DAFTAR PUSTAKA

Agrawal, A., Agarwal, A., Kedia, N., Mohan, J., Kundu, S., Kwatra, N., Ramjee, R., & Tumanov, A. (2024). *Etalon: Holistic Performance Evaluation Framework for LLM Inference Systems*. http://arxiv.org/abs/2407.07000

Agustini, K., Sindu, I. G. P., & Kusuma, K. A. (2019). The effectiveness of content based on dynamic intellectual learning with visual modality in vocational school. *Jurnal Pendidikan Vokasi*, *9*(1), 11–20. https://doi.org/10.21831/jpv.v9i1.21629

Akheel, S. A. (2025). Guardrails for Large Language Models: A Review of Techniques and Challenges. *Journal of Artificial Intelligence, Machine Learning and Data Science*, *3*(1), 2504–2512. https://doi.org/10.51219/jaimld/syed-arham-akheel/536

Al Thaher, J., Radas, J., & Al, J. (2025). *Exploring Retrieval-Augmented Generation for Large Language Models: Enhancing University IT Support with a RAG-based Chatbot*.

Alibaba Cloud Community. (2022, January 10). *Introduction to the Python Flask framework. Alibaba Cloud*. Alibaba Cloud.

Anassai, B. R., & Josaphat, P. (2024). *Pembangunan Chatbot Sistem Informasi KBLI dan KBJI Berbasis LLM (Development of LLM-Based KBLI and KBJI Information System Chatbot)*.

Anglada. (2007). *Gambar 1: Tahapan ADDIE Model*. ResearchGate.

Arthana, I. K. R., Dewi, N. P. N. P., Saskara, G. A. J., Pradnyana, I. M. A., & Indrayani, L. (2026). Real-time intelligent virtual assistant based on retrieval augmented generation. *IAES International Journal of Artificial Intelligence (IJ-AI)*, *15*(1), 237. https://doi.org/10.11591/ijai.v15.i1.pp237-246

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). *SELF*-RAG*: LEARNING TO RETRIEVE, GENERATE, AND CRITIQUE THROUGH SELF-REFLECTION*. https://selfrag.github.io/.

Brooke, J. (1996). SUS: A “quick and dirty” usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability evaluation in industry* (pp. 189–194). Taylor & Francis.

Cheng, X., Wang, X., Zhang, X., Ge, T., Chen, S.-Q., Wei, F., Zhang, H., & Zhao, D. (2024). *xRAG: Extreme Context Compression for Retrieval-augmented Generation with One Token*. http://arxiv.org/abs/2405.13792

Dasmen, R. N., Fatoni, F., Wijaya, A., Tujni, B., & Nabila, S. (2021). Pelatihan uji kegunaan website menggunakan System Usability Scale (SUS). *ABSYARA: Jurnal Pengabdian Pada Masyarakat*, *2*(2), 146–158.

Elkiran, H., & Rasheed, J. (2025). EvaRAG: Evaluating Advanced RAG Techniques With Indexing and Distance Metrics. *IEEE Access*, *13*, 215724–215747. https://doi.org/10.1109/ACCESS.2025.3646665

Elysia, S., & Herianto. (2024). Chatbot Berbasis Retrieval Augmented Generation (RAG) untuk Peningkatan Layanan Informasi Sekolah. *Journal TIFDA (Technology Information and Data Analytic)*, *1*(2), 52–58. https://doi.org/10.70491/tifda.v1i2.52

Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2025). *Ragas: Automated Evaluation of Retrieval Augmented Generation*. http://arxiv.org/abs/2309.15217

Fitria Hidayat, & Muhamad Nizar. (2021). MODEL ADDIE (ANALYSIS, DESIGN, DEVELOPMENT, IMPLEMENTATION AND EVALUATION) DALAM PEMBELAJARAN PENDIDIKAN AGAMA ISLAM ADDIE (ANALYSIS, DESIGN, DEVELOPMENT, IMPLEMENTATION AND EVALUATION) MODEL IN ISLAMIC EDUCATION LEARNING. *JIPAI; Jurnal Inovasi Pendidikan Agama Islam*, *1(1)_*.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). *Retrieval-Augmented Generation for Large Language Models: A Survey*. http://arxiv.org/abs/2312.10997

Google. (2025). *Batas tarif | Google AI for Developers*. Google. https://ai.google.dev/gemini-api/docs/rate-limits?hl=id#free-tier

Google AI Blog. (2023). *Introducing Gemini: A new paradigm in Large Language Models*. Google. HTTPs://AI.googleblog.com

Google DeepMind. (2023). *Gemini: The next generation AI model*. Google. HTTPs://deepmind.com/research/Gemini

Hadi, M. U., Tashi, Q. Al, Qureshi, R., Shah, A., Muneer, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Hassan, S. Z., Shoman, M., Wu, J., Mirjalili, S., & Shah, M. (2025). *Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects*. https://doi.org/10.36227/techrxiv.23589741.v8

Handayani, D. S. , Irwansyah, I., Sondang, S. , & Kaunang, R. ,. (2024). Manfaat dan Potensi Masalah Penggunaan Kecerdasan Buatan (AI) dalam Komunikasi Publik. *Co-Value: Jurnal Ekonomi, Koperasi & Kewirausahaan*, *14*(12), 1–20.

Hidayat, L. R., Pasek, G., Wijaya, S., & Dwiyansaputra, R. (n.d.). *Optimalisasi Layanan Sistem Informasi Mahasiswa dengan Integrasi Telegram : Chatbot Retrieval-Augmented-Generation berbasis Large Language Model (Optimization of Student Information System Services with*

Telegram Integration : Chatbot Retrieval-Augmented Generation based on Large Language Model). Retrieved http://jtika.if.unram.ac.id/index.php/JTIKA/

Ichsanudin, I., Pratama, R., & Sisephaputra, B. (2024). Pengembangan Sistem Helpdesk Menggunakan Chatbot Dengan Metode Retrieval-Augmented Generation (RAG). *Journal of Informatics and Computer Science*, *06*.

Izacard, G., & Grave, E. (2021). *Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering*.

Jeong, C. (n.d.). *A Study on the Implementation of Generative AI Services Using an Enterprise Data-Based LLM Application Architecture*.

Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., & Neubig, G. (2023). *Active Retrieval Augmented Generation*. http://arxiv.org/abs/2305.06983

Kristiyanto, E. N. (2016). Urgensi_Keterbukaan_Informasi_dalam_Peny (1). *Jurnal Penelitian Hukum De Jure*, *16(2)*, 231–244.

Kurniawan, D., & Triloka, J. (2025). *Penerapan Teknologi Langchain dan LLM pada Sistem Question Answering Berbasis Chatbot Telegram: Literature Review*.

Lehto Bachelor, T., & June, T. (2024). *Developing LLM-powered Applications Using Modern Frameworks*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2025). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. https://github.com/huggingface/transformers/blob/master/

Li, H., Huang, J., Ji, M., Yang, Y., & An, R. (2025). Use of Retrieval-Augmented Large Language Model for COVID-19 Fact-Checking: Development and Usability Study. *Journal of Medical Internet Research*, *27*(1). https://doi.org/10.2196/66098

Li, Z., Li, C., Zhang, M., Mei, Q., & Bendersky, M. (2024). *Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach*.

Lubis, A. T. U. BR., Harahap, N. S., Agustian, S., Irsyad, M., & Afrianty, I. (2024). Question Answering System pada Chatbot Telegram Menggunakan Large Language Models (LLM) dan Langchain (Studi Kasus UU Kesehatan). *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(3), 955–964. https://doi.org/10.57152/malcom.v4i3.1378

Madaan, A., & Yazdanbakhsh, A. (2022). *Text and Patterns: For Effective Chain of Thought, It Takes Two to Tango*. http://arxiv.org/abs/2209.07686

Maiasnita, Y. , & Darwis, Y. (2024). Pentingnya Sistem Pelayanan Informasi Bagi Pemerintah dan Publik (Studi Kasus Aplikasi Kaba AIa Perumda Padang). . *Jurnal Ilmu Komunikasi Dan Sosial Politik*, *1*(3), 394–401.

Mell, P. M., & Grance, T. (2011). *The NIST definition of cloud computing*. https://doi.org/10.6028/NIST.SP.800-145

Muhajir, M. D. A. , Koeshardianto, M., & Prastiti, N. ,. (2025). Implementasi Chatbot Menggunakan Framework Langchain Berbasis LLM GPT (Studi Kasus: Panduan Akademik Universitas Trunojoyo). *JATI (Jurnal Mahasiswa Teknik Informatika)*, *9*(2).

Mulyawan, M., Dana, R. D., Bahtiar, A., & Ali, I. (2024). Optimalisasi Layanan Kesehatan di Puskesmas Melalui Pengembangan Chatbot Berbasis Web Menggunakan Flowise AI. *JTIM : Jurnal Teknologi Informasi Dan Multimedia*, *6*(3), 376–391. https://doi.org/10.35746/jtim.v6i3.617

Neupane, S., Hossain, E., Keith, J., Tripathi, H., Ghiasi, F., Golilarz, N. A., Amirlatifi, A., Mittal, S., & Rahimi, S. (2024). *From Questions to Insightful Answers: Building an Informed Chatbot for University Resources*. http://arxiv.org/abs/2405.08120

Nur'aini, I. (2024). SISTEM CHATBOT SEBAGAI LAYANAN INFORMASI KESEHATAN MENTAL PADA REMAJA MENGGUNAKAN METODE LARGE LANGUAGE MODEL (LLM). *(Doctoral Dissertation, Universitas Islam Sultan Agung Semarang).*

Öztürk, E., & Mesut, A. (2024). PERFORMANCE ANALYSIS OF CHROMA, QDRANT, AND FAISS DATABASES. *UNITECH – SELECTED PAPERS*. https://doi.org/10.70456/tbrn3643

Pan, G., Chodnekar, V., Roy, A., & Wang, H. (2025). *A Cost-Benefit Analysis of On-Premise Large Language Model Deployment: Breaking Even with Commercial LLM Services*. http://arxiv.org/abs/2509.18101

Pujiono, I., Agtyaputra, I. M., & Ruldeviyani, Y. (2024). IMPLEMENTING RETRIEVAL-AUGMENTED GENERATION AND VECTOR DATABASES FOR CHATBOTS IN PUBLIC SERVICES AGENCIES CONTEXT. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *10*(1), 216–223. https://doi.org/10.33480/jitk.v10i1.5572

Richardson, L. , & Ruby, S. (2007). *RESTful Web Services: Web Services for the Real World.* O'Reilly Media.

Rizal, S. (2025). Meningkatkan Minat Baca Siswa melalui Program Reading Classroom di SDN 2 Mamben Daya. *AS-SABIQUN*, *7*(2), 265–277. https://doi.org/10.36088/assabiqun.v7i2.5632

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2025). *A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications*. http://arxiv.org/abs/2402.07927

Setyawan, A. , Sugiartawan, I. P., & Prasetiyo, B. (2023). Pemanfaatan layanan chatbot sebagai media informasi pada Fakultas Ilmu Komputer Universitas Brawijaya. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *7*(3), 1234–1242.

Sheikhalishahi, S., Haddadi, A., Sadeghipour, S., Rafiei, F., & Soltani, H. (2026). Comparative performance of ChatGPT-4o, ChatGPT-5, and gemini 2.5 flash on Persian internal medicine subspecialty board exams. *Scientific Reports*, *16*(1). https://doi.org/10.1038/s41598-025-31251-3

Steybe, D., Poxleitner, P., Aljohani, S., Herlofson, B. B., Nicolatou-Galitis, O., Patel, V., Fedele, S., Kwon, T. G., Fusco, V., Pichardo, S. E. C., Obermeier, K. T., Otto, S., Rau, A., & Russe, M. F. (2025). Evaluation of a context-aware chatbot using retrieval-augmented generation for answering clinical questions on medication-related osteonecrosis of the jaw. *Journal of Cranio-Maxillofacial Surgery*, *53*(4), 355–360. https://doi.org/10.1016/j.jcms.2024.12.009

Sugiyarto, M. F., Yasa, R. N., Girinoto, G., Setiawan, H., & Wicaksono, H. R. (2025). Spaticrypt : Platform Edukasi Kriptografi Berbasis Web dengan Konsep Gamifikasi Capture-the-Flag dan Integrasi Chatbot Kecerdasan Buatan sebagai Asisten Virtual. *Info Kripto*, *19*(1), 39–47. https://doi.org/10.56706/ik.v19i1.120

Sugiyono. (2014). *Metode Penelitian Kuantitatif, Kualitatif dan R & D*. Alfabeta.

Swacha, J., & Gracel, M. (2025). Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications. In *Applied Sciences (Switzerland)* (Vol. 15, Number 8). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/app15084234

Taipalus, T. (2024). Vector database management systems: Fundamental concepts, use-cases, and current challenges. *Cognitive Systems Research*, *85*. https://doi.org/10.1016/j.cogsys.2024.101216

Tanyildiz, D., Ayvaz, S., & Amasyali, M. F. (2024). Enhancing Retrieval-Augmented Generation Accuracy with Dynamic Chunking and Optimized Vector Search. *Orclever Proceedings of Research and Development*, *5*(1), 215–225. https://doi.org/10.56038/oprd.v5i1.516

Tribber, Y. , Asfi, M., & Kusnadi. (2024). Implementasi Retrieval Augmented Generation untuk Layanan Informasi Kampus dengan Chatbot Berbasis AI. . *Prosiding Seminar Nasional Sistem Informasi Dan Teknologi (SISFOTEK) Ke-8*, 594–600.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need*. http://arxiv.org/abs/1706.03762

Villena, F., Bravo-Marquez, F., & Dunstan, J. (2025). NLP modeling recommendations for restricted data availability in clinical settings. *BMC*

Medical Informatics and Decision Making, *25*(1). https://doi.org/10.1186/s12911-025-02948-2

Wijaya, S. M. , Prawita, F. N., M. E. Q. A., & Zahra. (2025). MicroLingo: Chatbot WhatsApp Berbasis AI untuk Pelatihan Bahasa Inggris dengan Pendekatan Microlearning untuk Pelaku UMKM Generasi Z di Indonesia. *E-Proceeding of Applied Science*, *11*(1), 72–76.

Yusuf, M. , Raihan, M. R., & Anam, A. ,. (2025). Sistem Pakar Penentuan Komposisi Skincare Berdasarkan Masalah Kulit Wajah Menggunakan Natural Language Processing (NLP). *JIRE: Jurnal Informatika & Rekayasa Elektronika* , *8*(1), 22–30.

Yusuf, M., Ganis, D., Murpri, S., & Ernas, K. A. (n.d.). *Sistem Pakar Menggunakan Metode Natural Language Processing Untuk Mendiagnosa Penyakit Tanaman Cabai Merah Keriting*.

Zhang, Y., Liu, S., & Wang, J. (2024). *Are There Fundamental Limitations in Supporting Vector Data Management in Relational Databases? A Case Study of PostgreSQL*. https://github.com/YunanZzz/VecDB-ICDE24.

Zheng, C., Liu, Z., Xie, E., Li, Z., & Li, Y. (2024). *Progressive-Hint Prompting Improves Reasoning in Large Language Models*. http://arxiv.org/abs/2304.09797