generation: Techniques, metrics, and challenges," *Big Data Cogn. Comput.*, vol. 9, no. 12, p. 320, Dec. 2025, doi: 10.3390/bdcc9120320.

[17] R. Yang, M. Fu, C. Tantithamthavorn, C. Arora, L. Vandenhurk, and J. Chua, "RAGVA: Engineering retrieval augmented generation-based virtual assistants in practice," *J. Syst. Softw.*, vol. 226, p. 112436, 2025, doi: 10.1016/j.jss.2025.112436.

[18] G. Bubaš, A. Čižmešija, and A. Kovačić, "Development of an assessment scale for measurement of usability and user experience characteristics of Bing Chat conversational AI," *Future Internet*, vol. 16, no. 1, p. 4, Dec. 2023, doi: 10.3390/fi16010004.

[19] F. Saory, J. H. Jaman, and Carudin, "Sistem monitoring server menggunakan Prometheus dan Grafana dengan integrasi Slack di PT. Atlas Lintas Indonesia," *J. Inform. dan Tek. Elektro Terap. (JITET)*, vol. 13, no. 3S1, 2025, doi: 10.23960/jitet.v13i3S1.8024.

[20] L. S. T. Nguyen and T. T. Quan, "URAG: Implementing a unified hybrid RAG for precise answers in university admission *chatbot*s—a case study at HCMUT," in

UNIVERSITAS
GADJAH MADA
**Implementasi dan Analisis Kinerja Layanan Chatbot Berbasis Large Language Model dengan Arsitektur
Retrieval Augmented Generation (Studi Kasus Layanan UGM Anjem)**
Anugrah Dwiki Ardana, Ir. Yuris Mulya Saputra, S.T., M.Sc., Ph.D., IPM.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Commun. Comput. Inf. Sci. (CCIS), vol. 2351, Springer, 2025, doi: 10.1007/978-981-96-4285-4_7.

[21] C. W. Okonkwo and A. Ade-Ibijola, "*Chatbot*s applications in education: A systematic review," *Comput. Educ. Artif. Intell.*, vol. 2, p. 100033, 2021, doi: 10.1016/j.caeai.2021.100033.

[22] E. Adamopoulou and L. Moussiades, "An overview of *chatbot* technology," in *Proc. IFIP Int. Conf. Artif. Intell. Appl. Innov. (AIAI)*, 2020, pp. 373–383, doi: 10.1007/978-3-030-49186-4_31.

[23] T. Brown *et al.*, "Language models are few-shot learners," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 1877–1901, 2020.

[24] OpenAI, "GPT-4o mini: Advancing cost-efficient intelligence," OpenAI Blog, Jul. 2024. [Online]. Available: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[25] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 9459–9474, 2020.

[26] A. Asai *et al.*, "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," in *Proc. ICLR*, 2024. [Online]. Available: https://arxiv.org/abs/2310.11511

[27] Y. Wang et al., "Science Checker Reloaded: A bidirectional paradigm for transparency and logical reasoning," arXiv:2410.02819, 2024. [Online]. Available: https://arxiv.org/abs/2410.02819

[28] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 26, 2013.

[29] OpenAI, "New embedding models and API updates," OpenAI Blog, Jan. 2024. [Online]. Available: https://openai.com/index/new-embedding-models-and-api-updates/

[30] D. Merkel, "Docker: Lightweight Linux containers for consistent development and deployment," *Linux J.*, vol. 2014, no. 239, p. 2, 2014.

UNIVERSITAS
GADJAH MADA
**Implementasi dan Analisis Kinerja Layanan Chatbot Berbasis Large Language Model dengan Arsitektur
Retrieval Augmented Generation (Studi Kasus Layanan UGM Anjem)**
Anugrah Dwiki Ardana, Ir. Yuris Mulya Saputra, S.T., M.Sc., Ph.D., IPM.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

[31] R. Likert, “A technique for the measurement of attitudes,” *Arch. Psychol.*, vol. 22, no. 140, pp. 1–55, 1932.

[32] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, “RAGAS: Automated evaluation of retrieval augmented generation,” *arXiv:2309.15217*, 2023. [Online]. Available: https://arxiv.org/abs/2309.15217

[33] J. Scheuner and P. Leitner, “Function-as-a-service performance evaluation: A multivocal literature review,” *J. Syst. Softw.*, vol. 170, p. 110708, 2020, doi: 10.1016/j.jss.2020.110708.

[34] Locust Contributors, "Locust Documentation." [Online]. Available: https://docs.locust.io/en/stable/

[35] S. Ramírez, “FastAPI,” 2019. [Online]. Available: https://fastapi.tiangolo.com