P-ISSN: 2656-1743 | E-ISSN: 2656-1735 **JURNAL RISET INFORMATIKA**
DOI: https://doi.org/10.34288/jri.v8i4.590 Vol. 8, No. 4. September 2026
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

IMPLEMENTATION AND EVALUATION OF A RETRIEVAL-AUGMENTED GENERATION CHATBOT FOR EXPORT CUSTOMER SERVICE AUTOMATION

Jenri Dewany Widya Sunandar[-1], Mutaqin Akbar[-2]

Program Studi Informatika
Fakultas Teknologi Informasi
Universitas Mercu Buana Yogyakarta
221110017@student.mercubuana-yogya.ac.id[-1], mutaqin@mercubuana-yogya.ac.id[-2]

Abstract

This study discusses the issue of customer service in export companies, which is still limited in terms of service availability and the ability to handle recurring inquiries from international customers. Limitations in human resources result in responses to customer inquiries not being performed optimally, especially outside of operating hours. This study aims to design, implement, and evaluate a Retrieval-Augmented Generation (RAG)-based chatbot to support the automation of customer service in export companies using the Botika Agentic Platform. The study contributes by implementing and quantitatively evaluating a RAG-based chatbot on a Platform as a Service (PaaS) environment using the DeepEval framework with the LLM-as-a-Judge approach. The methods used in this study include four stages, namely data collection through interviews and analysis of company documents, system design, implementation through the Botika platform, and testing using the DeepEval framework with the LLM-as-a-Judge approach. The testing dataset consists of 93 questions covering four knowledge base domains, namely company profile, export and shipping procedures, partnership information, and product information. Evaluation results indicate that the chatbot achieved an average Answer Relevancy score of 0.98, Faithfulness of 1.00, and Contextual Relevancy of 0.88. The system performs well, with high Answer Relevancy and Faithfulness scores, while Contextual Relevancy is relatively lower, particularly in the product information domain. This is due to the dense structure of product documents, which contain many product variations within a single document, thereby affecting the accuracy of the retrieval process. Overall, the proposed RAG-based chatbot demonstrates its potential to support customer service automation in export companies by providing accurate and contextually grounded responses, while highlighting the importance of knowledge base organization for improving retrieval performance.

Keywords**:** Retrieval-Augmented Generation, Chatbot, Customer Service, Export Company, DeepEval, Botika Agentic Platform

Abstrak

Penelitian ini membahas permasalahan layanan pelanggan pada perusahaan ekspor yang masih terbatas dalam hal ketersediaan waktu layanan dan kemampuan menangani pertanyaan berulang dari pelanggan internasional. Keterbatasan sumber daya manusia menyebabkan respon terhadap pertanyaan pelanggan tidak dapat dilakukan secara optimal, terutama di luar jam operasional. Penelitian ini bertujuan untuk merancang, mengimplementasikan dan mengevaluasi chatbot berbasis Retrieval-Augmented Generation (RAG) untuk mendukung otomatisasi layanan pelanggan pada perusahaan ekspor menggunakan Botika Agentic Platform. Metode yang digunakan dalam penelitian ini meliputi empat tahapan, yaitu pengumpulan data melalui wawancara dan analisis dokumen perusahaan, perancangan sistem, implementasi melalui platform Botika serta pengujian menggunakan framework DeepEval dengan pendekatan LLM-as-a-Judge. Dataset pengujian terdiri dari 93 pertanyaan yang mencakup empat domain knowledge base yaitu profil perusahaan, prosedur ekspor dan pengiriman, informasi kemitraan dan informasi produk. Hasil evaluasi menunjukkan bahwa chatbot memperoleh nilai rata-rata Answer Relevancy sebesar 0.98, Faithfulness sebesar 1.00, dan Contextual Relevancy sebesar 0.88. Secara keseluruhan, sistem menunjukkan kinerja yang baik dengan nilai Answer Relevancy dan Faithfulness yang tinggi, serta Contextual Relevancy yang relatif lebih rendah terutama pada domain informasi produk. Hal ini disebabkan oleh struktur dokumen produk yang padat dan mengandung banyak variasi produk dalam satu dokumen, sehingga mempengaruhi ketepatan

proses retrieval. Secara keseluruhan, chatbot berbasis RAG menunjukkan potensi dalam mendukung otomatisasi layanan pelanggan pada perusahaan ekspor dengan menghasilkan respons yang akurat dan berlandaskan konteks, sekaligus menegaskan pentingnya pengelolaan knowledge base yang baik untuk meningkatkan kinerja proses retrieval.

Kata kunci: Retrieval-Augmented Generation, chatbot, layanan pelanggan, perusahaan ekspor, DeepEval, Platform Agentic Botika

INTRODUCTION

Major changes in customer service across various business lines are being influenced by the development of artificial intelligence technology, including in the international trade sector (Ozturk, 2024). In Indonesia, building relationships with international buyers is often hampered by communication. (Nathalie et al., 2025) identified cultural differences, language barriers, and time zone disparities as major obstacles to effective international business communication. This situation can be exacerbated by limited workforce and the unavailability of services outside of business hours, which can reduce customer satisfaction and trust (Abel Uzoka, Emmanuel Cadet, & Pascal Ugochukwu Ojukwu, 2024).

Preliminary interviews conducted in February 2026 with an export company located in Sleman, Yogyakarta, revealed recurring customer service challenges related to repetitive inquiries from international clients and limited service availability outside business hours. These challenges may lead to delayed responses, increased staff workload and reduced service efficiency, particularly when customers require immediate access to company specific information.

Conventional customer service solutions, including rule based chatbots, often struggle to handle diverse customer inquiries and provide responses based on company specific knowledge (Lin, Wang, Shao, & Taylor, 2024). In export business, customer inquiries frequently involve product details, export procedures, company policies and operational information that must remain consistent with organizational standards. Consequently, a more intelligent solution is required to deliver accurate, context aware and policy-complaint responses while reducing staff workload and maintaining service availability beyond business hours.

Chatbot technology has evolved significantly from rule-based approaches to machine learning algorithms such as Naive Bayes and Support Vector Machine, and further to transformer-based Large Language Models (LLM) (Muhammad Abdul Gofur & Akbar, 2024; Okonkwo & Ade-Ibijola, 2021; Pria Purnama & Akbar, 2026) However, LLMs have fundamental limitations including knowledge cutoff boundaries and the tendency to generate inaccurate information, commonly referred to as hallucination (Ji et al., 2023). Retrieval Augmented Generation (RAG) addresses these limitations by combining LLM generative capabilities with an external knowledge base, enabling contextually accurate responses grounded in specific organizational documents (Denis Rothman, 2024; Klesel & Wittmann, 2025).

Previous research has demonstrated the applicability of RAG-based chatbots in various Indonesian institutional contexts (Pujiono, Agtyaputra, & Ruldeviyani, 2024), and academic service domains (Hariyadi & Soraya, 2025), as well as its effectiveness in customer service for small-scale enterprises (Marcineková, Sujová, & Ďurica, 2025). These studies indicate that RAG-based chatbot can improve response accuracy and reduce hallucination by grounding generated responses in domain specific knowledge sources.

However, most existing studies focus on educational and institutional settings, while applications in export business remain limited. Export business presents unique challenges, including multilingual communication, international customer interactions and the need to provide accurate information regarding products and export procedures (Nathalie et al., 2025; Ozturk, 2024). Furthermore, relatively few studies gave systematically evaluated RAG-based chatbot using automated evaluation framework such as DeepEval with the LLM-as-a-Judge approach. To address these gaps, this study employs the Botika Agentic Platform, which was selected because it provides integrated support for knowledge base management, retrieval workflows, and LLM deployment without requiring custom infrastructure development. The study designs, implements, and evaluates a RAG-based chatbot for customer service automation in an export company. Chatbot performance is evaluated using the DeepEval framework based on Answer Relevancy, Faithfulness and Contextual Relevancy.

P-ISSN: 2656-1743 | E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i4.590
JURNAL RISET INFORMATIKA
Vol. 8, No. 4. September 2026
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

RESULTS AND DISCUSSION

The RAG-based chatbot was successfully implemented on the Botika Agentic Platform. The system integrates a knowledge base containing company profile information, export and shipping procedures, partnership information, and product information. The chatbot is accessible through a webchat interface and supports interactions in both Indonesian and English. In addition, separate interaction flows were implemented for client and supplier user segments, while service requests such as Request for Quotation (RFQ) submissions and supplier registration were integrated with Google Sheets for automated data recording.

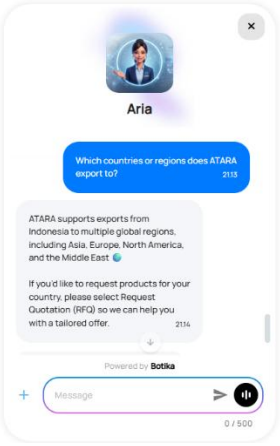

Figure 4. Implemented Chatbot Interface for Buyer

Figure 5. Implemented Chatbot Interface for Supplier

Figure 4 and 5 present examples of the implemented chatbot interface for buyer and supplier users. Figure 4 demonstrates an interaction with a buyer using an English query, while Figure 5 presents an interaction with a supplier using an Indonesian query. In both cases, the chatbot generates responses that are consistent with the information stored in the company knowledge base, demonstrating that the Retrieval Augmented Generation (RAG) mechanism successfully retrieves relevant information and provides contextually appropriate responses. These examples also show that the implemented chatbot supports multilingual communication while maintaining response consistency across different user segments (Lin et al., 2024).

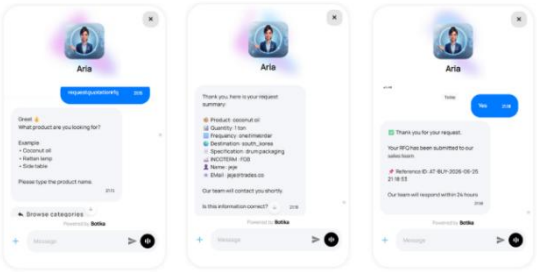

Figure 6. RFQ Submission Workflow through the Chatbot

Figure 6 illustrates the Request for Quotation (RFQ) submission workflow implemented in the chatbot. The workflow guides users through a structured data collection process by requesting product category, quantity, order frequency, destination country, specifications, incoterm, and contact information. After all required information has been collected, the chatbot generates a summary for user confirmation before submitting the request. Once confirmed, the submitted data are automatically recorded in Google Sheets, enabling company staff to process quotation requests without manual data entry.

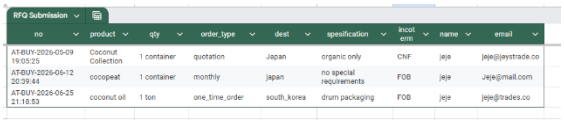

Figure 7. RFQ data in Google Sheets

Figure 7 presents the Google Sheets integration used to record RFQ submissions. Each confirmed request is automatically stored as a new record containing customer information and quotation details. This integration enables structured request management and demonstrates that the chatbot supports not only information retrieval but also customer service process automation.

Evaluation Results

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v8i4.590
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

The implemented chatbot was evaluated using the DeepEval framework following the LLM-as-a-Judge methodology, with Gemini (gemini-3.1-flash-lite) serving as the evaluation model. Evaluation was performed using 93 test cases covering four knowledge domains: company profile, export and shipping, partnership, and product information. Each test case was assessed using three evaluation metrics: Answer Relevancy, Faithfulness, and Contextual Relevancy. A performance threshold of 0.80 was adopted for all evaluation metrics.

Table 2. Overall DeepEval Evaluation Results

Metric	Avg	Min	Max
Answer Relevancy	0.98	0.5	1.00
Faithfulness	1.00	0.5	1.00
Contextual Relevancy	0.88	0.11	1.00

The overall evaluation results indicate that the chatbot achieved high performance across all three evaluation metrics. Faithfulness obtained the highest average score (1.00), followed by Answer Relevancy (0.98), while Contextual Relevancy achieved an average score of 0.88. These results suggest that the chatbot consistently generated responses grounded in the retrieved knowledge base while maintaining high semantic relevance to user queries.

Considering the three evaluation metrics simultaneously, 70 out of 93 test cases (75.27%) satisfied the predefined threshold (≥ 0.80) for Answer Relevancy, Faithfulness, and Contextual Relevancy. The remaining 23 test cases (24.73%) failed to satisfy at least one of the evaluation metrics. Performance across each knowledge domain is summarized in Table 3.

Table 3. Evaluation Results per Category

Category	N	CR	AR	F	Pass
Company Profile	12	0.97	1.00	1.00	11
Export & Shipping	24	0.97	0.99	1.00	23
Partnership	12	0.97	0.95	1.00	11
Product Info	45	0.80	0.97	1.00	25

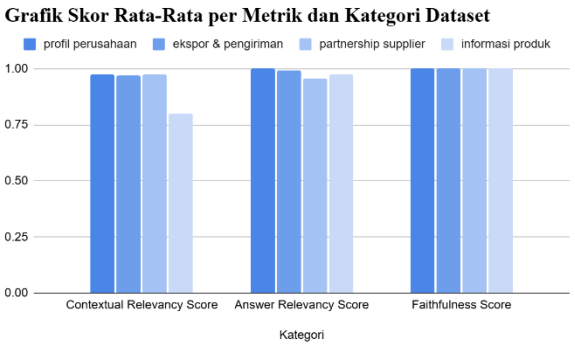

Figure 8. Chart of Average Scores per Metric and Dataset Category

Overall, the chatbot demonstrated consistently strong performance in the Company Profile, Export and Shipping, Partnership domains, with average metrics scores exceeding 0.95. In contrast, the Product information domain recorded the lowest Contextual Relevancy score (0.80), indicating that retrieving relevant information for product-related queries posed greater challenges than for other knowledge domains.

Discussion

The system achieved an average Answer Relevancy score of 0.98, corresponding to a very high-performance level on the adopted evaluation scale. This high performance across all four categories indicates that the instruction prompt configuration on the Botika Agentic Platform contributes to guiding the model to understand question context and generate appropriately targeted responses. Failures in this metric were primarily observed in questions requiring specific, concise answers where the chatbot produced overly detailed responses containing peripheral information. In these cases, the retrieved context contained additional product information beyond the user's request, reducing Contextual Relevancy despite the generated response remaining factually correct.

Faithfulness averaged 1.00, the highest score among the three metrics, indicating that chatbot responses are consistently grounded in the knowledge base and do not contain hallucinated information. This result suggests that the instruction prompt helps guide the GPT-4o model to rely primarily on retrieved context, consistent with the principle that RAG architectures reduce hallucination risk by anchoring generation to retrievable knowledge (Klesel & Wittmann, 2025).

Contextual Relevancy recorded the lowest average score (0.88), with the Product Information category obtaining the lowest average score (0.80). The distribution of Contextual Relevancy scores is presented in Table 4.

Table 4. Contextual Relevancy Score Distribution

Score Range	Category	Number of Cases	Percentage
0.80 – 1.00	Very Good	71	76.34%
0.60 – 0.79	Good	10	10.75%
0.40 – 0.59	Fair	9	9.68%
0.20 – 0.39	Poor	2	2.15%
0.00 – 0.19	Very Poor	1	1.08%

As shown in Table 4, 22 of the 93 test cases (23.66%) scored below the predefined threshold of 0.80. Most of these cases were classified as Good (10 cases), Fair (9 cases), indicating that retrieval quality generally remained acceptable but lacked sufficient precision for certain queries. This pattern is attributed to the dense structure of product catalogue documents, where a single document contains multiple product variants with overlapping specifications. Examination of the retrieval logs showed that the retrieved context frequently contained complete product catalogue sections rather than passages specific to the requested product variant. Consequently, when users ask about a specific variant, the retriever tends to retrieve the entire document block, resulting in lower context precision (Gao et al., 2024).

The combination of high Faithfulness (1.00) and comparatively lower Contextual Relevancy (0.88) indicates that the retrieval and generation components contribute differently to the overall system performance. Although the retriever occasionally provides broader context than necessary, the GPT-4o model remains capable of generating responses that are consistent with the retrieved information while remaining faithful to the retrieved context (Khasanova Zafar kizi & Suh, 2025; Klesel & Wittmann, 2025). These findings demonstrate that the proposed RAG-based chatbot is capable of supporting customer service automation for export companies by generating accurate and contextually grounded responses (Abel Uzoka et al., 2024). At the same time, the results identify knowledge base organizations as the primary factor affecting retrieval quality, suggesting that restructuring product information into smaller and more granular documents could further improve the contextual retrieval performance (Gao et al., 2024; Klesel & Wittmann, 2025; Yang et al., 2025).

Failure analysis showed that most unsuccessful cases were associated with product-related queries containing multiple product variants within a single knowledge document. In these cases, the retrieval component tended to return broader document sections instead of passages specific to the requested product, reducing Contextual Relevancy while maintaining high Faithfulness. Examination of the retrieval logs confirmed that retrieval errors mainly originated from document organization rather than incorrect response generation.

The implementation also has several practical limitations. Because the chatbot was deployed on the Botika Agentic Platform, low-level configuration such as embedding models, retrieval parameters, vector indexing, and inference settings are managed internally by the platform. Consequently, optimization in this study primarily focused on knowledge base organization, workflow configuration, and prompt design. Furthermore, computational efficiency, inference latency, and scalability were not quantitatively evaluated because these aspects are abstracted by the platform infrastructure and fall outside the scope of this study.

CONCLUSIONS AND SUGGESTIONS

Conclusion

This study demonstrates that a RAG chatbot was successfully designed and implemented using the Botika Agentic Platform to support customer service automation in an export company. The developed system is capable of providing continuous responses to customer inquiries through a web-based interface without operational time limitations.

The knowledge base constructed from verified company documents and interviews enables the chatbot to generate responses that are consistent with organizational information. The integration of the RAG mechanism ensures that all generated responses are grounded in the available knowledge base.

The evaluation using the DeepEval framework with LLM-as-a-Judge approach indicates strong overall system performance

across all metrics. The result shows that the chatbot is able to produce relevant, contextually grounded, and reliable responses. However, performance differences in contextual relevancy suggest that further improvement is required in the structure of the knowledge base, particularly for product-related information with higher variability and complexity.

This study contributes to the growing body of research on Retrieval-Augmented Generation by demonstrating the implementation and evaluation of a RAG-based chatbot in the export business domain using the Botika Agentic Platform. From a practical perspective, the proposed chatbot provides an automated customer service solution capable of supporting multilingual information services and structured service requests, thereby reducing manual workload and improving service availability for export companies.

Suggestion

Future work should focus on continuously updating and expanding the knowledge base to ensure broader coverage of user inquiries and alignment with evolving company information. Previous studies have shown that the quality and completeness of the knowledge base directly influence retrieval effectiveness in RAG systems (Gao et al., 2024; Klesel & Wittmann, 2025). Furthermore, product-related documents should be reorganized into smaller and more granular knowledge units to improve retrieval precision, particularly for product information containing multiple variants and overlapping specifications (Yang et al., 2025).

Further research may also explore comparative evaluations of different Large Language Models available within the Botika Agentic Platform, such as OpenAI, Groq, and other LLM providers, to identify the most suitable model for customer service applications in export business. Moreover, expanding the evaluation dataset with more diverse and complex conversational scenarios would provide a more comprehensive assessment of system performance. Recent studies recommend evaluating RAG systems using diverse benchmark datasets and realistic user queries to obtain more representative performance assestments (Es et al., 2024; Zheng et al., 2023).

Future research may also investigate advanced RAG techniques, such as hybrid retrieval or knowledge graph integration, in deployment environments that support these capabilities to further improve retrieval quality.

REFERENCES

Abel Uzoka, Emmanuel Cadet, & Pascal Ugochukwu Ojukwu. (2024). Leveraging AI-Powered chatbots to enhance customer service efficiency and future opportunities in automated support. *Computer Science & IT Research Journal*, *5*(10), 2485–2510. https://doi.org/10.51594/csitrj.v5i10.1676

Denis Rothman. (2024). *RAG-Driven Generative AI: Build custom retrieval augmented generation pipelines with LlamaIndex, Deep Lake, and Pinecone.* Packt Publishing Ltd.

Es, S., James, J., Espinosa-Anke, L., Schockaert, S., & Gradients, E. (2024). RAGAS: Automated Evaluation of Retrieval Augmented Generation. In *https://aclanthology.org/2024.eacl-demo.16.pdf*. https://doi.org/doi: 10.18653/v1/2024.eacl-demo.16

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., … Wang, H. (2024). *Retrieval-Augmented Generation for Large Language Models: A Survey*. Retrieved from http://arxiv.org/abs/2312.10997

Hariyadi, I. N. R., & Soraya, D. U. (2025). JEPIN (Jurnal Edukasi dan Penelitian Informatika). *Https://Jurnal.Untan.Ac.Id/Index.Php/Jepin/Article/View/101635*, *11*(3), 442–447. https://doi.org/10.26418/jp.v11i3.101635

Jeffrey Ip. (2025). RAG Evaluation Metrics: Assessing Answer Relevancy, Faithfulness, Contextual Relevancy, And More - Confident AI. Retrieved May 25, 2026, from https://www.confident-ai.com/blog/rag-evaluation-metrics-answer-relevancy-faithfulness-and-more website: https://www.confident-ai.com/blog/rag-evaluation-metrics-answer-relevancy-faithfulness-and-more

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., … Fung, P. (2023, December 31). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, Vol. 55. Association for Computing Machinery. https://doi.org/10.1145/3571730

Khasanova Zafar kizi, M., & Suh, Y. (2025). Design and Performance Evaluation of LLM-Based RAG Pipelines for Chatbot Services in International Student Admissions. *Electronics (Switzerland)*, *14*(15).

https://doi.org/10.3390/electronics14153095

Klesel, M., & Wittmann, H. F. (2025). Retrieval-Augmented Generation (RAG). *Business and Information Systems Engineering*, *67*(4), 551–561. https://doi.org/10.1007/s12599-025-00945-3

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., … Kiela, D. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. Retrieved from https://github.com/huggingface/transformers/blob/master/

Lin, X., Wang, X., Shao, B., & Taylor, J. (2024). How Chatbots Augment Human Intelligence in Customer Services: A Mixed-Methods Study. *Journal of Management Information Systems*, *41*(4), 1016–1041. https://doi.org/10.1080/07421222.2024.2415773

Marcineková, K., Sujová, A. J., & Ďurica, R. (2025). Implementing AI Chatbots in Customer Service Optimization—A Case Study in Micro-Enterprise. *Information (Switzerland)*, *16*(12). https://doi.org/10.3390/info16121078

Muhammad Abdul Gofur, & Akbar, M. (2024). Prototipe chatbot pada aplikasi e-commerce berbasis web (studi kasus: toko seni gypsum). *Jurnal RESTIKOM : Riset Teknik Informatika Dan Komputer*, *6*, 240–250. https://doi.org/10.52005/restikom.v6i2.338

Nathalie, J., Paramita, S., Komunikasi, H., Di, B., Ekspor, P., Kasus, S., … Logam, D. (2025). *PT Sinar Dunia Logam)*. https://doi.org/10.24912/pr.v9i2.33186

Okonkwo, C. W., & Ade-Ibijola, A. (2021, January 1). Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence*, Vol. 2. Elsevier B.V. https://doi.org/10.1016/j.caeai.2021.100033

Ozturk, O. (2024). The Impact of AI on International Trade: Opportunities and Challenges. *Economies*, *12*(11). https://doi.org/10.3390/economies12110298

Pria Purnama, A., & Akbar, M. (2026). Pengembangan Chatbot Pintar Untuk Layanan Konsultasi Hukum Menggunakan OpenAI GPT-4. *SMARTICS Journal*. https://doi.org/10.21067/smartics.v12i1.13817

Pujiono, I., Agtyaputra, I. M., & Ruldeviyani, Y. (2024). IMPLEMENTING RETRIEVAL-AUGMENTED GENERATION AND VECTOR DATABASES FOR CHATBOTS IN PUBLIC SERVICES AGENCIES CONTEXT. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *10*(1), 216–223. https://doi.org/10.33480/jitk.v10i1.5572

Shahriar, S., Lund, B. D., Mannuru, N. R., Arshad, M. A., Hayawi, K., Bevara, R. V. K., … Batool, L. (2024). Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency. *Applied Sciences (Switzerland)*, *14*(17). https://doi.org/10.3390/app14177782

Yang, R., Fu, M., Tantithamthavorn, C., Arora, C., Vandenhurk, L., & Chua, J. (2025). RAGVA: Engineering retrieval augmented generation-based virtual assistants in practice. *Journal of Systems and Software*, *226*. https://doi.org/10.1016/j.jss.2025.112436

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., … Stoica, I. (2023). *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. Retrieved from http://arxiv.org/abs/2306.05685