

DAFTAR PUSTAKA

Grattafiori, A. *dkk.* (2024) “The Llama 3 Herd of Models.” arXiv. Tersedia pada: <https://doi.org/10.48550/arXiv.2407.21783>.

Lewis, P. *dkk.* (2021) “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” arXiv. Tersedia pada: <https://doi.org/10.48550/arXiv.2005.11401>.

Vaswani, A. *dkk.* (2023) “Attention Is All You Need.” arXiv. Tersedia pada: <https://doi.org/10.48550/arXiv.1706.03762>.

Asai, A. et al. (2023) “Self-RAG: Learning to Retrieve, Generate, dan Critique through Self-Reflection.” arXiv. Available at: <https://doi.org/10.48550/arXiv.2310.11511>.

Bornmann, L. dan Mutz, R. (2015) “Growth rates of modern science: A bibliometric analysis based on the number of publications dan cited references,” *Journal of the Association for Information Science dan Technology*, 66(11), pp. 2215–2222. Available at: <https://doi.org/10.1002/asi.23329>.

Brown, T.B. et al. (2020) “Language Models are Few-Shot Learners.” arXiv. Available at: <https://doi.org/10.48550/arXiv.2005.14165>

Devlin, J. et al. (2019) “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” arXiv. Available at: <https://doi.org/10.48550/arXiv.1810.04805>.

Dosovitskiy, A. et al. (2021) “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” arXiv. Available at: <https://doi.org/10.48550/arXiv.2010.11929>.

Eman, E.T., Fatyanosa, T.N. dan Aji, A.F. (2026) “Analisis Perbandingan Metode *Chunking* dalam *Chatbot* Berbasis *Retrieval-Augmented Generation* Rekomendasi Terapi Nutrisi Medis Pasien.” *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 10(01).

Es, S. et al. (2023) “RAGAS: Automated Evaluation of *Retrieval Augmented Generation*.” arXiv. Available at: <https://doi.org/10.48550/arXiv.2309.15217>

Gao, Y. et al. (2024) “*Retrieval-Augmented Generation* for Large Language Models: A Survey.” arXiv. Available at: <https://doi.org/10.48550/arXiv.2312.10997>.

Grattafiori, A. et al. (2024) “The Llama 3 Herd of Models.” arXiv. Available at: <https://doi.org/10.48550/arXiv.2407.21783>.

Guntoro, G., Loneli Costaner dan Lisnawita, L. (2020) “Aplikasi *Chatbot* untuk Layanan Informasi dan Akademik Kampus Berbasis Artificial Intelligence Markup Language (AIML),” *Digital Zone: Jurnal Teknologi Informasi dan Komunikasi*,

11(2), pp. 291–300. Available at: <https://doi.org/10.31849/digitalzone.v11i2.5049>.

H.Gomaa, W. dan A. Fahmy, A. (2013) “A Survey of Text Similarity Approaches,” *International Journal of Computer Applications*, 68(13), pp. 13–18. Available at: <https://doi.org/10.5120/11638-7118>.

Johnson, J., Douze, M. dan Jégou, H. (2017) “Billion-scale similarity search with GPUs.” *arXiv*. Available at: <https://doi.org/10.48550/arXiv.1702.08734>.

Lewis, P. et al. (2021a) “*Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.*” *arXiv*. Available at: <https://doi.org/10.48550/arXiv.2005.11401>.

Lewis, P. et al. (2021b) “*Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.*” *arXiv*. Available at: <https://doi.org/10.48550/arXiv.2005.11401>.

Lewis, P. et al. (2021c) “*Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.*” *arXiv*. Available at: <https://doi.org/10.48550/arXiv.2005.11401>.

Liu, F., Kang, Z. dan Han, X. (2024) “A Case Study with Locally Deployed Ollama Models.” *arXiv*. Available at: <https://doi.org/10.48550/arXiv.2408.05933>

Liu, S., McCoy, A.B. dan Wright, A. (2025) “Improving large language model applications in biomedicine with *retrieval-augmented generation*: a systematic review, meta-analysis, dan clinical development guidelines,” *Journal of the American Medical Informatics Association*, 32(4), pp. 605–615. Available at: <https://doi.org/10.1093/jamia/ocaf008>.

M, H. dan M.N, S. (2015) “A Review on Evaluation Metrics for Data Classification Evaluations,” *International Journal of Data Mining & Knowledge Management Process*, 5(2), pp. 01–11. Available at: <https://doi.org/10.5121/ijdkp.2015.5201>.

Manning, C., Raghavan, P. dan Schuetze, H. (2009) “*Introduction to Information Retrieval.*” Cambridge: Cambridge University Press. Available at: <https://doi.org/10.1017/CBO9780511809071>

Miao, J. et al. (2024) “Integrating *Retrieval-Augmented Generation* with Large Language Models in Nephrology: Advancing Practical Applications,” *Medicina*, 60(3), p. 445. Available at: <https://doi.org/10.3390/medicina60030445>.

National Information Standards Organization JATS Standardizing Committee (no date) “ANSI/NISO Z39.96-2024, JATS: Journal Article Tag Suite, version 1.4.” NISO. Available at: <https://doi.org/10.3789/ansi.niso.z39.96-2024>.

OpenAI et al. (2024) “GPT-4 Technical Report.” *arXiv*. Available at: <https://doi.org/10.48550/arXiv.2303.08774>.

Otter, D.W., Medina, J.R. dan Kalita, J.K. (2019) "A Survey of the Usages of Deep Learning in Natural Language Processing." arXiv. Available at: <https://doi.org/10.48550/arXiv.1807.10854>.

Ouyang, L. et al. (2022) "Training language models to follow instructions with human feedback." arXiv. Available at: <https://doi.org/10.48550/arXiv.2203.02155>

Radford, A. et al. (2018) "Improving Language Understanding by Generative Pre-Training." Available at: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Rumelhart, D.E., Hinton, G.E. dan Williams, R.J. (1986) "Learning representations by back-propagating errors," *Nature*, 323(6088), pp. 533–536. Available at: <https://doi.org/10.1038/323533a0>.

Sabrina, C. (2025) "CHATBOT TELEGRAM BERBASIS KECERDASAN BUATAN MENGGUNAKAN LARGE LANGUAGE MODEL QWEN 2.5 DAN RETRIEVAL AUGMENTED GENERATION STUDI KASUS PPID UNIVERSITAS LAMPUNG"

Tharwat, A. (2021) "Classification assessment methods," *Applied Computing dan Informatics*, 17(1), pp. 168–192. Available at: <https://doi.org/10.1016/j.aci.2018.08.003>.

Tkaczyk, D. et al. (2015) "CERMINE: automatic extraction of structured metadata from scientific literature," *International Journal on Document Analysis dan Recognition (IJDAR)*, 18(4), pp. 317–335. Available at: <https://doi.org/10.1007/s10032-015-0249-8>.

Touvron, H. et al. (2023) "LLaMA: Open dan Efficient Foundation Language Models." arXiv. Available at: <https://doi.org/10.48550/arXiv.2302.13971>.

Vaswani, A. et al. (2023) "Attention Is All You Need." arXiv. Available at: <https://doi.org/10.48550/arXiv.1706.03762>.

Virgiansyah, M.R. dan Pribadi, M.R. (2025) "Sistem *Chatbot* Helpdesk WhatsApp Berbasis Large Language Model Gemini dengan Fine-Tuning Menggunakan *Retrieval-Augmented Generation* dan Qdrant Vector Database (Studi Kasus: Universitas Multi Data Palembang)," 10(2).

Zhang, J. (2023) "Research on transformer dan attention in applied algorithms," *Applied dan Computational Engineering*, 13(1), pp. 221–228. Available at: <https://doi.org/10.54254/2755-2721/13/20230737>.

Zhao, W.X. et al. (2023) "A Survey of Large Language Models." arXiv. Available at: <https://doi.org/10.48550/ARXIV.2303.18223>.