Technologia Journal: Jurnal Informatika
E-ISSN : 3046-9163
Vol.2.No.4, November 2025
DOI :
https://doi.org/10.62872/xe3zrt53

Application of Large Language Models (LLMs) for Optimising Indonesian Language-Based Public Service Chatbots

Ali Ibrahim[1]□, Sitti Rachmaeati Yahya[2], Iwan Adhicandra[3]
Universitas Negeri Makassar[1], Universitas Siber Asia (UNSIA)[2], Universitas Bakrie[3]
e-mail: * aliibrahimok@gmail.com, sitti.rachma@gmail.com, iwan.adhicandra@bakrie.ac.id

Inputed : October 21, 2025 *Revised : October 15, 2025*
Accepted : November 20, 2025 *Published : November 29, 2025*

ABSTRACT

This study examines the potential of Large Language Models to optimise Indonesian language-based public service chatbots by integrating linguistic, technological, and administrative perspectives. Using a mixed-method approach that combines a systematic literature review with secondary benchmarking of state-of-the-art LLMs, the research evaluates model performance in Indonesian semantic comprehension, contextual reasoning, and domain adaptability. The findings show that LLMs can significantly improve chatbot accuracy, inclusivity, and responsiveness, outperforming rule-based systems that struggle with informal expressions, multi-intent queries, and policy-specific terminology. Benchmarking highlights that GPT-4 and PaLM-2 achieve high contextual coherence and low hallucination rates, while Indonesian-centric models such as IndoGPT demonstrate strong local language adaptability. However, risks related to data privacy, bias, hallucination, and governance limitations present substantial challenges for implementation. The study proposes a strategic framework that emphasizes AI governance, interoperable data infrastructure, institutional capacity building, hybrid retrieval–generation design, and citizen engagement to ensure responsible adoption. Overall, the integration of LLM-powered chatbots has the potential to transform Indonesia's digital public service landscape, provided that deployment is accompanied by robust oversight, ethical safeguards, and sustainable technological planning.

Keywords: *artificial intelligence, chatbots, Indonesian language, large language models, public service.*

INTRODUCTIONS

Digital transformation in public administration has accelerated rapidly over the past decade, driven by growing societal expectations for efficient, responsive, and accessible public services. Governments worldwide have increasingly adopted artificial intelligence technologies to support citizen engagement, automate administrative processes, and enhance service delivery.

A significant technological shift has emerged with the development of Large Language Models, which demonstrate advanced capabilities in natural language understanding, contextual reasoning, and dialogue generation (Zhu et al., 2024). Public service institutions in countries such as Singapore, South Korea, and the United States have integrated LLM-powered chatbots to streamline information dissemination, reduce administrative burden, and address service delays. These advancements illustrate that LLMs are redefining digital governance by enabling more human-like interaction and personalized support for citizens seeking government information or assistance.

In Indonesia, the government has invested in extensive digitalization initiatives through programs such as SPBE (Sistem Pemerintahan Berbasis Elektronik). Despite this progress, many public service chatbots remain rule-based, poorly contextualized, and limited in their ability to process natural Indonesian language variation. Studies show that Indonesian government chatbots often struggle with idiomatic expressions, regional dialects, paraphrased questions, and multi-turn conversations (Ma'rup, 2024). These limitations create barriers for citizens, particularly those with lower digital literacy or those unfamiliar with formal bureaucratic terminology. As a result, public service chatbots frequently fail to deliver accurate or timely responses, leading to user frustration and reduced trust in digital public services.

Globally, LLMs such as GPT, PaLM, LLaMA, and BLOOM have demonstrated exceptional capabilities in multilingual language processing, including low-resource languages. However, Indonesian presents unique linguistic challenges, including agglutination, code-switching with English, morphological complexity, and nuanced socio-cultural expressions. Early evaluations indicate that LLMs trained with multilingual corpora perform reasonably well in Indonesian, but performance varies significantly depending on dialect, situational context, and public-sector terminology (Cahyawijaya et al., 2021). This nuance suggests that while LLMs have enormous potential, their application to Indonesian public service settings requires careful optimization to ensure contextually appropriate and culturally aligned responses.

From a public administration perspective, the effectiveness of digital services depends on the extent to which technological systems can provide clarity, accuracy, and equitable access. Indonesian citizens frequently seek information related to identity cards, healthcare insurance registration, taxation, social welfare assistance, and education services. These domains involve complex policies, evolving regulations, and high contextual dependency. Rule-based chatbots are unable to process these complexities because they rely on fixed patterns and cannot dynamically interpret citizen queries. Consequently, service delivery inefficiencies persist, particularly in times of crisis. During the COVID-19 pandemic, for example, Indonesian government digital platforms experienced overwhelming demand, revealing the inadequacy of existing chatbot systems in handling large-scale, multi-intent inquiries (Wagola et al., 2023). LLMs could significantly mitigate such issues by

providing adaptive responses capable of interpreting diverse questions and generating accurate information based on updated datasets.

Another challenge concerns inclusivity. Indonesia is characterized by linguistic diversity, with more than 700 local languages influencing daily communication patterns. Even when users speak Indonesian, differences in vocabulary, syntax, and pragmatics emerge across regions. Public service chatbots must therefore interpret queries expressed in varied styles, such as informal Indonesian, mixed-language expressions, or localized vocabulary. LLMs equipped with large multilingual corpora and fine-tuning datasets can better handle such linguistic variation by capturing deep semantic structures rather than relying on rigid rules. Studies show that conversational AI systems supported by LLMs perform more effectively in multilingual and informal language environments, enhancing accessibility for populations with diverse linguistic backgrounds (Siddiqui et al., 2023). This makes LLMs particularly valuable for countries with high linguistic heterogeneity such as Indonesia.

Despite these advantages, the adoption of LLMs in Indonesian public services presents several risks and barriers. Concerns include data privacy, model bias, hallucination, and the ethical implications of using AI-generated content in official communications. Data privacy is especially critical because public service chatbots frequently collect citizen information linked to administrative records. Without strong safeguards and transparent data governance frameworks, LLM integration may expose the government to cybersecurity risks or undermine citizen trust. Model hallucination, where an LLM produces inaccurate or fabricated responses, also poses a serious challenge for public-sector use because misinformation in areas such as taxation, legal documentation, or health services can have substantial consequences. Studies warn that LLMs require strict validation pipelines to ensure reliability in high-stakes environments (Waldo & Boussard, 2024).

Another concern relates to institutional readiness. Many Indonesian government agencies lack the technical expertise, computational resources, and data infrastructure needed to implement advanced AI models. LLM deployment requires high-performance computing, large-scale cloud environments, and secure APIs, which are not yet widely adopted in public institutions. Furthermore, bureaucratic resistance to technological change may slow adoption. Public servants may be reluctant to integrate LLM-assisted tools due to perceived threats to job roles or unfamiliarity with AI processes. Addressing these institutional challenges requires capacity building, regulatory frameworks, and collaborative partnerships between government, academia, and the private sector.

Existing research on Indonesian public service chatbots has largely focused on usability, interface design, and rule-based performance evaluation. For instance, Nadzif & Soelistijadi (2024) examined citizen satisfaction with municipal chatbots but did not analyze AI-enhanced models. Meanwhile, Keith (2024) assessed the limitations of rule-based systems but did not evaluate LLM

integration. Other studies like Judjianto & Vandika explored Indonesian NLP models but focused on linguistic performance rather than governance applications. These studies reveal a research gap in understanding how LLMs can specifically optimize Indonesian public service chatbots from both technological and administrative perspectives.

This article provides novelty by offering an integrated examination of how LLMs can improve semantic understanding, response quality, inclusivity, and operational efficiency in Indonesian-language public service chatbots. Additionally, this research introduces a mixed-method analytical framework combining academic literature, secondary benchmarking of major LLM architectures, and Indonesian public sector case considerations. Unlike previous works that evaluate isolated technical or service-delivery aspects, this study bridges AI capabilities with public administration demands and Indonesian linguistic contexts. Therefore, the objective of this study is to critically analyze the potential and challenges of LLM application for Indonesian-language public service chatbots, evaluate structural and linguistic constraints, and propose strategic recommendations to optimize government adoption.

METHODOLOGY

This study adopts a mixed-method research design combining a systematic literature review with secondary benchmarking analysis of Large Language Models relevant to Indonesian language processing. The literature review systematically identifies peer-reviewed journal articles published between 2018 and 2024 concerning LLMs, natural language processing for low-resource languages, Indonesian-language computational linguistics, and public sector chatbot deployments. Databases such as Scopus, IEEE Xplore, and Google Scholar were searched using keywords including Indonesian NLP, public service chatbots, large language models, and multilingual transformers. Articles were selected based on methodological rigor, relevance to public administration, and technological depth. This approach ensures comprehensive coverage of theoretical advances while grounding the study in empirical findings across multiple technological and administrative domains (Siddiqui et al., 2023).

The second methodological component involves secondary benchmarking analysis drawing on existing research datasets that evaluate LLM performance in Indonesian or multilingual contexts. These benchmarking sources include comparative studies on GPT variants, PaLM, BLOOM, IndoBERT, and local Indonesian transformer models such as INDOLLM and IndoGPT. Relevant performance metrics such as perplexity, accuracy, multilingual understanding, and contextual coherence were examined from published evaluation datasets in reputable NLP journals and AI conference proceedings. This comparative approach allows the study to assess the strengths and limitations of current LLM architectures in processing Indonesian language structures and public service terminology without conducting

original model training. By synthesizing secondary benchmarking results, the study can objectively assess which architectural features contribute to improved chatbot responsiveness and accuracy in Indonesian public sector contexts (Zhu et al., 2024).

The mixed-method integration is achieved by triangulating insights from the literature review and benchmarking analysis with Indonesian public service needs identified through policy documents and previous e-government studies. Policy documents from the Indonesian Ministry of Communication and Informatics, SPBE evaluation reports, and digital transformation studies were incorporated to contextualize LLM applicability within nationwide administrative reforms. Triangulation enhances analytical robustness by ensuring that technical findings correspond with practical public service realities. This integrative method provides a comprehensive foundation for formulating recommendations that align linguistic capability, technological feasibility, and governance priorities in optimizing Indonesian language-based public service chatbots (Wagola et al., 2023).

RESULTS AND DISCUSSION

The Transformative Potential of LLMs for Indonesian Public Service Chatbots

The emergence of Large Language Models has fundamentally reshaped natural language processing, enabling machines to engage in more human-like conversation and interpret increasingly complex linguistic structures. For Indonesian public service environments, the integration of LLMs represents an opportunity to address longstanding technological constraints associated with rule-based and retrieval-based chatbot systems. Traditional chatbots deployed in various Indonesian government platforms are limited by static pattern-matching rules, narrow vocabulary coverage, and poor contextual reasoning. These limitations prevent them from responding accurately to paraphrased questions, handling multi-turn interactions, or generating explanations based on conversational history. LLMs, by contrast, are trained on massive corpora containing multilingual textual data, allowing them to build deep semantic representations that facilitate more natural and contextually grounded dialogue (Zhu et al., 2024). This capability is crucial for public administration, where citizen queries often include implicit meaning, narrative descriptions, or informal phrasing.

One of the most transformative aspects of LLMs lies in their ability to interpret Indonesian language variation. Indonesian is characterized by diverse registers, including formal bureaucratic language, conversational everyday speech, and hybrid expressions incorporating English or regional vocabulary. Public service users often express questions using informal Indonesian or regionally influenced patterns that rule-based systems fail to recognize. LLMs trained on multilingual or Indonesian-enriched corpora have demonstrated improved performance in processing such variation, enabling chatbots to interpret a wider range of citizen expressions with greater accuracy. Empirical

evaluations of IndoGPT and multilingual GPT models show that they achieve strong performance in tasks involving paraphrase detection, sentiment classification, and question answering in Indonesian contexts (Cahyawijaya et al., 2021). These linguistic capabilities expand accessibility, ensuring that citizens with diverse language backgrounds can engage with digital public services without needing to conform to rigid administrative terminology.

Beyond linguistic comprehension, LLMs significantly enhance the contextual reasoning required for high-quality public service interaction. Many government processes involve multi-step instructions, eligibility requirements, or conditional pathways. Citizens often ask questions that combine multiple intents, such as "How do I register for BPJS if I just moved cities and don't have a new ID card yet?" Rule-based models typically fail to process such compound queries, resulting in irrelevant or fragmented responses. LLMs, however, are capable of decomposing multi-intent questions, identifying the user's underlying needs, and generating cohesive explanations that reflect regulatory complexities. This capacity aligns with research demonstrating the effectiveness of LLMs in multi-turn task-oriented dialogue, where contextual coherence and memory are crucial for user satisfaction (Siddiqui et al., 2023). Improved reasoning enables LLM-powered chatbots to become more reliable sources of public administrative guidance, reducing dependence on human officers and shortening service queues.

Another important dimension of LLM integration relates to domain adaptation. Indonesian public service institutions manage diverse policy domains including healthcare, identity administration, social welfare, taxation, education, and transportation. Each domain contains specialized terminology that must be understood accurately to prevent misinformation. LLMs can be fine-tuned or instruction-tuned using domain-specific datasets, enabling chatbots to interpret highly technical language while maintaining conversational naturalness. For example, fine-tuning an LLM on BPJS health insurance regulations or passport application guidelines would enable the model to provide more detailed and authoritative responses. Recent research in multilingual domain adaptation confirms that LLMs can integrate new domain knowledge through parameter-efficient tuning without degrading overall language performance (Waldo & Boussard, 2024). This adaptability makes LLMs particularly advantageous for Indonesian public agencies whose regulatory content frequently changes due to policy reforms.

In addition to enhancing accuracy and contextuality, LLMs have the potential to increase efficiency within public sector operations. Chatbots powered by LLMs can manage large volumes of repetitive citizen inquiries, enabling human officers to focus on more complex or sensitive cases. During peak administrative periods, such as national identity registration deadlines, school enrollment seasons, or social assistance campaigns, LLM chatbots can absorb demand surges by providing immediate responses, reducing waiting times and administrative bottlenecks. Studies in digital governance have shown that AI-driven automation can reduce operational workloads by up to 40

percent and increase service speed, particularly in high-volume environments (Keith, 2024). These efficiencies align with Indonesia's broader SPBE goals of improving institutional performance and digital service readiness.

Another transformative potential lies in improving inclusivity and democratizing access to information. Many citizens, particularly those in rural areas or with lower socioeconomic status, face barriers to understanding bureaucratic procedures due to limited digital literacy or unfamiliarity with formal language. LLMs can mitigate these barriers by generating simpler explanations, clarifying bureaucratic terminology, and adjusting linguistic register based on user input. This aligns with research on adaptive language modeling, which shows that LLMs can adjust complexity levels in real-time to match user comprehension needs (Judjianto & Vandika (2025)). Through adaptive responses, LLM-powered chatbots can become tools not only for service access but also for public education, supporting citizen empowerment and governance transparency.

Despite these potentials, the transformative nature of LLMs is accompanied by significant ethical and technical concerns. A primary risk is model hallucination, where an LLM generates responses that appear coherent but are factually incorrect. In public service contexts, hallucination poses serious risks because misinformation related to legal requirements, administrative deadlines, or social assistance eligibility can cause harm. Studies indicate that hallucination rates increase when models are prompted with ambiguous or domain-specific queries, which are common in public service interactions (Siddiqui et al., 2023). To mitigate this risk, LLMs must be integrated within controlled architectures that combine retrieval of verified government documents with generative reasoning.

Bias is another critical concern. LLMs may reproduce harmful stereotypes or unintended policy interpretations embedded in training data. This is particularly relevant to Indonesia's socio-cultural diversity, where biases based on region, ethnicity, gender, or socioeconomic status may surface in subtle linguistic patterns. Research in fairness-aware NLP emphasizes the need for bias detection pipelines and dataset curation to prevent discrimination in AI-assisted services (Zhu et al., 2024). Bias mitigation is essential to ensure that LLM-powered chatbots uphold principles of equality and neutrality expected from public institutions.

Finally, LLMs must be aligned with data governance and cybersecurity standards. Public service chatbots often process sensitive information, such as personal identification numbers or household data. Therefore, LLM adoption requires secure architectures capable of preventing unauthorized access. Policy frameworks must enforce data minimization, encryption, and auditability protocols to ensure compliance with national data protection regulations. Studies in AI governance highlight that trust in digital public services is contingent on transparency and accountability mechanisms that regulate AI-generated content (Wagola et al., 2023). Without robust governance, citizens may distrust LLM-powered services, undermining the intended benefits.

Overall, the transformative potential of LLMs for Indonesian public service chatbots is substantial. Their linguistic flexibility, contextual reasoning, domain adaptability, and efficiency gains present opportunities to improve service quality and inclusivity. Nevertheless, such transformative potential must be accompanied by responsible governance, robust supervision, and careful system design to ensure that LLMs serve public interest in a safe, ethical, and sustainable manner.

Benchmarking LLM Capabilities for Indonesian Public Service Chatbots

The application of Large Language Models in Indonesian public service chatbots requires a clear understanding of how these models perform across linguistic, functional, and contextual dimensions. Benchmarking analysis is therefore essential to evaluate whether current LLM architectures can realistically meet the demands of Indonesian e-government services. Public sector communication relies heavily on accuracy, regulatory consistency, and the ability to interpret procedural requirements. Existing studies comparing multilingual language models provide valuable insights into performance gaps and strengths that influence chatbot reliability. For instance, multilingual models such as GPT-4, PaLM-2, and BLOOM exhibit strong accuracy in Indonesian tasks, yet they vary significantly in terms of contextual coherence and domain-specific reasoning (Zhu et al., 2024). By analyzing secondary benchmarking datasets, this discussion assesses the suitability of LLMs for Indonesian public service deployment, focusing on semantic understanding, conversational coherence, classification accuracy, and ability to follow policy-based instructions.

A central component of benchmarking is semantic comprehension, which determines whether an LLM can interpret queries written in diverse linguistic forms. Indonesian public service questions frequently involve informal phrasing, colloquial expressions, and mixed-language structures influenced by English or regional dialects. Studies using Indonesian NLP benchmarks such as IndoNLI, IndoQA, and MT-Bench ID demonstrate that LLMs differ considerably in how effectively they interpret semantic nuance. GPT-4 has consistently achieved high accuracy in intent recognition and paraphrase detection, whereas models such as BLOOM and LLaMA-2 require domain adaptation to achieve similar performance (Cahyawijaya et al., 2021). The ability to generalize across Indonesian language variations is therefore uneven across models, suggesting that model selection must be tailored to the linguistic diversity of Indonesian citizens.

Contextual reasoning represents another key dimension in evaluating LLM suitability for public service chatbots. In Indonesian administrative processes, queries are often multi-intent, requiring models to infer procedural dependencies or regulatory exceptions. For example, citizenship administrative questions may involve complex conditional elements related to document validity or residency status. Benchmarking studies show that GPT-4 and PaLM-2 outperform other models in tasks requiring multi-step reasoning, due to their

large-scale training and sophisticated instruction-tuning pipelines (Siddiqui et al., 2023). Meanwhile, smaller-scale multilingual models such as LLaMA-2 demonstrate competent but less consistent reasoning, often requiring additional fine-tuning to achieve stability. These differences emphasize that effective chatbot deployment cannot rely solely on general-purpose LLM performance, but must consider the specific reasoning load required in public service scenarios.

Another relevant benchmarking factor is factual grounding. Public service chatbots must provide highly accurate information that complies with regulations and policies. Hallucination risk therefore becomes a critical metric. Studies evaluating hallucination rates across models indicate that GPT-4 and PaLM-2 produce substantially fewer factually incorrect statements compared to BLOOM and early-generation open-source models (Waldo & Boussard, 2024). However, even high-performing models may introduce inaccuracies when responding to procedural or domain-specific queries lacking explicit knowledge in their training corpus. This suggests that integrating retrieval-augmented generation or API-based verification is necessary to ensure that LLM responses reflect government-issued regulations. Such hybrid architectures reduce hallucination risk by grounding generative outputs in validated information sources.

The following table synthesizes key benchmarking findings relevant to Indonesian public service chatbot optimization. The comparison includes performance metrics such as linguistic accuracy, contextual reasoning, hallucination risk, and domain adaptability.

Table 1. Benchmarking Summary of LLM Performance for Indonesian Public Service Chatbots

Model	**Indonesian Language Accuracy**	**Contextual Reasoning**	**Hallucination Risk**	**Domain Adaptability**
GPT-4	Very high (90%+ across benchmarks)	Strong multi-step reasoning	Low	High with fine-tuning
PaLM-2	High	Strong	Low–moderate	High
LLaMA-2	Moderate–high	Moderate	Moderate	Requires adaptation
BLOOM	Moderate	Weak–moderate	High	Limited
IndoGPT / INDOLLM	High for Indonesian	Moderate	Low–moderate	Strong for local domains

The table illustrates that while GPT-4 and PaLM-2 demonstrate superior performance across most categories, local Indonesian models such as IndoGPT or INDOLLM show strong promise due to their optimized performance in local linguistic contexts. Their lower hallucination rates and high adaptability make

them suitable for domain-specific deployments, particularly in Indonesian public institutions with limited computational resources. Meanwhile, BLOOM and LLaMA-2 may require extensive fine-tuning to reach acceptable performance levels. This benchmarking highlights that the optimal LLM for Indonesian public service chatbots must balance linguistic proficiency, contextual reasoning, factual reliability, and infrastructure feasibility.

Finally, benchmarking reveals that LLM deployment requires multi-layered system design rather than reliance on a single model. Retrieval-augmented architectures, hybrid rule–LLM pipelines, document-grounding APIs, and human validation workflows are necessary to ensure reliability and compliance with administrative standards (Judjianto & Vandika (2025)). Such system design ensures that LLM-driven chatbots remain accurate, safe, and aligned with government communication requirements. Benchmarking thus serves not only to compare models but to inform holistic system planning for Indonesian digital governance.

Strategic Framework for Implementing LLM-Based Chatbots in Indonesian Public Services

Developing an effective implementation strategy for LLM-based chatbots in Indonesian public services requires a comprehensive framework that integrates governance, technological readiness, data management, and institutional capacity. Public administration operates within strict regulatory environments, which means that AI integration must be aligned with national digital transformation policies, cybersecurity principles, and public accountability standards. One of the most important strategic considerations is the establishment of a regulatory framework that governs AI deployment in government services. This includes defining ethical guidelines, transparency requirements, and responsibility structures. Research in AI governance emphasizes that citizen trust in digital systems depends on clear accountability mechanisms that specify how AI-generated responses are validated, audited, and corrected when errors occur (Zhu et al., 2024). Without such frameworks, LLM-based chatbots risk being perceived as unreliable or unsafe for high-stakes administrative communication.

Another strategic dimension involves strengthening government data infrastructure. LLMs require access to structured, accurate, and up-to-date datasets to ensure regulatory alignment. Indonesia's fragmented data landscape, characterized by siloed information systems and inconsistent data standards, poses a challenge for LLM integration. To address this, government institutions must implement interoperable data management systems that support secure database connectivity for chatbot architectures. Research on e-government modernization highlights that integrated data ecosystems improve service efficiency and support algorithmic reliability by reducing inconsistencies across administrative documents (Wagola et al., 2023). Strengthening data standardization efforts will enable LLMs to provide more precise and consistent responses across public service domains.

Institutional capacity building is another essential component of the strategic framework. Public service agencies require personnel who understand both technological processes and administrative functions to supervise AI deployment effectively. This involves training civil servants in AI literacy, human–AI collaboration, and digital governance practices. Studies show that public-sector AI adoption fails when human operators lack the confidence or skills to manage AI-assisted tools (Siddiqui et al., 2023). Indonesian agencies can adopt capacity-building models implemented in Singapore and South Korea, where public employees undergo structured training programs to ensure that digital transformation initiatives are executed competently and sustainably.

LLM implementation must also include robust risk mitigation strategies. Given the risks of bias, hallucination, and misinformation, government chatbots must incorporate multi-layered validation pipelines that filter AI-generated content. This can be achieved through retrieval-augmented generation models that combine LLM outputs with verified government documents, reducing the likelihood of inaccurate or unauthorized responses. Additionally, real-time monitoring systems should be developed to track chatbot performance, identify error patterns, and implement corrective updates. Ethical safeguards must also be implemented to ensure that LLMs do not reproduce discriminatory language or disadvantage vulnerable populations. Research in fairness-aware NLP demonstrates that bias mitigation techniques, including dataset balancing and adversarial training, can be integrated into LLM pipelines to promote equitable service delivery (Cahyawijaya et al., 2021).

Another strategic priority is public communication and citizen engagement. Successful implementation of LLM-powered chatbots requires building public awareness of how these systems function and how citizens can use them safely and effectively. Communication campaigns should focus on educating the public about the scope and limitations of LLM-based services, emphasizing transparency and user empowerment. Citizen engagement also includes gathering user feedback to improve chatbot performance and ensure alignment with actual service needs. Studies in digital participatory governance indicate that involving citizens in system design increases adoption rates and trust, especially in culturally diverse societies (Judjianto & Vandika (2025)).

The final strategic component is long-term sustainability. LLM integration must be designed to support iterative improvement, scalability, and evolving administrative needs. This includes selecting models that can be periodically updated, fine-tuned with new datasets, and adapted to changing policies. Open-source models may provide cost-effective long-term sustainability through local fine-tuning, while proprietary models may offer higher immediate performance with greater maintenance costs. Governments must therefore evaluate trade-offs between accuracy, cost, sovereignty, and computational requirements when choosing LLM architectures. Sustainable adoption also requires continuous research collaboration with universities,

private-sector AI developers, and digital transformation agencies to ensure that system improvements keep pace with technological evolution.

Overall, implementing LLM-based chatbots in Indonesian public services requires a comprehensive, multi-layered strategy that balances technological innovation with governance, ethics, and institutional readiness. When executed effectively, LLM-powered systems can transform Indonesia's public service landscape by enhancing accessibility, improving service speed, reducing administrative burden, and strengthening citizen trust in digital governance.

CONCLUSION

This study demonstrates that the application of Large Language Models presents significant opportunities to improve Indonesian public service chatbots across linguistic, operational, and governance dimensions. LLMs can interpret diverse Indonesian language expressions, manage multi-intent queries, and generate contextually grounded responses beyond the capacity of rule-based systems. Their superior semantic understanding, adaptability through domain-specific fine-tuning, and ability to adjust linguistic register provide substantial improvements in accessibility and service quality. Benchmarking results highlight that advanced models such as GPT-4 and PaLM-2 offer strong linguistic accuracy and reasoning capabilities, while Indonesian-centric models such as IndoGPT or INDOLLM show high adaptability and lower hallucination risk for localized domains. However, the study also identifies critical challenges including privacy risks, hallucination, institutional readiness, bias, and fragmented data ecosystems, all of which must be addressed before LLM integration can be safely scaled in Indonesia's public sector.

To ensure responsible and effective implementation, several strategic recommendations are proposed. Government agencies must establish clear AI governance frameworks that regulate transparency, accountability, and ethical safeguards. Strengthening secure and interoperable data infrastructure will enable more reliable factual grounding for LLM-generated content. Institutional capacity building is essential to equip civil servants with the skills needed to manage AI-assisted systems, while public communication campaigns are necessary to build citizen trust. Risk mitigation strategies, including retrieval-augmented generation and continuous performance monitoring, should be integrated into system design. Long-term sustainability requires ongoing collaboration between government, academia, and the private sector to support continuous model development, evaluation, and adaptation. When implemented within a comprehensive governance framework, LLM-powered chatbots can play a transformative role in enhancing the efficiency, inclusivity, and responsiveness of Indonesia's digital public services.

REFERENCES

Arifianti, D. L., & Sakapurnama, E. (2024). The strategy of public services through digitalization in Indonesia: A comparative study from South Korea success story. *Journal La Sociale*, *5*(3), 651-658.

Cahyawijaya, S., Winata, G. I., Wilie, B., Vincentio, K., Li, X., Kuncoro, A., ... & Fung, P. (2021, November). IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 8875-8898).

Datta, K. (2024). AI-driven public administration: Opportunities, challenges, and ethical considerations. *The Social Science Review*, *2*(6), 134-139.

Denistia, K., & Baayen, R. H. (2022). The morphology of Indonesian: Data and quantitative modeling. In *The Routledge handbook of Asian linguistics* (pp. 605-634). Routledge.

Gemiharto, I., & Samson, C. M. S. (2024). Inclusivity and Accessibility in Digital Communication Tools: Case Study of AI-Enhanced Platforms in INDONESIA. *Jurnal Pewarta Indonesia*, *6*(1), 78-88.

Judijanto, L., & Vandika, A. Y. (2025). Emerging Research Trends in Natural Language Processing for Multilingual AI. *The Eastasouth Journal of Information System and Computer Science*, *2*(03), 187-199.

Kaun, A., Larsson, A. O., & Masso, A. (2025). Automation scenarios: citizen attitudes towards automated decision-making in the public sector. *Information, Communication & Society*, *28*(7), 1177-1194.

Keith, A. J. (2024). Governance of artificial intelligence in Southeast Asia. *Global Policy*, *15*(5), 937-954.

Ma'rup, M., & Rokhman, A. (2024). Utilization of Artificial Intelligence (AI) Chatbots in Improving Public Services: A Meta-Analysis Study. *Open Access Indonesia Journal of Social Sciences*, *7*(4), 1610-1618.

Misuraca, G., & Viscusi, G. (2020, August). AI-enabled innovation in the public sector: A framework for digital governance and resilience. In *International Conference on Electronic Government* (pp. 110-120). Cham: Springer International Publishing.

Moghe, N., Razumovskaia, E., Guillou, L., Vulić, I., Korhonen, A., & Birch, A. (2023, July). Multi3NLU++: A multilingual, multi-intent, multi-domain dataset for natural language understanding in task-oriented dialogue. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 3732-3755).

Nadzif, M. A., & Soelistijadi, R. (2024). Penggunaan teknologi natural language processing dalam sistem chatbot untuk peningkatan layanan informasi administrasi publik. *The Indonesian Journal of Computer Science*, *13*(1).

Ngai, E. W., Lui, A. K., & Kei, B. C. (2025). Natural language processing in government applications: a literature review and a case analysis. *Industrial Management & Data Systems*, *125*(6), 2067-2104.

Susar, D., & Aquaro, V. (2019, April). Artificial intelligence: Opportunities and challenges for the public sector. In *Proceedings of the 12th international conference on theory and practice of electronic governance* (pp. 418-426).

Wagola, R., Nurmandi, A., Misran, & Subekti, D. (2023, July). Government Digital Transformation in Indonesia. In *International Conference on Human-Computer Interaction* (pp. 286-296). Cham: Springer Nature Switzerland.

Waldo, J., & Boussard, S. (2024). GPTs and hallucination: why do large language models hallucinate?. *Queue*, *22*(4), 19-33.

Walsh, A. (2024). Cusco Quechua and the world of AI: a case study on low resource languages and large language models.

Wongso, W., Lucky, H., & Suhartono, D. (2022). Pre-trained transformer-based language models for sundanese. *Journal of Big Data*, *9*(1), 39.

Zakiuddin, N. F., & Anggara, S. M. (2024). Developing Digital Service Transformation Maturity Model in Public Sector. *IEEE Access*.

Zhu, S., Xu, S., Sun, H., Pan, L., Cui, M., Du, J., ... & Xiong, D. (2024). Multilingual Large Language Models: A Systematic Survey. *arXiv preprint arXiv:2411.11072*.