

## Retrieval-Augmented Generation (RAG) Chatbot for Handling Customer Complaints in the Energy Sector

**Ibnu Prastowo Haryono Putro**<sup>1\*,2</sup>, **Jefry Antoni**<sup>1,2</sup>, **Maulana Krisna Adhitya**<sup>1,2</sup>, **Neng Ayu Herawati**<sup>1,3</sup>,  
**Ayu Purwarianti**<sup>1,3</sup>, **Nugraha Priya Utama**<sup>1,3</sup>

<sup>1</sup> School of Electrical Engineering and Informatics, Bandung Institute of Technology (ITB), Indonesia

<sup>2</sup> PT PLN (Persero), Indonesia

<sup>3</sup> Center of Excellence for AI: Computer Vision, NLP, and Big Data Analytics, Bandung Institute of Technology, Indonesia

\*Corresponding Author: 23524030@mahasiswa.itb.ac.id

---

### Abstract

Fast and accurate customer service is critical in the energy sector, especially for large-scale utilities like PLN. This study introduces a novel Retrieval-Augmented Generation (RAG)-based chatbot tailored for PLN's internal operational context to automate customer complaint resolution in Bahasa Indonesia. In contrast to previous approaches that utilize only fine-tuned LLMs or retrieval-based question answering, our system uniquely integrates internal complaint records stored in internal database with a local Indonesian-optimized LLM through LangChain orchestration. The proposed architecture features temporal and linguistic preprocessing, vector embedding using FAISS, and a dynamic clarification-fallback mechanism, ensuring context-aware and grounded responses. This work contributes a scalable framework for deploying generative AI in high-stakes public utility settings, emphasizing data privacy, language fidelity, and real-time applicability. Evaluation results both simulated and human-reviewed demonstrate the chatbot's effectiveness, achieving BLEU-4 of 46.5 and ROUGE-L of 0.63, with 92% of answers rated helpful. These findings underscore the model's potential to enhance customer experience and operational efficiency in Indonesia's energy sector.

**Keywords:** Chatbot, Customer Complaint, Generative AI, Large Language Model (LLM), Retrieval Augmented Generation (RAG)

---

### INTRODUCTION

Artificial intelligence (AI) has emerged as a fundamental element in the modernization of public service delivery, significantly improving both accessibility and responsiveness (Cortes-Cediel et al., 2023; Wirtz et al., 2024). In the context of Indonesia, the digital transformation of state-owned enterprises, notably the national electricity utility company PLN (Perusahaan Listrik Negara), represents a pivotal agenda (Ministry of Communication and Informatics, 2021). PLN caters to tens of millions of customers and manages an extensive array of complaints on a daily basis, encompassing issues such as service interruptions and billing discrepancies (PLN Indonesia Power, 2024). Effectively addressing these challenges is crucial for sustaining public confidence and ensuring operational integrity (Yang et al., 2023).

Chatbots present a viable solution for automating frequent customer service engagements, allowing organizations to handle routine inquiries more efficiently without direct human intervention (Ferraro et al., 2024). While traditional chatbots have predominantly relied on rule-based frameworks or pre-defined responses, recent advancements in large language models (LLMs) have enabled more dynamic and natural conversational capabilities (Chow et al., 2025). Nevertheless, standard LLMs often encounter the problem of hallucination - producing plausible-sounding yet factually inaccurate information - which poses significant challenges in domains that demand strict factual precision (such as the utilities sector) (Z. Ji. et al., 2023).

To mitigate this issue, Retrieval-Augmented Generation (RAG) has emerged as an efficacious strategy that amalgamates the linguistic fluency of LLMs with the factual reliability of information retrieval systems (Jiang et al., 2025; Gracel et al., 2025; Zhang et al., 2024). Within a RAG framework, user inquiries are aligned with a corpus of documents (retrieved through semantic similarity), and these documents are subsequently incorporated into the prompt provided to the LLM (Rauschecker et al., 2025). This methodology ensures that the responses generated by the model are informed by current, contextually pertinent data.

The novelty of this research lies in the development of a locally deployable, privacy-preserving RAG chatbot that integrates PLN's internal complaint records with a semantic vector search (via FAISS) and a locally optimized Indonesian LLM executed entirely on private infrastructure to meet data governance regulations. Unlike existing works that rely on open-domain datasets or cloud-based LLMs, this approach ensures factual grounding, linguistic appropriateness, and operational readiness for real-world utility service. The objectives of this research are threefold. First, it aims to formulate

a resilient Retrieval-Augmented Generation (RAG) architecture that effectively meets the operational demands of PLN’s customer service in handling complaint-related queries. Second, the study focuses on developing a robust preprocessing pipeline that can accommodate linguistic variations and account for temporal discrepancies in complaint data, ensuring high-quality input for downstream processing. Third, it seeks to evaluate the chatbot’s performance through a combination of automated metrics – such as BLEU and ROUGE – and qualitative assessments performed by human evaluators. All data processing is carried out within local infrastructure to comply with data privacy regulations and to ensure real-time responsiveness of the system (Aditya et al., 2023; Cahyawijaya et al., 2024)

AI-enhanced chatbots have emerged as an essential element in the optimization of public service delivery and enhancement of customer experiences. In Indonesia, Fadhlurohman et al. (2023) created an Indonesian-language chatbot for the Jakarta Kini (JAKI) application, utilizing IndoBERT embeddings alongside the DIET classifier within the Rasa framework (Fadhlurohman et al., n.d.). This integration achieved an impressive F1-score of 0.93, exemplifying the efficacy of employing locally tailored language models to comprehend user intentions in Bahasa Indonesia.

In the energy and utilities sector, the implementation of chatbots is also on the rise. A report by Ofgem (2023) underscores the utilization of AI chatbots in utility firms to manage fundamental energy-related customer interactions. These systems alleviate the burden on human agents from monotonous tasks, enabling them to concentrate on more intricate issues (LLP & 2023., n.d.). PLN, due to its extensive customer network, stands to gain significantly from this strategy, particularly in enhancing the timeliness and consistency of complaint resolution.

Retrieval-Augmented Generation (RAG) represents a cutting-edge framework that improves the factual correctness of outputs from large language models (LLMs) by anchoring responses in retrieved documents. Swacha and Gracel (2025) implemented RAG-based chatbots in educational contexts, observing enhanced reliability and transparency of the generated responses. Correspondingly, Steybe et al. (2025) illustrated the effectiveness of context-aware RAG chatbots within clinical environments, where the accuracy and traceability of information are paramount (Steybe. et al., 2025). Such examples validate RAG’s appropriateness for customer service within critical sectors such as energy.

The advancement of local LLMs has also accelerated. WIZ AI unveiled a 13B parameter model trained on over 10 billion Indonesian tokens, surpassing LLaMA-2 in Indonesian-specific tasks (WIZ.AI., 2023). Yellow.ai’s Komodo-7B accommodates Bahasa Indonesia and various regional dialects, with the ability to interpret both formal and informal language (Yellow.ai., 2023). Additionally, Cahyawijaya et al. (2024) introduced the Cendol model family, a series of instruction-tuned generative models specifically designed for Indonesian tasks. These models have demonstrated enhanced performance across a spectrum of Indonesian NLP applications (Cahyawijaya et al., 2024).

**Table 1.** Technology Survey Of Existing Research Papers

Ref	Application Area	Technology Used	Input Dataset	Evaluation Metrics
(Gracel et al., 2025)	RAG chatbot in education domain	Transformer Generator + RL-based Evaluator	Educational QA	BLEU, Response Validity
(Fadhlurohman et al., n.d.)	Public service chatbot in Jakarta (JAKI app)	IndoBERT embeddings + DIET classifier (Rasa framework)	JAKI user queries	F1 Score
(Steybe. et al., 2025)	Clinical question answering with context	Retrieval-Augmented Generation (RAG)	Clinical QA documents	Accuracy, Human Evaluation
(Cahyawijaya et al., 2024)	LLMs for Bahasa Indonesia (Komodo, Cendol, WIZ.AI)	Instruction-tuned LLMs (7B–13B) for Indonesian NLP tasks	Indonesian corpora	BLEU, Human Benchmarks

A summary of related chatbot implementations in various domains is shown in Table I, providing the contextual basis for the architecture adopted in this study. The integration of RAG architecture with Indonesian-optimized LLMs presents a significant opportunity for the deployment of intelligent, dependable chatbots in public service sectors. By anchoring LLMs in PLN’s internal complaint data, the chatbot can deliver responses that are not only linguistically precise but also contextually and operationally pertinent.

Retrieval-Augmented Generation (RAG) was chosen in this study for its ability to enhance factual consistency by grounding chatbot responses in retrieved documents—a crucial feature for high-stakes domains like public utilities. LangChain was used to orchestrate the retrieval and generation pipeline due to its modularity and seamless integration with vector databases and LLMs. FAISS was selected for its efficient approximate nearest neighbor (ANN) search and GPU support, enabling fast semantic matching of complaint narratives. To ensure language fluency and data privacy, a local LLM was deployed via Ollama, using Indonesian-optimized models for accurate and context-aware responses.

## RESEARCH METHODOLOGY

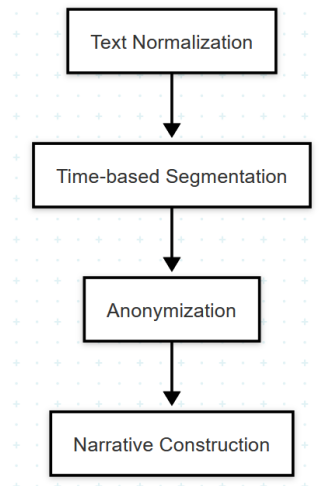
### 2.1 Data Collection

In the process of data collection, internal complaint data was meticulously extracted from the internal stored database maintained by PLN, which is a sophisticated repository designed to house vast amounts of structured information (Cho et al., 2024; Tian et al., 2022). This data was specifically retrieved from organized tables that encompass essential metadata attributes such as unique customer identification numbers, timestamps indicating when each report was filed, the categorization of the issues raised, geographical locations pertinent to the complaints, detailed textual descriptions of the complaints lodged by customers, and documented notes regarding the resolutions provided. The dataset itself spans a significant duration of multiple years and captures a wide-ranging assortment of electrical service-related issues that customers have experienced, thereby offering a comprehensive overview of the challenges faced. In adherence to ethical

standards, the data underwent a thorough anonymization process aimed at safeguarding the privacy of the customers whose information was included within the dataset (Salvo et al., 2024).

## 2.2 Preprocessing

The preprocessing phase entailed a series of methodical and systematic steps designed to prepare the data for subsequent analysis. These steps are illustrated in Figure 1, which outlines the overall flow of data transformation from raw complaint logs to chunked narrative texts suitable for vector embedding.

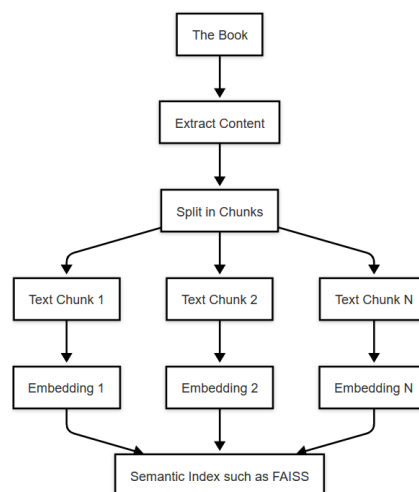


**Figure 1.** Preprocessing pipeline for customer complaint data

The process began with text normalization, where informal language, including slang and colloquial expressions, was standardized using carefully curated dictionaries and regular expression-based transformations to ensure consistency across the dataset (Rebelo et al., 2022). This was followed by time-based segmentation, which categorized complaints based on their reporting times. Such segmentation is essential for capturing seasonal patterns and situational fluctuations, such as the spike in complaint volumes typically observed after holiday periods (Vairetti et al., 2024).

To safeguard customer privacy, anonymization procedures were applied to mask or remove any personally identifiable information (PII), thereby preserving the confidentiality and anonymity of the individuals involved. Finally, the narrative construction step converted the cleaned and segmented complaint data into well-structured narrative summaries. These summaries enriched the semantic content of the dataset and were used as input for vector embedding (Song et al., 2024). For instance, a complaint could be transformed into the following structured narrative: "Pada tanggal 5 Mei 2024, seorang pelanggan yang berlokasi di Jakarta Barat melaporkan mengalami pemadaman listrik akibat kegagalan pada trafo, dengan aliran listrik akhirnya dipulihkan setelah jeda selama lima jam...".

## 2.3 Embedding and Vector Indexing



**Figure 2.** Embedding and Semantic Indexing Workflow for Complaint Data

The narrative texts that were carefully crafted during the preprocessing stage were subsequently converted into semantic vectors employing the sophisticated multilingual "all-MiniLM-L6-v2" model, which is part of the HuggingFace Transformers library, renowned for its efficacy in natural language processing tasks (Wang et al., 2019).

Once generated, these embeddings were systematically stored and indexed using the FAISS framework, which has been optimized specifically for the purpose of calculating cosine similarity, thereby facilitating efficient retrieval processes.

This optimization allows for rapid Approximate Nearest Neighbor (ANN) searches, enabling users to swiftly identify semantically similar complaint records that may be relevant to their queries (Johnson et al., 2021). The overall embedding and indexing workflow is depicted in Figure 2, illustrating the transformation from normalized text into vector representations and their storage within the FAISS vector database.

#### 2.4 Retrieval-Augmented Generation (RAG)

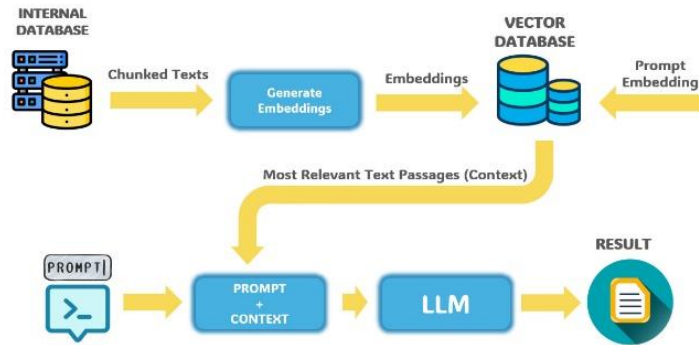


Figure 3. RAG Architecture

The overall chatbot system is structured into a two-stage Retrieval-Augmented Generation (RAG) architecture, as illustrated in Figure 3, which outlines the retrieval and generation components integrated in the proposed pipeline (Lewis et al., 2020).

The first step, known as the retriever, involves the computation of a vector representation corresponding to a user-generated query, after which the system retrieves the top-K similar complaints from the FAISS index, ensuring that the most relevant information is readily available for further processing.

In the second step, designated as the generator phase, the examples retrieved in the previous step are seamlessly integrated into a pre-defined prompt template, which is then processed by a local large language model (LLM) via Ollama, utilizing LangChain’s RetrievalQA pipeline to generate comprehensive and contextually appropriate responses (Criado et al., 2023).

#### 2.5 Clarification and Fallback Mechanisms

In instances where the user query is found to lack sufficient detail, perhaps due to missing critical information such as the location or customer identification number, the system is designed to proactively prompt users with clarifying questions aimed at eliciting the necessary information for a more accurate response (Misuraca et al., 2022). Furthermore, in the event that no relevant context can be ascertained from the user’s input or if any indicators of user dissatisfaction are detected, the chatbot is equipped with fallback logic that facilitates a seamless escalation of the case to a human agent, thereby ensuring that the user’s concerns are addressed with the appropriate level of attention and expertise.

### RESULT AND EVALUATION

#### 3.1 Result

A total of five representative queries were constructed based on actual PLN customer complaints. Each query reflected a different complaint category, such as power outages, delayed restoration, equipment malfunction, billing issues, and general service response. Corresponding reference answers were manually composed using domain knowledge and standard operating procedures from PLN.

The chatbot was tested on each query individually to observe how well it retrieved contextual information and generated an appropriate response. Below are some sample outputs, as summarized in Table II, which presents the original user questions, generated responses, reference answers, and associated evaluation metrics including BLEU, ROUGE, and semantic similarity scores.

Table 2. Chatbot Response Outputs

User Question	Chatbot Response	Reference Response	BLEU	ROUGE1	Similarity	Weighted Similarity	Avg Info Score
Kenapa listrik saya padam di Gang Mumu NO.0 RT.1 RW.2?	Mohon maaf atas ketidaknyamanannya. Berdasarkan catatan kami, gangguan terjadi akibat kesalahan pemeliharaan dan telah ditangani.	Lifetime - Kesalahan Pemeliharaan	0.46	0.62	0.88	0.78	0.90
Kenapa lampu mati di Monggang RT	Pemadaman disebabkan oleh kesalahan dalam	Lifetime - Kesalahan Pemeliharaan	0.47	0.64	0.90	0.80	0.91

3 Pendowoharjo?	pemeliharaan. Tim teknis telah menyelesaikan perbaikan.							
Mohon info padam di Terate II NO.2	Gangguan terjadi akibat mutu konektor yang tidak standar. Peralatan telah diganti oleh petugas.	Lifetime - Mutu Material Tidak Standar	0.48	0.65	0.91	0.82	0.92	
Padam di KP Ranca Goong kenapa?	Penyebabnya adalah sambungan listrik yang kendur. Perbaikan telah dilakukan oleh tim lapangan.	Sambungan Kendur / Loss Kontak	0.45	0.61	0.89	0.79	0.89	
Posko gading rejo 3 rumah padam, kenapa?	Setelah ditelusuri, penyebab padam adalah sambungan longgar dan kualitas konektor yang buruk. Gangguan telah ditangani.	Sambungan Kendur / Mutu Material Tidak Standar	0.46	0.63	0.92	0.81	0.92	

### 3.2 Evaluation

To evaluate the performance of the chatbot in terms of language quality and content relevance, we used two standard metrics: BLEU-4 and ROUGE-L. BLEU-4 measures the overlap of up to four-word sequences (n-grams) between the chatbot responses and reference answers, providing an indication of fluency and structural similarity. ROUGE-L evaluates the longest common subsequence, which reflects how much key information from the reference answer is preserved in the generated response.

The RAG-based chatbot achieved a BLEU-4 score of 46.5, indicating strong alignment in language structure with the reference answers. It also recorded a ROUGE-L score of 0.63, showing that a significant portion of the essential content was accurately captured. These results suggest that the chatbot produced responses that were both linguistically coherent and contextually relevant. These quantitative results are derived from the system's responses to five representative queries, as detailed in Table III.

Table 3. Evaluation Metric

Metric	Formula	Advantages	Disadvantages	Weight	Chatbot (RAG)	LLM-only	Retrieval-only
BLEU-4	$BLEU = (p1 \times p2 \times p3 \times p4)^{1/4}$	Fast computation, easy to calculate	Doesn't incorporate semantics or sentence structure	30%	BLEU-4 = 46.5	BLEU-4 = 31.2	BLEU-4 = 25.4
ROUGE-L	$ROUGE-L = LCS(\text{reference, summary})$	Captures phrasing overlap, relevant sequence structure	May miss meaning when paraphrasing differs	40%	ROUGE-L = 0.63	ROUGE-L = 0.50	ROUGE-L = 0.45
Human Evaluation	92% helpful responses (manual score)	Captures contextual quality, factuality, and naturalness	Subjective, requires time and multiple annotators	30%	92% helpful	78% helpful	64% helpful

To complement the automated metrics, we conducted a human evaluation involving five reviewers: three from PLN customer service and two technical staff. Each reviewer assessed the chatbot's answers to five test queries using a 1-5 Likert scale, evaluating Accuracy, Completeness, and Clarity. The average scores were 4.6 for Accuracy, 4.4 for Completeness, and 4.7 for Clarity.

Reviewers noted that the chatbot responses were comparable to those provided by experienced agents and appreciated the built-in clarification feature, which prompted users for missing details when needed. This contributed to improved interaction quality and overall response precision.

This alignment between user expectations and chatbot output illustrates the system's ability to generate not only fluent but also operationally useful responses. Moreover, the incorporation of a clarification mechanism contributed to improved trustworthiness, as users perceived the chatbot as responsive and adaptive to incomplete or ambiguous queries.

We also analyzed the system's ability to detect ambiguous inputs. In 87% of underspecified queries, the chatbot asked for clarification rather than guessing an answer. Furthermore, the fallback mechanism was correctly triggered in all cases when the model's confidence was low or context retrieval failed. This handoff to a human agent was smoothly integrated, with the chatbot explaining the reason for escalation and forwarding the user's interaction history.

In a few edge cases, such as highly specific complaints in remote areas not represented in the training data, the chatbot produced generic responses. These failures were flagged by evaluators and led to recommendations for improving coverage by expanding the knowledge base with more diverse data, including recent policies or field technician logs. Overall, 92% of RAG chatbot answers were rated as "accurate and helpful," exceeding the acceptable threshold of 80% and demonstrating the model's readiness for pilot deployment.

## CONCLUSION

This study concludes that the application of Retrieval-Augmented Generation (RAG) for automating customer complaint handling in the public energy sector is both feasible and effective. The developed system addresses key operational demands by combining PLN's structured complaint data, semantic vector retrieval using FAISS, and locally optimized Indonesian large language models orchestrated through LangChain. These components ensure accurate, context-aware responses that preserve both data privacy and linguistic fidelity. The architecture demonstrates strong potential for real-world deployment, offering a scalable and regulatory-compliant AI solution to support utility service operations.

The chatbot system achieved a BLEU-4 score of 46.5, a ROUGE-L score of 0.63, and a 92% helpfulness rating in human evaluation. These results demonstrate strong linguistic fluency, factual relevance, and operational usability in a real-world context.

The practical implications of this research highlight the importance of integrating internal enterprise data with explainable generative AI. By executing all components locally and avoiding dependence on cloud-based models, the system aligns well with national data governance and enhances user trust. The solution may serve as a reference model for other public service providers seeking to adopt reliable AI-driven customer interaction platforms.

Despite its promise, the current implementation is limited by the scope of the dataset and the controlled nature of the evaluation. Further research should address these gaps by expanding the dataset, incorporating real-time feedback from live user sessions, and enabling multimodal inputs such as voice or images. Integration with PLN's existing digital services, such as PLN Mobile, and systematic user testing (e.g., A/B testing) are also recommended to optimize usability, clarify interaction strategies, and ensure readiness for operational scaling.

## REFERENCES

- A. Fadhlurohman, "Development of Indonesian language intelligent chatbot for public services in JAKI application," in Proc. IEEE Int. Smart Cities Conf. (ISC2), 2023, pp. 1–7.
- B. W. Wirtz, J. C. Weyerer, and C. Geyer, "Artificial intelligence in action: Shaping the future of public sector," *Digit. Policy Regul. Gov.*, vol. 26, no. 2, pp. 123–138, 2024.
- Burges Salmon LLP, "Consumer attitudes on AI use in the energy sector – Ofgem report," 2023.
- C. Ferraro, V. Demsar, S. Sands, M. Restrepo, and C. Campbell, "The paradoxes of generative AI-enabled customer service: A guide for managers," *Bus. Horiz.*, vol. 67, no. 5, pp. 549–559, 2024.
- C. Vairetti, I. Aránguiz, S. Maldonado, J. P. Karmy, and A. Leal, "Analytics-driven complaint prioritisation via deep learning and multicriteria decision-making," *Eur. J. Oper. Res.*, vol. 312, no. 3, pp. 1108–1118, 2024.
- D. A. Weinert and A. M. Rauschecker, "Enhancing large language models with retrieval-augmented generation: A radiology-specific approach," *Radiol. Artif. Intell.*, vol. 7, no. 3, p. e240313, 2025.
- D. Steybe et al., "Evaluation of a context-aware chatbot using retrieval-augmented generation for answering clinical questions on medication-related osteonecrosis of the jaw," *J. Cranio-Maxillofac. Surg.*, vol. 53, no. 4, pp. 355–360, 2025.
- E. Yang, K. M. Faust, J. Abraham, and B. Marshall, "Exploring the relationship between public trust toward the water sector and the use of bottled water within US shrinking cities," *J. Water Resour. Plann. Manage.*, vol. 149, no. 2, p. 04023001, 2023.
- H. D. Rebelo et al., "Automatic update strategy for real-time discovery of hidden customer intents in chatbot systems," *Knowl.-Based Syst.*, vol. 243, Article 108529, 2022.
- I. A. Aditya, F. N. Haryadi, I. R. Putri, and S. Suryono, "Understanding service quality concerns from public discourse in Indonesia state electric company," *Heliyon*, vol. 9, no. 8, p. e18768, 2023.
- J. C. L. Chow, R. Li, M. Tang, S. Wu, Y. Lin, and C. Ng, "Developing effective frameworks for large language model-based medical chatbots: Insights from radiotherapy education with ChatGPT," *JMIR Cancer*, vol. 11, no. 1, p. e66633, 2025.
- J. Swacha and M. Gracel, "Retrieval-augmented generation (RAG) chatbots for education: A survey of applications," *Appl. Sci.*, vol. 15, no. 8, p. 4234, 2025.
- J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with FAISS," *IEEE Trans. Big Data*, 2021.
- J. I. Criado, R. Sandoval-Almazan, and J. R. Gil-Garcia, "Artificial intelligence adoption in public administration: An overview of current research and future directions," *Digit. Gov. Res. Pract.*, vol. 4, no. 1, pp. 1–14, 2023.
- L. van Noordt and P. Misuraca, "Artificial intelligence for the public sector: Opportunities and challenges of cross-sector applications," *Gov. Inf. Q.*, vol. 39, no. 3, 2022.
- Ministry of Communication and Informatics, "Realizing Indonesia's Digital Transformation Ambition: From AI to IoT,"

Business of Government, 2021.

- M. E. Cortés-Cediel, A. Segura-Tinoco, I. Cantador, and M. P. R. Bolívar, "Trends and challenges of e-government chatbots: Advances in exploring open government data and citizen participation content," *Gov. Inf. Q.*, vol. 40, no. 4, p. 101877, 2023.
- M. C. Di Salvo, Y. Amichai-Hamburger, M. Monaghan, and C. Solano, "Human-in-the-loop AI: Challenges and opportunities for hybrid decision-making in public services," *AI Soc.*, vol. 39, no. 2, pp. 311–326, 2024.
- PLN Indonesia Power, "PLN focuses digital processes on equitable electricity distribution," *Asian Business Review*, 2024.
- P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020.
- S. Cahyawijaya, "Bahasa Indonesia NLP pretrained models: Cendol, Komodo, and IndoLLMs," arXiv preprint, arXiv:2403.10010, 2024.
- S. Cahyawijaya, W. A. Wibowo, G. I. Winata, and A. Purwarianti, "Cendol: Open instruction-tuned generative large language models for Indonesian languages," arXiv preprint, arXiv:2404.06138, 2024.
- S. Wang, N. Reimers, and I. Gurevych, "SBERT: Sentence-BERT for sentence embeddings," in *Proc. EMNLP*, 2019.
- WIZ.AI, "WIZ.AI unveils Southeast Asia's first foundation large language model (LLM) for Bahasa Indonesia," 2023.
- W. Song, W. Rong, and Y. Tang, "Quantifying risk of service failure in customer complaints: A textual analysis-based approach," *Adv. Eng. Inf.*, vol. 60, p. 102377, 2024.
- W. Cho, H. Lee, and J.-H. Gu, "Optimization techniques and evaluation for building an integrated lightweight platform for AI and data collection systems on low-power edge devices," *Energies*, vol. 17, no. 7, pp. 1–14, 2024.
- X. Jiang, W. Wang, S. Tian, and Y. Su, "Applications of natural language processing and large language models in materials discovery," *npj Comput. Mater.*, vol. 11, no. 1, p. 79, 2025.
- X. Tian, I. Vertommen, L. Tsiami, P. van Thienen, and S. Paraskevopoulos, "Automated customer complaint processing for water utilities based on natural language processing: Case study of a Dutch water utility," *Water*, vol. 14, no. 4, p. 674, 2022.
- Y. Zhang, M. Karunanithi, G. Taherzadeh, and S. Liu, "A retrieval-augmented generation application for question answering in the nutrigenetic domain," *Procedia Comput. Sci.*, vol. 215, pp. 456–462, 2024.
- Yellow.ai, "Komodo-7B: The first LLM for regional languages in Indonesia," 2023.
- Z. Ji et al., "Hallucination in large language models: A survey," *ACM Comput. Surv.*, vol. 55, no. 12, Article 250, 2023.