

DAFTAR PUSTAKA

- Adamopoulou, E., & Moussiades, L. (2020). *An Overview of Chatbot Technology* (hlm. 373–383). https://doi.org/10.1007/978-3-030-49186-4_31
- Ahadi, R., Safaat Harahap, N., Fikry, M., & Kurnia, F. (2025). Retrieval-Augmented Generation in a Web-Based Question Answering System for Fiqh Books. *Journal of Artificial Intelligence and Software Engineering*, 5(2), 626–635. <https://doi.org/10.30811/jaise.v5i2.7005>
- Anam, M. R., Agus Subhan Akbar, & Heru Saputro. (2024). QnA Chatbot with Mistral 7B and RAG method: Traffic Law Case Study. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, 15(03), 207–218. <https://doi.org/10.24843/LKJITI.2024.v15.i03.p06>
- Aumüller, M., Bernhardsson, E., & Faithfull, A. (2020). ANN-Benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. *Information Systems*, 87, 101374. <https://doi.org/10.1016/j.is.2019.02.006>
- Bangor, A., Kortum, P., & Miller, J. (2009a). Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. Dalam *Journal of Usability Studies* (Vol. 4).
- Bangor, A., Kortum, P., & Miller, J. (2009b). Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. Dalam *Journal of Usability Studies* (Vol. 4).
- Baur, D., Ansorg, J., Heyde, C.-E., & Voelker, A. (2025). Development and Evaluation of a Retrieval-Augmented Generation Chatbot for Orthopedic and Trauma Surgery Patient Education: Mixed-Methods Study. *JMIR AI*, 4, e75262–e75262. <https://doi.org/10.2196/75262>
- Bayer, M. (2024). *SQLAlchemy Documentation*. <https://www.sqlalchemy.org>
- Bocklisch, T., Faulkner, J., Pawlowski, N., & Nichol, A. (2017). *Rasa: Open Source Language Understanding and Dialogue Management*.
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2022). *On the Opportunities and Risks of Foundation Models*.
- Brooke, J. (1996). SUS: A “Quick and Dirty” Usability Scale. Dalam *Usability Evaluation In Industry* (hlm. 207–212). CRC Press. <https://doi.org/10.1201/9781498710411-35>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-

- Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodi, D. (2020). *Language Models are Few-Shot Learners*.
- Chapman, P., Julian, C., Randy, K., Thomas, K., Thomas, R., Colin, S., & Rüdiger, W. (2000). *Step-by-step data mining guide*. DaimlerChrysler.
- Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., Marris, L., Petulla, S., Gaffney, C., Aharoni, A., Lintz, N., Pais, T. C., Jacobsson, H., Szpektor, I., Jiang, N.-J., ... Helmholtz, W. (2025). *Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities*. <http://arxiv.org/abs/2507.06261>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*.
- Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.-M., Chen, W., Yi, J., Zhao, W., Wang, X., Liu, Z., Zheng, H.-T., Chen, J., Liu, Y., Tang, J., Li, J., & Sun, M. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3), 220–235. <https://doi.org/10.1038/s42256-023-00626-4>
- Docker Inc. (2024). *Docker Documentation*. <https://docs.docker.com>
- Es, S., James, J., Espinosa Anke, L., & Schockaert, S. (2024). RAGAs: Automated Evaluation of Retrieval Augmented Generation. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 150–158. <https://doi.org/10.18653/v1/2024.eacl-demo.16>
- Fanani, I. (2025). Implementasi Retrieval Augmented Generation Untuk Proposal Tugas Akhir Mahasiswa. *Jurnal Teknologi Komputer dan Informatika*, 3(2). <https://doi.org/10.59820/tekomin.v3i2.336>
- Fenniak, M. (2024). *PyPDF2 Documentation*. <https://pypdf2.readthedocs.io/en/3.x/>
- Følstad, A., & Brandtzaeg, P. B. (2020). Users' experiences with chatbots: findings from a questionnaire study. *Quality and User Experience*, 5(1), 3. <https://doi.org/10.1007/s41233-020-00033-2>
- 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*.
- Gemini Team, Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., Mariooryad, S., Ding, Y., Geng, X., Alcober, F., Frostig, R., Omernick, M., Walker, L., Paduraru, C., Sorokin, C., ...

- Vinyals, O. (2024). *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*.
- Holmes, S., Moorhead, A., Bond, R., Zheng, H., Coates, V., & Mctear, M. (2019). Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? *Proceedings of the 31st European Conference on Cognitive Ergonomics*, 207–214. <https://doi.org/10.1145/3335082.3335094>
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*.
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. Le, Lavril, T., Wang, T., Lacroix, T., & Sayed, W. El. (2023). *Mistral 7B*.
- Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense Passage Retrieval for Open-Domain Question Answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781. <https://doi.org/10.18653/v1/2020.emnlp-main.550>
- LangChain. (2024). *LangChain Docs*. <https://docs.langchain.com/>
- Lewis, J. R. (2018). The System Usability Scale: Past, Present, and Future. *International Journal of Human-Computer Interaction*, 34(7), 577–590. <https://doi.org/10.1080/10447318.2018.1455307>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*.
- Lin, C.-Y. (2004). *ROUGE: A Package for Automatic Evaluation of Summaries*.
- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12, 157–173. https://doi.org/10.1162/tacl_a_00638
- Llama Team, Llama Team, Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., ... Ma, Z. (2024). *The Llama 3 Herd of Models*.
- 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>

- Ma, X., Sun, K., Pradeep, R., & Lin, J. (2021). *A Replication Study of Dense Passage Retriever*.
- Malkov, Y. A., & Yashunin, D. A. (2020). Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4), 824–836. <https://doi.org/10.1109/TPAMI.2018.2889473>
- Malkov, Yu. A., & Yashunin, D. A. (2018). *Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs*.
- Meta. (2024). *WhatsApp Documentation*. <https://developers.facebook.com/docs/whatsapp/>
- Meta. (2025). *The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation*. Meta AI Blog. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed Representations of Words and Phrases and their Compositionality*.
- Muludi, K., Fitria, K. M., Triloka, J., & -, S. (2024). Retrieval-Augmented Generation Approach: Document Question Answering using Large Language Model. *International Journal of Advanced Computer Science and Applications*, 15(3). <https://doi.org/10.14569/IJACSA.2024.0150379>
- Nakash, M., & Dan, B. (2024). *How Much Time does the Workforce Spend Searching for Information in the “new normal”?*
- Nogueira, R., & Cho, K. (2020). *Passage Re-ranking with BERT*.
- Ockerman, S., Gueroudji, A., Oh, S. Y., Underwood, R., Chia, N., Chard, K., Ross, R., & Venkataraman, S. (2025). *Exploring Distributed Vector Databases Performance on HPC Platforms: A Study with Qdrant*.
- OpenAI. (2024). *Tiktoken: BPE tokeniser for use with OpenAI's models*. <https://github.com/openai/tiktoken>
- OpenAI, :, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., Mądry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., ... Malkov, Y. (2024). *GPT-4o System Card*.
- OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., ... Zoph, B. (2024). *GPT-4 Technical Report*.
- Pan, J. J., Wang, J., & Li, G. (2023). *Survey of Vector Database Management Systems*.

- Pan, J. J., Wang, J., & Li, G. (2024). Survey of vector database management systems. *The VLDB Journal*, 33(5), 1591–1615. <https://doi.org/10.1007/s00778-024-00864-x>
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 311. <https://doi.org/10.3115/1073083.1073135>
- Pratama, A. Z. (2025). *Model Chatbot Assistant Kesehatan Menggunakan Large Language Model Meta AI (LLaMA) dengan Pendekatan Retrieval-Augmented Generation*. Institut Teknologi Kalimantan (ITK).
- Pressman, R. S., & Maxim, B. (2014). *Software Engineering*.
- Qdrant. (2024a). *Qdrant Benchmarks*. 2024. <https://qdrant.tech/benchmarks/>
- Qdrant. (2024b). *Qdrant Documentation*. <https://qdrant.tech/documentation/>
- Ramírez, S. (2024). *FastAPI Documentation*. <https://fastapi.tiangolo.com/>
- Redis Ltd. (2024). *Redis Documentation*. <https://redis.io/documentation>
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3980–3990. <https://doi.org/10.18653/v1/D19-1410>
- Samudra, G., Ahmad Turmudi Zy, & Ermanto. (2025). Implementasi Retrieval Augmented Generation (RAG) Dalam Perancangan Chatbot Kesehatan Pencernaan. *JSAI (Journal Scientific and Applied Informatics)*, 8(1), 181–188. <https://doi.org/10.36085/jsai.v8i1.7678>
- Sanjani, L. A., Sarno, R., Sungkono, K. R., Haryono, A. T., Septiyanto, A. F., & Sunaryono, D. (2025). Performance Analysis of LLM Models with RAG and Fine-Tuning T5 for Chatbot Optimization in Call Centers. *2025 International Conference on Computer Sciences, Engineering, and Technology Innovation (ICoCSETI)*, 152–157. <https://doi.org/10.1109/ICoCSETI63724.2025.11018908>
- Sharfina, Z., & Santoso, H. B. (2016). An Indonesian adaptation of the System Usability Scale (SUS). *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 145–148. <https://doi.org/10.1109/ICACSIS.2016.7872776>
- Smith, R. (2024). *Tesseract OCR*. <https://github.com/tesseract-ocr/tesseract>

- Sue, H. (2025, April 24). *Social Media Statistics for Indonesia [Updated 2024]*. Meltwater. <https://www.meltwater.com/en/blog/social-media-statistics-indonesia>
- Sugiyono. (2019). *METODE PENELITIAN KUANTITATIF*.
- Topsakal, O., & Akinci, T. C. (2023). Creating Large Language Model Applications Utilizing LangChain: A Primer on Developing LLM Apps Fast. *International Conference on Applied Engineering and Natural Sciences, 1*(1), 1050–1056. <https://doi.org/10.59287/icaens.1127>
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., ... Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need*.
- Vue.js. (2024). *Vue.js Documentation*. <https://waha.devlike.pro/docs/overview/>
- Waha. (2024). *Waha Documentation*. 2024. <https://waha.devlike.pro/docs/overview/>
- Wallace, R. S. (2009). The Anatomy of A.L.I.C.E. Dalam *Parsing the Turing Test* (hlm. 181–210). Springer Netherlands. https://doi.org/10.1007/978-1-4020-6710-5_13
- Wang, J., Yi, X., Guo, R., Jin, H., Xu, P., Li, S., Wang, X., Guo, X., Li, C., Xu, X., Yu, K., Yuan, Y., Zou, Y., Long, J., Cai, Y., Li, Z., Zhang, Z., Mo, Y., Gu, J., ... Xie, C. (2021). Milvus: A Purpose-Built Vector Data Management System. *Proceedings of the 2021 International Conference on Management of Data*, 2614–2627. <https://doi.org/10.1145/3448016.3457550>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*.
- Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45. <https://doi.org/10.1145/365153.365168>
- White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*.
- Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020).
BERTScore: Evaluating Text Generation with BERT.



Intelligentia - Dignitas