Komodo-7B with Hybrid Retrieval and Q-LoRA for Indonesian Population Administration Question Answering

Anindo Saka Fitri[1], Abdul Rezha Efrat Najaf[1], Eko Wahyudi[2], Adelia Azizatul Haq[3], Sugiarto[4] & I Gede Susrama Mas Diyasa[5,*]

[1]Department of Information System, Faculty of Computer Science, University of Pembangunan Veteran Jawa Timur, Rungkut Madya, Surabaya 60294, Indonesia
[2]Department of Law, Faculty of Law, University of Pembangunan Veteran Jawa Timur, Rungkut Madya, Surabaya 60294, Indonesia
[3]Department of Data Science, Faculty of Computer Science, University of Pembangunan Veteran Jawa Timur, Rungkut Madya Surabaya 60294, Indonesia
[4]Department of Digital Business, Faculty of Computer Science, University of Pembangunan Veteran Jawa Timur, Rungkut Madya Surabaya 60294, Indonesia
[5]Department of Master Information Technology, Faculty of Computer Science, University of Pembangunan Veteran Jawa Timur, Rungkut Madya, Surabaya 60294, Indonesia
*E-mail: igsusrama.if@upnjatim.ac.id

Abstract. Fast, responsive, and informative public services are societal demands that must be fulfilled by government agencies, among which the Department of Population and Civil Registration of Surabaya City. To enhance service quality, this study developed a Large Language Model (LLM)-based Question Answering (QA) system to address public inquiries regarding Identity Card (ID) and Family Card (FC) services. The proposed system utilizes the Komodo-7B model, which was customized using Quantized Low-Rank Adaptation (Q-LoRA) fine-tuning and integrated with a Retrieval-Augmented Generation (RAG) approach to improve the accuracy and relevance of the generated responses. The training process leveraged a real-world complaint dataset from Disdukcapil alongside the open-source MS MARCO dataset. Furthermore, the RAG implementation employs sentence vectorization via SentenceTransformer and cosine similarity-based context retrieval. System performance was evaluated using ROUGE and METEOR metrics across four scenarios: Komodo-7B Base, RAG Komodo-7B Base, Fine-Tuned Komodo-7B, and RAG Fine-Tuned Komodo-7B. The results show that the RAG Fine-Tuned Komodo-7B configuration delivered the best performance, achieving F1-Scores of 0.3554 for ROUGE-1, 0.3096 for ROUGE-L, and 0.2886 for METEOR.

Keywords: *Komodo-7B; large language models; Q-LoRA; question answering; retrieval-augmented generation; population administration; public services*.

Received August 26th, 2025, Revised June 5th, 2026, Accepted for publication June 11th, 2026.

1 Introduction

Public service delivery frequently serves as a benchmark for the success of government agencies. High quality service fosters a positive image of the government bureaucracy, aligning with the principles of good governance [1]. This paradigm compels Regional Apparatus Organizations (OPD), as public service providers, to continuously enhance service quality through innovative approaches that meet public expectations. Furthermore, rapid digital transformation in Indonesia supports governmental innovation in public service delivery [2]. The Department of Population and Civil Registration (Disdukcapil) in one of the major cities in Indonesia operates as a regional agency responsible for Population Administration and Civil Registration [3]. Accordingly, Disdukcapil is obligated to continuously elevate the quality of its population services.

One of the facilities provided by Disdukcapil is a complaint service accessible via social media and call centers. This service functions as a primary communication channel for the general public, village officials, and subdistrict officers encountering obstacles or requiring information regarding population administration and civil registration. The agency consistently receives a significantly high volume of questions and complaints, with the most frequent inquiries pertaining to Identity Cards (ID) and Family Cards (FC). Given this high volume of inquiries, an automated question answering (QA) system is urgently required to alleviate the administrative burden on officers. The implementation of such a QA system directly supports the broader objective of realizing good governance.

Recent advancements in QA systems have been largely driven by the development of Large Language Models (LLMs), which overcome the limitations of rule-based methods and traditional machine learning approaches that rely heavily on manual feature engineering and extensive labeled datasets [4]. Trained on large scale text corpora using the Transformer architecture, LLMs are capable of capturing contextual dependencies, syntactic structures, and semantic representations with profound depth [5-7]. Although several studies have explored the application of LLMs in specific domains [8-11], deploying LLMs without domain adaptation often generates inaccurate and biased responses, despite appearing highly convincing. In the context of population administration, such inaccuracies pose significant risks to citizens' administrative rights and legal processes. Therefore, an approach is required that not only adapts the LLM to the target domain but also ensures strict accuracy regarding formal Indonesian administrative terminology, while simultaneously accommodating the computational infrastructure constraints of local government institutions.

Consequently, this study proposes a domain specific QA system that integrates three core components based on technical and practical considerations. First, Q-LoRA (Quantized Low-Rank Adaptation) [11] was selected as the fine-tuning method to address the computational infrastructure limitations commonly faced by local government institutions [12]. Second, Komodo-7B was chosen as the foundation model due to its specialized architecture for the Indonesian language and eleven regional dialects, facilitated by a localized tokenizer expansion [13]. Third, a Hybrid Retrieval mechanism combining BM25L (lexical search) and MPNet (semantic embedding) through Reciprocal Rank Fusion (RRF) was integrated as a Retrieval Augmented Generation (RAG) layer to maximize the precision and contextual relevance of the retrieved documents [14].

To the best of the authors' knowledge, no existing research has systematically integrated a BM25L and MPNet based Hybrid Retrieval mechanism with Q-LoRA fine-tuning on an Indonesian centric language model (Komodo-7B) for a public service QA. Thus, the main contributions of this study are:

1. *Establishment of the first benchmark in Indonesian GovTech*: This study empirically evaluated the combination of Komodo-7B, Q-LoRA, and Hybrid RAG for a public service QA system, thereby filling a literature gap regarding artificial intelligence in Indonesian Government Technology (AI GovTech).
2. *Curation of a domain-specific real-world dataset*: This research developed and released an evaluation dataset constructed from real citizen complaint logs at the regional Disdukcapil, which were augmented and subsequently combined with the MS MARCO dataset [15]. This provides a novel testing standard deeply rooted in the operational reality of Indonesian population administration.
3. *Systematic analysis and practical guidelines*: This study presents a comprehensive performance comparison of four model configurations (Komodo-7B Base, RAG Base, Fine-Tuned, and RAG Fine-Tuned), complemented by an error breakdown analysis. These findings provide practical guidelines for government institutions to implement LLMs safely while mitigating AI hallucination risks.

2 Related Works

In tandem with the rapid advancements in question answering (QA) research and attention mechanisms [16], several previous studies have explored approaches based on Large Language Models (LLMs). These studies provide a crucial foundation for research focused on enhancing LLM performance through the integration of hybrid retrieval and efficient parameter optimization. To illustrate this, several relevant studies are summarized in Table 1.

Table 1 Summary of related works on LLM-based question answering and optimization techniques.

Reference	Case Study	Method	Key Findings
Alawwad, *et al.* (2025) [8]	LLM for textbook QA (CK12-TQA, 26,260 questions from 1,076 science lessons)	Llama-2 + LoRA + RAG (Pinecone vector store + OpenAI text-embedding-ada-002, dot product similarity)	Accuracy increased from 35.87% (zero shot) to 84.24% following LoRA fine-tuning and RAG integration.
Hakim, *et al.* (2025) [9]	Curation of Indonesian ethical and unethical instructions ('Anak Baik' dataset, 5,298 instruction response pairs)	Comparison of Llama-2-7B derivatives (Cendol, Komodo) vs multilingual models (Sealion, Bactrian X) using LoRA fine-tuning	Komodo-7B demonstrated the best performance after fine-tuning with a BLEU score of 45.64 and ROUGE-L of 35.29, outperforming Bactrian X, which initially excelled in zero shot/five shot scenarios.
Chaubey, *et al.* (2024) [10]	Chatbot development based on the Open Assistant Guanaco dataset from Hugging Face	Comparative analysis of RAG, fine-tuning, and prompt engineering utilizing vector embeddings on data chunks	Fine-tuning delivered the best performance (87.5% accuracy, BLEU 0.81, HES 8.9), although a perplexity of 10.3 indicated room for improvement; a combination of RAG and fine-tuning is recommended.
Dettmers, *et al.* (2023) [11]	Efficient LLM fine-tuning of up to 65 billion parameters on a single 48 GB GPU (tested on Llama 7B/13B/33B/65B, T5, RoBERTa)	Q-LoRA: 4-bit NormalFloat (NF4) Quantization, Double Quantization, and Paged Optimizers on Low Rank Adapters	Q-LoRA performed on par with 16-bit full fine-tuning and 16-bit LoRA (~1 point difference on 5 shot MMLU); Guanaco 33B (4-bit) outperformed Vicuna 13B with a smaller memory footprint (21 GB vs 26 GB).

Based on the aforementioned review, this study systematically mitigates the limitations of prior research by combining three complementary components. First, selecting Komodo-7B as the foundation model addresses the constraints of

existing systems that predominantly focus on the English language [17, 18]. Second, the implementation of Q-LoRA [11] supersedes the standard LoRA utilized in those previous studies [8, 10]. Q-LoRA performs on par with 16-bit full fine-tuning at a substantially lower computational cost, presenting a viable solution for local government institutions operating under infrastructure constraints. Third, the integration of BM25L–MPNet Hybrid Retrieval using RRF expands upon earlier retrieval only approaches by embedding them into a Retrieval Augmented Generation (RAG) framework. This effectively resolves the shortcomings of relying on a single semantic retrieval method [8] and overcomes the isolated implementations of RAG and fine-tuning [10]. Consequently, this paper fills the research gap by integrating BM25L–MPNet Hybrid Retrieval with Q-LoRA fine-tuning on the Komodo-7B model specifically for Indonesian public service QA. Theoretically, synthesizing local linguistic adaptation, parameter efficiency, and enhanced retrieval precision and recall promises to yield an accurate, efficient, and contextually aware QA system for Indonesia's population administration domain.

3 Research Methodology

This section outlines the research methodology, which is divided into three subsections: system architecture, dataset, and research scheme.

3.1 System Architecture

This QA system adopts a hybrid RAG approach based on the Komodo-7B model, which was enhanced through the Q-LoRA method, as illustrated in Figure 1 [19]. The retrieval stage integrates BM25L lexical search ($k_1 = 1.5$; $b = 0.75$; $\delta = 0.5$) and MPNet semantic search [20] (paraphrase multilingual mpnet base v2) utilizing cosine similarity [21]. Both scores were linearly combined ($\alpha = 0.5$) to extract the top-three references with the highest aggregate scores.

For domain adaptation, Komodo-7B was trained using Q-LoRA (4-bit NF4 quantization) on an NVIDIA Tesla T4 15 GB GPU. Guided by established hyperparameter optimization research [22], the configuration focused on a combination of a 3e-4 learning rate, a LoRA rank of 256, and an alpha of 64 to maximize performance. Given infrastructure constraints, a step-based approach (max_steps 1,000, batch size 1) was employed instead of an epoch-based one to provide more granular control over weight updates. In the final stage, the top-three references were combined with the user's query into a structured prompt to be processed by the model (max_new_tokens = 200).

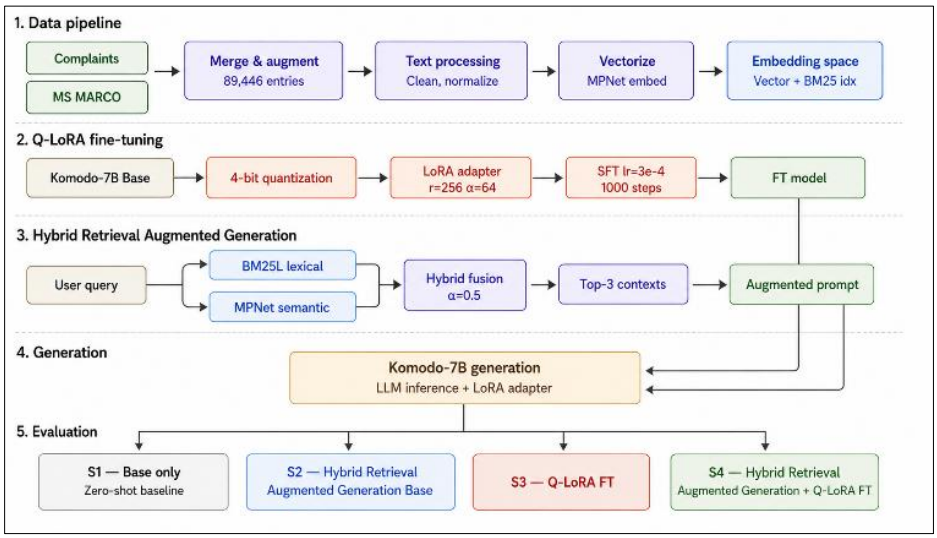

Figure 1 System design.

3.2 Data

This study utilized two primary data sources. The primary data consisted of complaint records from Disdukcapil Surabaya City, supported by business process reports (2023-2024), Surabaya Mayoral Regulation No. 38/2024, and related service regulations. The utilization of this data was officially authorized and validated by the CRM Division of Disdukcapil. To ensure citizen privacy, all Personally Identifiable Information (PII) attributes, such as National Identity Numbers (NIK), full names, specific addresses, and telephone numbers, were anonymized using automated placeholders prior to evaluation. As supporting data, 82,326 rows from the MS MARCO dataset [15] were used, translated into Indonesian, and formatted to ensure consistency with the primary data structure.

The preprocessing pipeline encompasses punctuation removal, case folding, and the normalization of abbreviations and slang terms. To overcome corpus limitations, 445 original records were retained as the test set, while the training data underwent multi-level augmentation: manual expansion yielded 1,780 entries, which were subsequently processed using Easy Data Augmentation (EDA) to produce 7,120 training entries. The final distribution and allocation of the dataset subsets are detailed in Table 2.

Table 2 Dataset subset distribution and allocation.

Subset	Source	Total Entries	Purpose
Train	Disdukcapil (augmented)	7,120	Q-LoRA fine-tuning
Train	MS MARCO (translated)	82,326	Q-LoRA fine-tuning
Total Train		**89,446**	
Test	Disdukcapil (original data)	445	Evaluation of all scenarios

3.3 Research Scheme

Four scenarios were designed to isolate the contributions of retrieval and fine-tuning: (S1) Komodo-7B Base without retrieval as a zero shot baseline; (S2) RAG Komodo-7B Base with top-three retrieval to measure the pure retrieval effect; (S3) Fine-Tuned Komodo-7B without retrieval to assess the pure domain adaptation effect; and (S4) RAG Fine-Tuned Komodo-7B with Top 3 retrieval as a combination of both. All scenarios were evaluated on the same test set (445 entries). Automatic evaluation employed ROUGE-1, ROUGE-L, and METEOR (using the rouge_score library with use_stemmer = true), calculating precision, recall, and F1-scores for each response reference pair [23, 24].

To address the limitations of surface-level quantitative metrics, model responses were manually evaluated against official answers validated by the regional Disdukcapil. The assessment was binary; a response was deemed correct if it aligned substantively with the reference, contained no factual administrative errors, and avoided mere repetition of the query. Accuracy was then calculated as the ratio of correct answers to the total test responses.

4 Results

This section presents the testing results to provide a comprehensive overview of the developed system's performance, starting from the response generation process and evaluation using quantitative metrics.

4.1 Model Evaluation

The evaluation was conducted across three dimensions: automatic metrics (ROUGE-1, ROUGE-L, METEOR), manual evaluation, and generation time efficiency. This multi perspective approach is necessary because surface level metrics and human assessments frequently yield divergent conclusions regarding the substantive quality of the answers.

Table 3 Model Performance, Manual Accuracy, and Generation Time Efficiency

	ROUGE-1			ROUGE-L			METEOR
Model	**Precision**	**Recall**	**F-1 Score**	**Precision**	**Recall**	**F-1 Score**	
Komodo-7B Base	0.1037	0.3084	0.1488	0.0899	0.2724	0.1300	0.1724
RAG Komodo-7B Base	0.0921	0.8358	0.1641	0.0863	0.7844	0.1539	0.3551
Fine-Tuned Komodo-7B	0.4754	0.2615	0.3159	0.3995	0.2193	0.2644	0.2166
RAG Fine-Tuned Komodo-7B	0.4437	0.3395	0.3554	0.3906	0.2950	0.3096	0.2886

RAG Fine-Tuned Komodo-7B achieved the highest F1-scores (ROUGE-1 0.3554; ROUGE-L 0.3096; METEOR 0.2886), whereas pure fine-tuning excelled in precision due to a tendency to select words closely matching the ground truth with minimal elaboration. Conversely, RAG Komodo-7B Base exhibited an anomaly with high recall (0.8358) and low precision (0.0921): the model merely copied the retrieved context, covering almost all reference n grams but lacking substance, resulting in an F1-score of only 0.16.

Since ROUGE is susceptible to awarding high scores for surface level copying, a binary reference based correctness check was performed based on: (i) substance alignment with the reference, (ii) absence of factual errors, and (iii) avoidance of verbatim repetition of the query. As shown in the previously referenced Table 3, the results highlight a stark discrepancy: although RAG Base performed well in terms of ROUGE recall, it failed to produce any valid answers (0/445). In contrast, RAG Fine-Tuned emerged as the most balanced configuration, achieving an accuracy of 60.00%.

Table 4 Manual test result.

Scenario	True	False	Accuracy
Komodo-7B Base	0	445	0.00%
RAG Komodo-7B Base	0	445	0.00%
Fine-Tuned Komodo-7B	251	194	56.40%
RAG Fine-Tuned Komodo-7B	267	178	60.00%

Table 5 Average generated duration (s).

Scenario	Duration
Komodo-7B Base	145.7362
RAG Komodo-7B Base	151.3587
Fine-Tuned Komodo-7B	25.1145
RAG Fine-Tuned Komodo-7B	32.5731

Fine-tuning significantly accelerated inference due to the PEFT nature of only adjusting a small subset of parameters. RAG fine-tuning added approximately 7 seconds due to the retrieval stage, yet it remained 4.6 times faster than the base model. A generation duration of 32.5 seconds is feasible for asynchronous scenarios such as COEX, Instagram, email, and Wargaku channels, though not yet ideal for real-time chats requiring sub-ten second responses. Consequently, this configuration is suitable for integration as a front office assistance layer for Disdukcapil, with further optimizations such as kv-cache, batching, and runtime quantization required before deployment in fully interactive scenarios.

4.2 RAG Fine-Tuned Komodo-7B Generation Results and Error Breakdown

The best performing scenario was analyzed to map response deficiencies. Qualitatively, the model successfully constructed contextual answers, such as referencing Klampid or district offices for Family Card printing procedures. However, issues remained, such as overly brief responses (e.g., a simple 'Yes' for complex 'split Family Card (KK)' inquiries), which reduced informativeness. To systematically categorize errors, an error breakdown was conducted on 445 test responses (Table 6).

Table 6 Error breakdown of RAG fine-tuned Komodo-7B (n = 445).

Category	Occurred	% Occurred
Factual Hallucination	158	35.51%
Overly Brief Response	118	26.52%
Repetitive Output	26	5.84%

The category Factual Hallucination remains the dominant weakness, encompassing errors in location, procedures, or administrative provisions. For example, the model inaccurately confirmed that nonresidents could change their Identity Card (ID) address in Surabaya, whereas this is restricted to local residents. The root causes are twofold: (i) high semantic similarity between administrative documents specifically between Identity Card (ID) and Family Card (FC) leads to the retrieval of contextually relevant but domain-incorrect documents, and (ii) the base model's limited reasoning capabilities when encountering questions requiring rule exceptions.

The category Overly Brief Responses reflects a tendency to respond with 'Yes/No' or short phrases (e.g., '7 working days') without supporting procedural context. This pattern is likely influenced by the training set, which is based on direct service logs, causing the model to prioritize brevity.

The category Repetitive Output includes the repetition of queries or system instructions. This percentage is relatively low, a significant improvement over the RAG Komodo-7B Base model, which was plagued by prompt echoing, indicating that Q-LoRA fine-tuning successfully mitigated but did not entirely eliminate these degenerative patterns. Overall, the primary bottleneck lies not in the retrieval which successfully identified relevant references for 401/445 (90.1%) of queries but within the generation layer. This underscores the necessity for supplementary mechanisms beyond basic fine-tuning and RAG.

4.3 Limitations

This study presents four primary limitations. First, the dominance of factual hallucinations (35.51%) within the 60.00% accuracy rate indicates that the system has not yet met the production-ready reliability threshold required for the legal administrative domain, particularly regarding inquiries involving rule exceptions. Second, the evaluation, which was restricted to a closed test set of 445 entries for ID and FC services in Surabaya, does not fully reflect the interactions by real users nor the generalization of regulations to other cities. Third, the absence of independent annotators precludes an objective measurement based on inter annotator agreement (e.g., Cohen's Kappa). Fourth, computational constraints utilizing a single Tesla T4 GPU limited a more comprehensive exploration of LoRA hyperparameter combinations.

4.4 Implications and Future Work

Practical Implications and SPBE Alignment. While the integration of LLMs aligns with the national agenda of the Presidential Regulation on the Electronic-based Government System (SPBE) No. 95/2018 and Q-LoRA offers a cost-effective computational solution, the current RAG Fine-Tuned Komodo-7B configuration is not yet optimal for immediate real world deployment. Due to its relatively low accuracy and the critical nature of legal administrative services, implementing the system as a direct first line responder on public complaint channels currently presents significant operational risks. These shortcomings reclassify the current system from a deployment ready solution to a foundational baseline. Operationally, its current capability is strictly limited to assisting human officers in managing low-risk, repetitive questions under tight supervision. Consequently, these limitations underscore the absolute necessity for substantial refinements before practical implementation can be realized.

Hallucination Mitigation and Future Research. To suppress the AI hallucination rate (35.51%) and address the study's limitations, several future research agendas are recommended: (1) knowledge graph integration to resolve semantic similarities among administrative entities; (2) application of Chain-of-Thought (CoT) and Natural Language Inference (NLI) verification to ensure logical

reasoning and proactively detect contradictions; (3) an abstention mechanism that trains the model to respond with "please wait for an officer" when confidence is below the threshold; (4) field testing (A/B Testing) and multi annotator evaluation to measure the deflection rate in real world scenarios; and (5) exploration of instruction-tuned models (such as Sahabat AI or Merak) coupled with cross city corpus expansion to enhance generalization capabilities.

5 Conclusion

This study demonstrated that Q-LoRA fine-tuning on the Komodo-7B model significantly enhanced system performance compared to the base model, as evidenced by consistent improvements in ROUGE and METEOR scores. The integration of Retrieval-Augmented Generation (RAG) into the fine-tuned model proved superior in increasing recall. However, this introduced a trade-off in the form of reduced precision, where the RAG Fine-Tuned configuration became prone to generating overly brief responses or experiencing factual inaccuracies. Computationally, fine-tuning was shown to significantly accelerate inference time; the additional latency introduced by the hybrid retrieval process in the RAG scenario remained substantially more efficient than the base model's computation. From a practical application perspective, while the system is not yet ready for fully autonomous deployment due to the critical nature of legal administrative services, it serves as an effective foundational baseline to assist human officers in managing low-risk, repetitive public inquiries. Furthermore, given that manual evaluation results revealed that automatic metrics do not fully capture the substantive quality of the answers, future development is recommended to pivot toward a graph-based knowledge base approach and incorporate human-in-the-loop validation. This approach is considered more relevant for precisely capturing the rules and structured relationships among population administration service entities.

Acknowledgement

We would like to express our sincere gratitude and appreciation to the Institute for Research and Community Service (LPPM) of UPN 'Veteran' Jawa Timur for the grant support and trust provided. This grant was an important support in encouraging the implementation of research and community service activities as well as strengthening the contribution of the academic community in producing innovative works and impactful outcomes. We hope that the collaboration and continuous support from LPPM UPN 'Veteran' Jawa Timur will continue to enhance research development and community engagement activities in the future.

References

[1] Irianto, H., Kurniawan, A. & Mulyono, A., *Optimizing Services to Achieve Good Governance at the Mini Public Service Mall in Sukodono District, Sidoarjo Regency*, Jurnal Intelektual Administrasi Publik dan Ilmu Komunikasi**, 8**(1), 2022. (Text in Indonesian)

[2] Andrew, B.F. & Mei, R.A., *The New Generation Klampid System Independently Addresses Population and Civil Registration Problems in Surabaya*, Masyarakat Mandiri: Jurnal Pengabdian dan Pembangunan Lokal, **1**(3), 2024. (Text in Indonesian)

[3] Hardi, W., Suprastiyo, A. & Retno, S.A., *Model of Implementation of Population and Civil Registration Services in the Surabaya City Government Indonesia*, Wseas Transactions on Environment and Development, **19**, 2023.

[4] Abbasiantaeb, Z. & Momtazi, S., *Text-based Question Answering from Information Retrieval and Deep Neural Network Perspectives: A Survey*, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, **11**(6), 2020.

[5] Shaikh, M.R, Nida, S., Khalid, M., Qurrat-ul-Ain, N. & Talha, M., *Transformers as the Foundation of Large Language Models: A Comprehensive Review*, International Journal of Innovations in Science & Technology, **7**(4), 2025.

[6] Ibomoiye, D.M., Nobert, J., George, O., Oyindamola, O.O., Ebenezer, E. & Cameron, M., *Large language models: an overview of foundational architectures, recent trends, and a new taxonomy*, Discover Applied Sciences, **7**(1027), 2025.

[7] Subhash, N., Bandyopadhyay, S., Zhang, J., et al., *Transformers and large language models in healthcare: A review*, Artif Intell Med, **154**, 102900, 2024.

[8] Alawwad, H.A., Alhothali, A., Naseem, U., Alkhathlan, A. & Jamal, A., *Enhancing textual textbook question answering with large language models and retrieval augmented generation*, Pattern Recognition, **162**(5), 111332, 2025.

[9] Hakim, S.A., Perdana, R.S. & Fatyanosa, T.N., *Anak Baik (Good Boy): A Low-Cost Approach to Curate Indonesian Ethical and Unethical Instructions*, Proceedings of the Second Workshop in South East Asian Language Processing, 2025.

[10] Chaubey, H. K., Tripathi, G., Ranjan, R. & and Gopalaiyengar, S.K., *Comparative Analysis of RAG, Fine-Tuning, and Prompt Engineering in Chatbot Development*, Proceedings of the International Conference on Future Technologies for Smart Society (ICFTSS), 2024.

[11] Dettmers, T., Pagnoni, A., Holtzman, A. & Zettlemoyer, L., *Q-LoRA: Efficient Finetuning of Quantized LLMs*, Proceedings of the Conference on Neural Information Processing Systems (NeurIPS), 2023.

[12] Nur, D., Kendry, W.D., & Puspitaningtyas, A., *Public Complaints Service Through the 'My Citizens of Surabaya' Application as a Manifestation of E-Governance in the City of Surabaya,* Triwikrama: Jurnal Ilmu Sosial, **4**, 2023. (Text in Indonesian)

[13] Puspita, S.R., *Komodo-7B: The Latest Multilingual AI Model for Regional Languages)* https://www.cloudcomputing.id/berita/komodo-7b-ai-multibahasa, 2024

[14] Maryamah, M., Wilsen, G., Suhalim, C.T., Septiana, R., Fajar, A. & Solihin, M.I., *Hybrid Information Retrieval with Masked and Permuted Language Modeling (MPNet) and BM25L for Indonesian Drug Data Retrieval*, IEEE Xplore, International Conference on Knowledge and Smart Technology (KST), 2024.

[15] Craswell, N., Mitra, B., Yilmaz, E., Campos, D. & Lin, J., *MS MARCO: Benchmarking Ranking Models in the Large-Data Regime*, in IEEE X-Plore, Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021.

[16] Brauwers, G. & Frasincar, F., *A General Survey on Attention Mechanisms in Deep Learning,* IEEE Trans Knowl Data Eng, **35**(4), pp. 3279-3298, 2023.

[17] Jiwandono, R., *Yellow.ai Launches Komodo-7B, Indonesia's First LLM Trained in 11 Regional Languages* https://www.techverse.asia/techno/6358/08032024/yellowai-meluncurkan-komodo-7b-llm-pertama-di-indonesia-yang-dilatih-11-bahasa-daerah, 2024. (Text in Indonesian)

[18] Owen, L., Tripathi, V., Kumar, A. & Ahmed, B., *Komodo: A Linguistic Expedition into Indonesia's Regional Languages*, Mar. 2024. Available: http://arxiv.org/abs/2403.09362

[19] Pujiono, I., Agtyaputra, I.M. & Ruldeviyani, Y., *Implementing Retrieval-Augmented Generation and Vector Databases for Chatbots in Public Services Agencies Context,* Jurnal Ilmu Pengetahuan dan Teknologi Komputer, **10**(1), pp. 216-223, 2024.

[20] Song, K., Tan, X., Qin, T., Lu, J. & Liu, T.Y., *MPNet: Masked and Permuted Pre-training for Language Understanding*, in Proceedings of the Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada, 2020

[21] Kurniawan, R.F. & Arif, M.F., *Implementation of Text Mining Using the Cosine Similarity Method for Classifying News Content in Posts on the Pasuruan Traffic and Crime Info Facebook Group*, JAMI: Jurnal Ahli Muda Indonesia, **3**(1), pp. 9-17, 2022. (Text in Indonesian)

[22] Tribes, C., Benarroch-Lelong, S., Lu, P. & Kobyzev, I., *Hyperparameter Optimization for Large Language Model Instruction-Tuning*, Jan. 2024, Available: http://arxiv.org/abs/2312.00949

[23] Walker II, S.M., *What is the ROUGE Score (Recall-Oriented Understudy for Gisting Evaluation)?,* https://klu.ai/glossary/rouge-score. 09 November 2024.

[24] Masdiyasa, I.G.S., Purnama, I.K.E. & Mauridhi, H. P., *A New Method to Improve Movement Tracking of Human Sperms*, IAENG International Journal of Computer Science, **45**(4), IJCS_45_4_05, 2020.