

DAFTAR PUSTAKA

- Alhanai, T., Kasumovic, A., Ghassemi, M. M., Zitzelberger, A., Lundin, J. M., & Chabot-Couture, G. (2025). *Bridging the Gap: Enhancing LLM Performance for Low-Resource African Languages with New Benchmarks, Fine-Tuning, and Cultural Adjustments*.
- Anam, K. (2025). *Punya 40 Juta User Lebih, Transaksi BRImo Tembus Rp1.599 T*. CNBC Indonesia. <https://www.cnbcindonesia.com/tech/20250428103647-37-629237/punya-40-juta-user-lebih-transaksi-brimo-tembus-rp1599-t>
- Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., ... Lin, J. (2025). *Qwen2.5-VL Technical Report* (No. arXiv:2502.13923). arXiv. <https://doi.org/10.48550/arXiv.2502.13923>
- Chen, B., Dao, T., Winsor, E., Song, Z., Rudra, A., & Ré, C. (2021). Scatterbrain: Unifying Sparse and Low-rank Attention. *Advances in Neural Information Processing Systems*, 34, 17413–17426. <https://proceedings.neurips.cc/paper/2021/hash/9185f3ec501c674c7c788464a36e7fb3-Abstract.html>
- Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024). *BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation* (No. arXiv:2402.03216). arXiv. <https://doi.org/10.48550/arXiv.2402.03216>
- Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., & Stoica, I. (2024). *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference* (No. arXiv:2403.04132). arXiv. <https://doi.org/10.48550/arXiv.2403.04132>
- Cho, J., Mahata, D., Irsoy, O., He, Y., & Bansal, M. (2024). *M3DocRAG: Multi-modal Retrieval is What You Need for Multi-page Multi-document Understanding* (No. arXiv:2411.04952). arXiv. <https://doi.org/10.48550/arXiv.2411.04952>
- Choromanski, K., Likhoshesterov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., Belanger, D., Colwell, L., & Weller, A. (2021). *RETHINKING ATTENTION WITH PERFORMERS*.
- Dao, T. (2023). *FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning* (No. arXiv:2307.08691). arXiv. <https://doi.org/10.48550/arXiv.2307.08691>
- Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLORA: efficient finetuning of quantized LLMs. *Proceedings of the 37th international conference on neural information processing systems*.
- Faysse, M., Sibille, H., Wu, T., Viaud, G., Hudelot, C., & Colombo, P. (2024). *ColPali: Efficient Document Retrieval with Vision Language Models* (No. arXiv:2407.01449; Versi 1). arXiv. <http://arxiv.org/abs/2407.01449>
- Ford, N., Richards, M., Sadalage, P., & Dehghani, Z. (2022). *Software Architecture: The Hard Parts*.
- 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* (No. arXiv:2312.10997). arXiv. <https://doi.org/10.48550/arXiv.2312.10997>
- GitHub. (2022). *Confusion about correct learning rate when running contrastive fine-tuning*. Issue #208. [huggingface/setfit](https://github.com/huggingface/setfit). GitHub. <https://github.com/huggingface/setfit/issues/208>
- GoTo. (2024). *GoToCompany/llama3-8b-cpt-sahabatai-v1-instruct*. Hugging Face. <https://huggingface.co/GoToCompany/llama3-8b-cpt-sahabatai-v1-instruct>
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., 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., ... Ma, Z. (2024). *The Llama 3 Herd of Models* (No. arXiv:2407.21783). arXiv. <https://doi.org/10.48550/arXiv.2407.21783>
- Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure* (No. arXiv:2203.05794). arXiv. <https://doi.org/10.48550/arXiv.2203.05794>
- Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., & Guo, J. (2025). *A Survey on LLM-as-a-Judge* (No. arXiv:2411.15594). arXiv. <https://doi.org/10.48550/arXiv.2411.15594>
- Hambarde, K. A., & Proenca, H. (2023). Information Retrieval: Recent Advances and Beyond. *IEEE Access*, 11, 76581–76604. <https://doi.org/10.1109/ACCESS.2023.3295776>
- 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* (No. arXiv:2106.09685). arXiv. <http://arxiv.org/abs/2106.09685>
- Hu, H., Wang, X., Zhang, Y., Chen, Q., & Guan, Q. (2024). A comprehensive survey on contrastive learning. *Neurocomputing*, 610, 128645. <https://doi.org/10.1016/j.neucom.2024.128645>
- Kang, B., Kim, Y., & Shin, Y. (2023). An Efficient Document Retrieval for Korean Open-Domain Question Answering Based on ColBERT. *Applied Sciences*, 13(24), 13177. <https://doi.org/10.3390/app132413177>
- Katharopoulos, A., Vyas, A., Pappas, N., & Fleuret, F. (2020). Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. *Proceedings of the 37th International Conference on Machine Learning*.
- Khade, O., Jagdale, S., Phaltankar, A., Takalikar, G., & Joshi, R. (2025). Challenges in Adapting Multilingual LLMs to Low-Resource Languages using LoRA PEFT Tuning. Dalam K. Sarveswaran, A. Vaidya, B. Krishna Bal, S. Shams, & S. Thapa (Ed.), *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)* (hlm. 217–222). International Committee on Computational Linguistics. <https://aclanthology.org/2025.chipsal-1.22/>
- Khattab, O., & Zaharia, M. (2020). *ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT* (No. arXiv:2004.12832). arXiv. <http://arxiv.org/abs/2004.12832>
- Kleppmann, M. (2017). *Designing Data-Intensive Applications*.
- Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., & Stoica, I. (2023). *Efficient Memory Management for Large*

- Language Model Serving with PagedAttention* (No. arXiv:2309.06180). arXiv. <https://doi.org/10.48550/arXiv.2309.06180>
- Le-Khac, P. H., Healy, G., & Smeaton, A. F. (2020). Contrastive Representation Learning: A Framework and Review. *IEEE Access*, 8, 193907–193934. <https://doi.org/10.1109/ACCESS.2020.3031549>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474. <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>
- Lin, W., Chen, J., Mei, J., Coca, A., & Byrne, B. (2023). *Fine-grained Late-interaction Multi-modal Retrieval for Retrieval Augmented Visual Question Answering*.
- Loshchilov, I., & Hutter, F. (2019). Decoupled Weight Decay Regularization. *International Conference on Learning Representations (ICLR)*.
- Ma, X., Gong, Y., He, P., Zhao, H., & Duan, N. (2023). *Query Rewriting for Retrieval-Augmented Large Language Models* (No. arXiv:2305.14283). arXiv. <https://doi.org/10.48550/arXiv.2305.14283>
- 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>
- Marcus, G. (2020). *The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence* (No. arXiv:2002.06177). arXiv. <https://doi.org/10.48550/arXiv.2002.06177>
- Mathew, M., Bagal, V., Tito, R. P., Karatzas, D., Valveny, E., & Jawahar, C. V. (2021). *InfographicVQA* (No. arXiv:2104.12756). arXiv. <https://doi.org/10.48550/arXiv.2104.12756>
- Mathew, M., Karatzas, D., & Jawahar, C. V. (2021). *DocVQA: A Dataset for VQA on Document Images* (No. arXiv:2007.00398). arXiv. <https://doi.org/10.48550/arXiv.2007.00398>
- Merkel, D. (2014). Docker: Lightweight Linux containers for consistent development and deployment. *Linux J*, 2014(239).
- Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2024). *A Comprehensive Overview of Large Language Models* (No. arXiv:2307.06435). arXiv. <https://doi.org/10.48550/arXiv.2307.06435>
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library* (No. arXiv:1912.01703). arXiv. <https://doi.org/10.48550/arXiv.1912.01703>
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*.
- Rahadika Diana, K. D., & Khodra, M. L. (2023). IndoSBERT: Enhancing Indonesian Sentence Embeddings with Siamese Networks Fine-tuning.

- 2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA), 1–6. <https://doi.org/10.1109/ICAICTA59291.2023.10390469>
- Ramírez, S. (2024). *FastAPI* [Software]. <https://github.com/fastapi/fastapi>
- Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval*, 3(4), 333–389. <https://doi.org/10.1561/15000000019>
- Shah, C., & Croft, W. B. (2004). Evaluating high accuracy retrieval techniques. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2–9. <https://doi.org/10.1145/1008992.1008996>
- Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., Grill, J., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., ... Hussenot, L. (2025). *Gemma 3 Technical Report* (No. arXiv:2503.19786). arXiv. <https://doi.org/10.48550/arXiv.2503.19786>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, 30. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547de91fbd053c1c4a845aa-Abstract.html
- Wirth, R., & Hipp, J. (1999). *CRISP-DM: Towards a Standard Process Model for Data Mining*.
- Wongso, W., Joyoadikusumo, A., Setiawan, D. S., & Limcorn, S. (2024). *LazarusNLP/indonesian-sentence-embeddings: V0.0.1* (Versi v0.0.1) [Software]. Zenodo. <https://doi.org/10.5281/ZENODO.10983756>
- Xia, P., Zhu, K., Li, H., Wang, T., Shi, W., Wang, S., Zhang, L., Zou, J., & Yao, H. (2024). Mmed-Rag: Versatile Multimodal Rag Sys- Tem for Medical Vision Language Models. *NeurIPS 2024 Workshop on Safe Generative AI*.
- Xu, C., Zhu, Z., Wang, J., Wang, J., & Zhang, W. (2024). *Understanding the Role of Cross-Entropy Loss in Fairly Evaluating Large Language Model-based Recommendation* (No. arXiv:2402.06216; Versi 2). arXiv. <https://doi.org/10.48550/arXiv.2402.06216>
- Yang, N. (2022). Financial Big Data Management and Control and Artificial Intelligence Analysis Method Based on Data Mining Technology. *Wireless Communications and Mobile Computing*, 2022, 1–13. <https://doi.org/10.1155/2022/7596094>
- Zhang, J., Huang, J., Jin, S., & Lu, S. (2024). *Vision-Language Models for Vision Tasks: A Survey* (No. arXiv:2304.00685). arXiv. <https://doi.org/10.48550/arXiv.2304.00685>
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). *BERTScore: Evaluating Text Generation with BERT* (No. arXiv:1904.09675). arXiv. <https://doi.org/10.48550/arXiv.1904.09675>
- Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2023). *Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models* (No. arXiv:2309.01219). arXiv. <https://doi.org/10.48550/arXiv.2309.01219>

- Zheng, Y., Zhang, R., Zhang, J., Ye, Y., Luo, Z., Feng, Z., & Ma, Y. (2024). *LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models* (No. arXiv:2403.13372). arXiv. <https://doi.org/10.48550/arXiv.2403.13372>
- Zhuang, W., Cairen, D., & Sun, Y. (2024). TIFD: Tibetan Instruction-Following Dataset for Large Language Models Supervised Fine-Tuning. *Data Intelligence*. <https://doi.org/10.3724/2096-7004.di.2024.0010>
- Hu, W., Gu, J.-C., Dou, Z.-Y., Fayyaz, M., Lu, P., Chang, K.-W., & Peng, N. (2025). Mrag-Bench: Vision-Centric Evaluation for Retrieval-Augmented Multimodal Models. The Thirteenth International Conference on Learning Representations.
- SAS. (1986). Introduction to Semma. https://s2.smu.edu/tfomby/eco5385_eco6380/data/SPSS/SAS%20_%20SEMMA.pdf
- Strand, A. T., Gautam, S., Midoglu, C., & Halvorsen, P. (2024). SoccerRAG: Multimodal soccer information retrieval via natural queries. 2024 international conference on content-based multimedia indexing (CBMI), 1–7. <https://doi.org/10.1109/CBMI62980.2024.10859209>
- Tran, Q.-L., Pham, N. N. D., Truong, Q. T., Nguyen, M. H., Le, H. C., Vu, D. K., Nguyen, V. M. T., Nguyen, V. K., Nguyen, L. P. N. L., Le, T., Dang, M. P., Nguyen, B., Jones, G. J. F., & Gurrin, C. (2025). A RAG Approach for Multi-Modal Open-ended Lifelog Question-Answering. Proceedings of the 2025 International Conference on Multimedia Retrieval, 1303–1312. <https://doi.org/10.1145/3731715.3733263>
- Xia, P., Zhu, K., Li, H., Zhu, H., Li, Y., Li, G., Zhang, L., & Yao, H. (2024). RULE: Reliable Multimodal RAG for Factuality in Medical Vision Language Models. Dalam Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Ed.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (hlm. 1081–1093). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.62>
- Yamanishi, H., Xiao, L., & Yamasaki, T. (2025). TourMLLM: A Retrieval-Augmented Multimodal Large Language Model for Multitask Learning in the Tourism Domain. Proceedings of the 2025 International Conference on Multimedia Retrieval, 1654–1663. <https://doi.org/10.1145/3731715.3733450>
- Yuan, J., Sun, S., Omeiza, D., Zhao, B., Newman, P., Kunze, L., & Gadd, M. (2024). RAG-Driver: Generalisable Driving Explanations with Retrieval-Augmented In-Context Learning in Multi-Modal Large Language Model. Robotics: Science and System XX.