Research Article

Digital System and Computing

An LLaMA 3.1-Based Chatbot with Retrieval-Augmented Generation (RAG) for Academic Services at UPN "Veteran" Yogyakarta

Farel Abid Yasser Prayanto, Rifki Indra Perwira

[The author informations are in the declarations section. This article is published by ETFLIN in Digital System and Computing, Volume 2, Issue 1, 2026, Page 19-26. DOI: 10.58920/dsc0201633]

Received: 13 March 2026
Revised: 25 May 2026
Accepted: 10 June 2026
Published: 17 June 2026

Editor: Ferian Fauzi Abdulloh

This article is licensed under a Creative Commons Attribution 4.0 International License. © The author(s) (2026).

Keywords: Retrieval-Augmented Generation, Academic Chatbot, Large Language Models, Hybrid Information Retrieval, Higher Education Information Systems.

Abstract: While universities heavily rely on digital information systems, static websites and manual administrative communication often limit accessibility and responsiveness for students seeking academic information. To address this, this study developed and evaluated an academic chatbot using the LLaMA 3.1 large language model integrated with a Retrieval-Augmented Generation (RAG) framework for Informatics students at Universitas Pembangunan Nasional "Veteran" Yogyakarta. Employing a Rapid Application Development approach, 263 institutional document chunks were processed to construct a knowledge base for a hybrid retrieval pipeline that combines BM25 lexical search and semantic vector similarity. The proposed system was comprehensively benchmarked against standalone lexical-only and semantic-only baselines using both RAG-specific and natural language generation (NLG) metrics. Experimental results demonstrated that the hybrid strategy achieved the highest answer faithfulness (0.712) and context recall (0.895), representing a 29.5% and 32.8% improvement in faithfulness over the respective standalone baselines, thereby ensuring superior factual consistency. Furthermore, the hybrid system recorded a Token F1 Score of 0.499, a BLEU score of 0.233, and a faster average response time of 7.64 seconds due to parallel query execution and context-size optimization. Finally, exploratory user evaluation yielded high satisfaction with an overall score of 4.46 out of 5.00, confirming its viability for real-world academic assistance.

Introduction

The rapid development of artificial intelligence (AI) and natural language processing (NLP) technologies has significantly transformed how users interact with digital information systems. In particular, the emergence of Large Language Models (LLMs) has enabled machines to understand and generate human-like text, facilitating the development of intelligent conversational agents capable of assisting users in various domains (1, 2). In higher education environments, access to accurate and timely academic information is essential for supporting student activities such as course registration, curriculum planning, academic regulations, and thesis administration. However, many academic information services in universities still rely on static websites or manual administrative communication, which may limit accessibility and responsiveness. These limitations often prevent students from obtaining information efficiently, particularly outside office hours. As digital transformation accelerates within educational institutions, there is an increasing demand for intelligent systems capable of providing real-time, automated academic assistance to students (3, 4).

Recent studies have explored the use of chatbots powered by LLMs to improve information accessibility and user interaction in educational environments (5, 6). These systems can process natural language queries and generate contextually appropriate responses, allowing students to access academic information more efficiently (7). Nevertheless, conventional LLM-based chatbots still face a critical limitation known as hallucination, where the model generates responses that appear plausible but are factually incorrect because they rely solely on pretrained knowledge without access to external data sources (8). This issue becomes particularly problematic in domains such as academic administration, where accurate and reliable information is essential.

To address this limitation, researchers have proposed Retrieval-Augmented Generation (RAG), an architecture that integrates information retrieval with generative

language models (9, 10). In this approach, relevant documents are first retrieved from a knowledge base and then provided as contextual input to the language model to generate more accurate and evidence-grounded responses (11). The RAG framework has been shown to significantly improve factual accuracy, contextual relevance, and reliability in knowledge-intensive tasks compared with traditional LLM systems (12 – 14).

Despite these advances, standard RAG architectures utilizing a single retrieval stream struggle in specialized academic environments. Academic documents possess a unique structural hierarchy and semantic density, containing exact codes (e. g., course codes like "123210052"), rigid numeric rules (e. g., "minimum 130 SKS" or "GPA of 2.00"), and specific nomenclature (e. g., lecturer names with intricate academic titles) alongside general academic jargon. Standalone semantic search utilizing dense vector embeddings often misses exact lexical matches for numeric constraints and course codes due to embedding compression. Conversely, standalone keyword-matching search (such as BM25) fails to capture semantic synonyms or conversational student phrasing (e. g., mapping "bimbingan skripsi" to "konsultasi tugas akhir"). Thus, a critical research gap exists regarding how to establish a robust local RAG system that balances lexical precision and semantic flexibility within localized, resource-constrained higher education settings.

Furthermore, to the best of our knowledge, most existing RAG-based chatbot studies for academic services focus on English-language corpora. This study addresses an underexplored gap by validating a hybrid RAG architecture on Indonesian-language institutional documents, where mixed formal-informal language patterns and domain-specific acronyms (e. g., SKS, MBKM, KRS) present distinct retrieval challenges not addressed by prior work.

Alongside the development of retrieval-augmented systems, recent progress in open-source LLM architectures has enabled institutions to implement advanced AI systems locally. One of the most influential models is the LLaMA family of language models, which offers competitive performance while allowing flexible deployment for domain-specific applications (15). Integrating LLaMA models with RAG architectures provides a promising approach for developing domain-specific conversational systems that utilize institutional knowledge sources effectively (16). Despite these advances, the implementation of RAG-based chatbots for academic service environments remains relatively limited, particularly in the context of integrating modern LLM architectures with hybrid retrieval mechanisms to improve information accuracy and accessibility. Therefore, this study aims to develop and evaluate an academic chatbot system using the LLaMA 3.1 large language model integrated with a Retrieval-Augmented Generation (RAG) framework to

Table 1. Document characteristics and adaptive chunking configurations.

No	Document Name	Pages	Category	Chunk Size	Overlap	Information Coverage
1	Dosen_dan_Kurikulum_Informatika. pdf (Informatics Lecturers and Curriculum Guidelines)	6	Lecturers & Curriculum	800	200	Lecturer profiles, curriculum structure, courses
2	Informasi_Fakultas_Teknik_Industri. pdf (Faculty of Industrial Technology Information Guide)	3	Facilities	500	100	Faculty profile, laboratories, facilities
3	Informasi_Umum_UPN. pdf (UPN General Information Handbook)	3	Profile	1000	300	History, vision and mission, organizational structure
4	Kerja_Praktik. pdf (Practical Work / Internship Guidelines)	5	Procedure	700	180	Internship guidelines, requirements, procedures
5	Peraturan_Akademik. pdf (Academic Regulations)	12	Regulations	600	150	Academic policies, rules, sanctions
6	Peraturan_Kemahasiswaan. pdf (Student Affairs and Conduct Regulations)	4	Regulations	600	150	Student rights and obligations, organizations
7	Program_Studi_Informatika. pdf (Informatics Study Program Profile)	4	Profile	1000	300	Program profile, vision and mission, strengths
8	Program_Studi_Sistem_Informasi. pdf (Information Systems Study Program Profile)	6	Profile	1000	300	Information Systems program profile, curriculum
9	Tugas_Akhir. pdf (Undergraduate Thesis / Final Project Regulations)	9	Procedure	700	180	Final project guidelines, requirements, stages
10	Website_Dokumen_Informatika. pdf (Informatics Document Portal and Web Information)	1	General	500	100	Website information, contacts, announcements

support academic information services for Informatics students at Universitas Pembangunan Nasional "Veteran" Yogyakarta. The proposed system utilizes institutional academic documents as its knowledge base and applies hybrid retrieval methods combining BM25 and semantic search to enhance information retrieval before response generation. The system architecture consists of a document store, a chatbot engine integrated with LLaMA through Ollama, a user interface built with Streamlit, and an evaluation module designed to measure system performance using multidimensional metrics such as context relevance, answer faithfulness, and response accuracy. Through this approach, the study aims to improve the accuracy, efficiency, and accessibility of academic information services while demonstrating the practical application of RAG-based LLM systems in higher education environments.

This study addresses these gaps by implementing and evaluating a locally deployed academic chatbot system using the LLaMA 3.1 large language model integrated with a hybrid RAG framework at Universitas Pembangunan Nasional "Veteran" Yogyakarta. The system runs completely locally to preserve data privacy and eliminate operational costs. We propose a hybrid retrieval strategy that combines BM25 lexical search with dense vector embeddings utilizing a multilingual semantic model, resolved through max-score linear scaling. Furthermore, we implement an adaptive category-specific prompt engineering mechanism to handle the diverse inquiries of informatics students. This paper presents a multidimensional evaluation framework comprising RAG-specific criteria (via RAGAS), traditional NLG metrics, robustness assessments, system latency profiling, and exploratory usability testing to validate the engineering choices and empirical performance of the local RAG architecture.

Methodology

Research Design

This study employed an experimental system development approach to design and evaluate an academic service chatbot for students of the Informatics Study Program at UPN "Veteran" Yogyakarta. The development process followed the Rapid Application Development (RAD) framework, emphasizing iterative prototyping and continuous refinement of system functionality. This approach enabled rapid integration between the retrieval component and the generative language model while ensuring the system could respond effectively to students' academic queries (17).

Data Sources and Preparation

The knowledge base was constructed using 10 official academic PDF documents obtained from the Informatics Study Program at UPN "Veteran" Yogyakarta, spanning academic guidelines, curricula, internship guidelines, and thesis regulations. Data preparation involved robust text extraction, cleaning, and normalization to remove formatting artifacts.

Following extraction, the text was segmented into smaller units using a RecursiveCharacterTextSplitter. To preserve semantic cohesion, adaptive chunking configurations were applied based on document category, as detailed in **Table 1**.

The segmentation pipeline yielded 263 high-coherence text chunks. Each chunk was indexed in parallel into two vector stores: a lexical store via BM25 and a semantic vector store utilizing the intfloat/multilingual-e5-small embedding model (384 dimensions) (18).

Hybrid Retrieval Mathematical Framework

To query the document store, the system implements a hybrid retrieval framework that dynamically merges keyword matching and vector-space similarities. Let Q represent the student's query and D represent a document chunk within collection C.

Lexical Component (BM25)

The lexical relevance of document D to query Q is computed using the BM25Okapi formula, with hyperparameters k_1 = 1.2 and b = 0.75 (standard defaults). This score captures exact token-level matches, making it particularly effective for queries containing precise course codes, acronyms, and numeric thresholds.

Semantic Component

Semantic similarity is calculated as the cosine similarity between the 384-dimensional dense vector representations of the query and the document chunk, generated by the multilingual-e5-small embedding model (18). This score captures conceptual relatedness and handles natural language variation in student queries.

Normalization and Linear Weight Fusion

Because lexical scores produced by BM25 are unbounded ([0, ∞)) and semantic scores are bounded cosine similarities ([−1, 1]), direct addition is mathematically invalid and would skew results. To address this, Maximum Score Linear Scaling is applied to normalize BM25 scores as seen in **Equation 1**. Once normalized, both scores occupy a consistent [0, 1] range. The final hybrid score was calculated using the **Equation 2**.

In this study, the mixing parameter was set to α = 0.3, allocating 30% weight to keyword matching (preserving precise codes and names) and 70% weight to semantic understanding (providing tolerance for natural variations in student phrasing). This weighting reflects the structural characteristics of the target corpus: while the documents contain entity-critical terminology requiring lexical anchoring, the predominant content consists of procedural descriptions and regulatory clauses that benefit more from

$$normalized\_BM25\,(D,\,Q) = \frac{Score\_BM25\,(D,\,Q)}{max\_\{d' \in C\}\,Score\_BM25\,(d',\,Q)} \quad \textbf{(Eq. 1)}$$

$$Score\_Hybrid\,(D,\,Q) = \alpha \times normalized\_BM25\,(D,\,Q) + (1-\alpha) \times Score\_Semantic\,(D,\,Q) \quad \textbf{(Eq. 2)}$$

semantic comprehension. This is consistent with findings demonstrating that dense semantic retrieval outperforms sparse lexical methods for natural-language queries (19), while the retained BM25 component prevents systematic failures on entity-dense queries (9). The top K (K = 10) chunks with the highest final scores are selected as the context window.

Generative Modeling and System Configuration

The generation module utilizes the LLaMA 3.1 large language model (8B parameter chat-instruct version), running locally through Ollama (20). To reduce hallucination, we implemented an adaptive prompt engineering mechanism. Depending on the classified query intent (e. g., "dosen", "kurikulum", "prosedur"), the system dynamically attaches specialized instructions to the system prompt. Crucially, the prompt is structured with a rigid negative constraint: if the retrieved context window does not contain information matching the student's query, the model is strictly instructed to state that the information is unavailable in the database rather than extrapolating or guessing from its parametric memory.

The model parameters were configured to prioritize factual determinism: temperature = 0.3, top_p = 0.9, top_k = 40, and the context window size (num_ctx) was restricted to 6144 tokens to prevent context dilution. The complete system was implemented in Python using LangChain as the orchestration framework, with the intfloat/multilingual-e5-small model from the Sentence Transformers library for embedding generation. Document vectors were indexed using FAISS (Facebook AI Similarity Search). The chunking parameters (chunk size and overlap per document category as shown in **Table 1**), retrieval top-K setting, and prompt templates are fully documented to support reproducibility.

Results and Discussion

Ablation Study of Retrieval Configurations

To evaluate the proposed hybrid retrieval configuration, we performed an ablation study by isolating individual components of the retrieval pipeline. Standard standalone lexical search (BM25 Only) and standalone semantic search (Semantic Only) were evaluated against the combined Hybrid (30: 70) configuration across 22 validated academic test cases. The performance was analyzed using the RAGAS framework (21) and is summarized in **Table 2**. Due to computational constraints of the local RAGAS evaluator running on consumer-grade hardware, the RAGAS framework metrics (faithfulness, answer relevancy, context precision, and context recall) were computed as batch-level aggregate scores across all 22 test cases simultaneously, consistent with the reference implementation by Es *et al.* (21).

The hybrid retrieval method achieved the highest faithfulness score (0.712), outperforming both the BM25-only (0.550) and semantic-only (0.536) baselines. This result aligns with the theoretical framework established by Lewis *et al.* (9), who demonstrated that response faithfulness in RAG systems is fundamentally constrained by retrieval completeness: when retrieved chunks fail to capture all relevant information, the LLM is compelled to supplement with parametric knowledge, thereby increasing hallucination risk. The hybrid approach, by achieving the highest Context Recall (0.895), provides the LLM with a more complete contextual grounding, directly translating to superior faithfulness which is a causal mechanism consistent with the retrieval-generation dependency described by Borgeaud *et al.* (11).

Standard semantic retrievers sometimes fetch contextually related but factually incorrect pages, whereas keyword-matching often fails to capture the overall meaning. Combining them ensures that the generated responses are well-supported by the extracted source text. This is particularly critical in the academic domain where precise regulatory information (SKS thresholds, credit requirements, procedural deadlines) must be reproduced accurately.

Resolving the Hybrid Latency Paradox

A noteworthy technical observation in **Table 2** is that the Hybrid approach (7.64 seconds) is significantly faster than the Semantic Only approach (13.23 seconds), despite executing both retrieval streams. Beyond the mean latency, the distribution reveals important characteristics: the standard deviation of 4.15 seconds (range: 0–21.44 seconds) reflects that simple factual queries (e. g., "Who is the head of department?") resolve in under 6 seconds, whereas complex multi-step procedural queries require up to 21.44 seconds due to longer generation sequences.

Table 2. Ablation analysis of retrieval strategies (RAGAS metrics).

Metric	BM25 Only	Semantic Only	Hybrid (30: 70)	Key Interpretation
Context Precision	0.848	0.942	0.862	Semantic is highly precise; Hybrid remains very competitive.
Context Recall	0.791	0.838	0.895	Hybrid excels at gathering all relevant chunks, minimizing missing information.
Answer Faithfulness	0.550	0.536	0.712	Hybrid yields the highest factual alignment (+29.5% vs BM25; +32.8% vs Semantic), significantly reducing hallucinations.
Answer Relevancy	0.826	0.857	0.844	High relevancy across all models; Semantic performs slightly better.
Token F1 Score	0.488	0.465	0.499	Hybrid exhibits the highest lexical overlap with reference answers.
Average Latency (s)	7.85	13.23	7.64	Hybrid achieves optimal speed, outperforming Semantic Only by 42%.

This performance behavior is due to two critical engineering factors:

Parallel Query Execution: The BM25 search and semantic embedding generation are processed asynchronously in separate threads, minimizing pipeline blockages.

Aggressive Context-Size Optimization: In the Semantic Only configuration, similarity matching must calculate cosine similarities across the entire high-dimensional vector space, which is computationally expensive on local consumer-grade CPUs. The Hybrid configuration utilizes the rapid, linear-time BM25 pipeline as a coarse pre-filtering heuristic, pruning irrelevant segments quickly and reducing both context length and downstream token generation time of the LLaMA model.

Evaluation of Generated Response Quality (NLG Metrics)

To verify the linguistic accuracy of the generated responses against human-validated ground truths, we evaluated the output using standard Natural Language Generation (NLG) metrics across 22 test cases. The results are detailed in **Table 3**.

Per-Category Performance and Error Analysis

The mean Token F1-score of 0.499 and BLEU score of 0.233 are moderate values that require contextual interpretation. In academic chatbot systems, lower lexical overlap does not necessarily indicate a lack of accuracy. Our analysis identified two main contributing factors:

Syntactic Synthesis and Paraphrasing: The LLaMA 3.1 model presents responses in natural, polite Indonesian with structured formatting (bulleted lists, headings), whereas the ground-truth references are concise text extracted directly from administrative PDFs. Lexical metrics penalize this elaboration even when the content is factually correct.

Granular Vocabulary Divergence: For example, a query about thesis seminar registration procedures generated a detailed, formatted multi-step response with clothing requirements and time instructions (Token F1: 0.261, BLEU: 0.009), yet the response was factually accurate (faithfulness: 0.712) and received high user satisfaction. This confirms that lexical metrics are insufficient for evaluating generative systems where elaboration is expected (8).

The evaluation framework also tracked a Mean Robustness Score of 0.467. This metric quantifies the model's structural resilience when handling non-standard student inputs, lexical variations, and minor typographical errors within the Indonesian academic domain. A score of 0.467 indicates that while the local system successfully circumvents catastrophic behavioral failure or hallucination under linguistic noise, localized variations in semantic alignment still occur due to the rigid constraint boundaries of the underlying document corpus. Per-category analysis further reveals a fairness gap across information domains as seen in **Table 4**.

The Token F1 fairness gap (min: 0.325 for Tugas Akhir, max: 0.891 for Dosen) is attributable to structural differences across information types, not retrieval failure. Lecturer data, presented as structured name-title pairs, achieves near-perfect lexical overlap. In contrast, thesis and curriculum categories require multi-document synthesis and naturally produce paraphrased, elaborated responses. This finding is consistent with Ji *et al.* (8) who noted that lexical overlap metrics are insufficient for evaluating generative systems where answer elaboration and paraphrase are expected behaviors.

User-Based Exploratory Usability Testing

To assess user satisfaction, an exploratory usability pilot study was conducted with five informatics students from UPN "Veteran" Yogyakarta (active second-semester and

Table 3. Natural language generation (NLG) and accuracy metrics (Hybrid Retrieval, n = 22).

Evaluation Metric	Mean Value	Std. Deviation	Min Value	Max Value
Token F1-Score	0.499	0.243	0.171	1.000
BLEU Score	0.233	0.282	0.000	1.000
ROUGE-1	0.415	0.242	0.087	1.000
ROUGE-2	0.268	0.259	0.000	1.000
ROUGE-L	0.413	0.243	0.087	1.000
Robustness Score	0.467	0.228	0.114	1.000

Table 4. Per-category token F1 performance (Hybrid Retrieval).

Category	Token F1 (mean)	Interpretation
Dosen (Lecturer)	0.891	Structured name-title data; both retrieval methods handle effectively.
Fasilitas (Facilities)	0.528	Moderate; list-based information with some formatting divergence.
Umum (General)	0.553	Moderate; short factual answers with high lexical overlap.
Peraturan (Regulations)	0.489	Moderate; procedural rules require some paraphrase.
Kurikulum (Curriculum)	0.361	Lower; course listings trigger formatted multi-line output.
Profil (Profile)	0.341	Lower; narrative historical content paraphrased by LLM.
Tugas Akhir (Thesis)	0.325	Lowest; complex multi-step procedures generate elaborated responses.

eighth-semester students). The small-scale pilot focused on identifying immediate usability issues and validating the interface. Given the limited sample size, these results are explicitly exploratory in nature and should be validated with a larger and more diverse participant pool in future work. Respondents interacted with the system and rated performance on a Likert scale (1 to 5). The aggregated results are summarized in **Table 5** (raw individual-level scores are provided in Appendix A).

The pilot evaluation indicated high user acceptance. The highest rating was achieved in Language Naturalness (4.80), confirming the effectiveness of LLaMA 3.1's instruction-following capabilities. The high score for Response Speed (4.79) reflects the real-time usability of the local deployment.

Security, Privacy, and Ethical Considerations

A critical advantage of the proposed local RAG architecture is its alignment with data privacy regulations and security standards. Standard cloud-based chatbot implementations rely on external APIs (such as OpenAI GPT or Anthropic Claude), which require transmitting query logs to third-party servers. In higher education, sending query logs externally can pose risks to student privacy and violate institutional data handling policies.

By deploying LLaMA 3.1 locally via Ollama on institutional hardware, the following benefits are achieved:

Data Containment: Student queries, conversation histories, and internal academic documents remain entirely within the university's local network infrastructure.

Mitigating Malicious Exploitation: Local deployment prevents external telemetry leaks and allows the university to implement strict rate-limiting and custom input-filtering heuristics to prevent adversarial injection and prompt-jailbreaking.

No External API Dependencies: Running the system locally avoids cost overheads and limits downtime risks associated with external API failures, making it a reliable and sustainable option for public universities.

Limitations and Future Outlook

While the hybrid RAG chatbot demonstrates strong performance, we acknowledge several engineering limitations and areas for future development:

Scale of Evaluation and Pilot Usability Testing: The dataset was restricted to 10 administrative PDFs, and user testing was limited to a pilot study of five respondents. Although the documents covered 100% of the active regulations for the department at the time of the study, expanding the knowledge base to encompass university-wide data and conducting larger-scale usability evaluations will be necessary for broader deployment.

Baseline Retrievers: The evaluation focused on standard BM25 and semantic search. Future iterations will evaluate modern dense retrievers (such as ColBERT or cross-encoder rerankers) to determine if they can further improve context precision without introducing significant latency.

Handling Complex PDF Formats: Like many standard RAG pipelines, our text splitter occasionally struggles with multi-column tables, diagrams, and scanned PDFs. Integrating advanced optical character recognition (OCR) and layout-aware table parsers (such as LayoutLM or TableTransformer) would help preserve the structure of complex academic documents.

Informal Language Handling: User feedback indicates the system is not yet fully optimized for handling student abbreviations and non-standard terms, which can affect retrieval quality and should be addressed through query expansion techniques in future work.

Conclusion

This study demonstrates that integrating the LLaMA 3.1 large language model with a Retrieval-Augmented Generation (RAG) framework is an effective approach for providing automated academic services in higher education. By combining BM25 lexical search and semantic vector similarity through a hybrid retrieval strategy with max-score linear scaling, the system achieved a faithfulness score of 0.712 and a context recall of 0.895, outperforming standalone lexical and semantic retrieval baselines. Local deployment via Ollama ensures complete data privacy and eliminates operational costs, while exploratory user testing produced a high satisfaction rating of 4.46 out of 5.00.

Declaration

Author Information

Farel Abid Yasser Prayanto
Department of Informatics, Faculty of Industrial Engineering, Universitas Pembangunan Nasional "Veteran" Yogyakarta .
Contribution: Data Curation, Formal Analysis, Visualization, Writing – Original Draft, Writing – Review & Editing.

Rifki Indra Perwira
*Corresponding author
Department of Informatics, Faculty of Industrial Engineering, Universitas Pembangunan Nasional "Veteran" Yogyakarta.
Contribution: Conceptualization, Funding Acquisition, Methodology, Project Administration, Resources,

Table 5. Aggregated usability pilot evaluation results.

Evaluation Aspect	Average Score	Std. Deviation	Quality Interpretation
Answer Quality	4.16	0.29	Very Good (highly accurate and relevant)
Response Speed	4.79	0.42	Outstanding (highly responsive local system)
Language Naturalness	4.80	0.40	Outstanding (clear, natural, and polite)
Overall Satisfaction	4.46	0.21	Very High (viable for daily academic assistance)

Supervision, Writing – Review & Editing.

Conflict of Interest
The authors declare no conflict of interest.

Data Availability
The datasets generated and/or analyzed during the current study are available in the https://github.com/Farelabid/academic-chatbot repository.

Ethics Statement
Not applicable.

Funding Information
This work received no external funding.

References
1. Bommasani R, Hudson DA, Adeli E, et al. On the Opportunities and Risks of Foundation Models. Published online July 12, 2022. http://arxiv.org/abs/2108.07258

2. Brown TB, Mann B, Ryder N, et al. Language Models are Few-Shot Learners. Published online May 28, 2020. http://arxiv.org/abs/2005.14165

3. Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences. 2023;103:102274. doi: https://doi.org/10.1016/j.lindif.2023.102274

4. Setyorini S. Implementasi sistem informasi akademik berbasis cloud untuk meningkatkan efisiensi administrasi akademik. Jatisi. 2025;12(2). doi: https://doi.org/10.35957/jatisi.v12i2.9227

5. Hwang GJ, Chang CY. A review of opportunities and challenges of chatbots in education. Interactive Learning Environments. 2021;31(7):4099-4112. doi: https://doi.org/10.1080/10494820.2021.1952615

6. Dwivedi YK, Kshetri N, Hughes L, Slade EL, Jeyaraj A, Kar AK, et al. Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management. 2023;71:102642. doi: https://doi.org/10.1016/j.ijinfomgt.2023.102642

7. Tarigan HP. Integrasi Chatbot Berbasis NLP pada Sistem Layanan Akademik Universitas. j.komputer. 2024;3(1):13-18. doi: https://doi.org/10.70963/jk.v3i1.110

8. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of Hallucination in Natural Language Generation. ACM Comput. Surv. 2023;55(12):1-38. doi: https://doi.org/10.1145/3571730

9. AL-Smadi M. QU-NLP at QIAS 2025 Shared Task: A Two-Phase LLM Fine-Tuning and Retrieval-Augmented Generation Approach for Islamic Inheritance Reasoning. Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks. 2025:892-898. doi: https://doi.org/10.18653/v1/2025.arabicnlp-sharedtasks.123

10. Gao Y, Xiong Y, Gao X, et al. Retrieval-Augmented Generation for Large Language Models: A Survey. Published online December 18, 2023. http://arxiv.org/abs/2312.10997

11. Borgeaud S, Mensch A, Hoffmann J, et al. Improving language models by retrieving from trillions of tokens. Published online December 8, 2021. http://arxiv.org/abs/2112.04426

12. Izacard G, Grave E. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021:874-880. doi: https://doi.org/10.18653/v1/2021.eacl-main.74

13. Husain ML, Wibisono Y, Anisyah A. Development of an Academic Services Chatbot Based on Retrieval-Augmented Generation (RAG). Brilliance. 2025;5(2):727-735. doi: https://doi.org/10.47709/brilliance.v5i2.6719

14. Yasmin SM, Fudholi DH. Pengembangan Chatbot Informasi Hukum Layanan Publik Berbasis Retrieval-Augmented Generation Menggunakan LangChain dan OpenAI di Ombudsman DIY. Jurnal Pendidikan dan Teknologi Indonesia. 2025;5(9):2548-2565. doi: https://doi.org/10.52436/1.jpti.995

15. Touvron H, Lavril T, Izacard G, et al. LLaMA: Open and Efficient Foundation Language Models. Published online February 27, 2023. http://arxiv.org/abs/2302.13971

16. Pujiono I, Agtyaputra IM, Ruldeviyani Y. Implementing retrieval-augmented generation and vector databases for chatbots in public services agencies context. jitk. 2024;10(1):216-223. doi: https://doi.org/10.33480/jitk.v10i1.5572

17. Adamopoulou E, Moussiades L. An Overview of Chatbot Technology. Cham: Springer International Publishing; 2020. doi: https://doi.org/10.1007/978-3-030-49186-4_31

18. Wang L, Yang N, Huang X, Yang L, Majumder R, Wei F. Multilingual E5 Text Embeddings: A Technical Report. Published online February 8, 2024. http://arxiv.org/abs/2402.05672

19. Karpukhin V, Oguz B, Min S, Lewis P, Wu L, Edunov S, et al. Dense Passage Retrieval for Open-Domain Question Answering. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020:6769-6781. doi: https://doi.org/10.18653/v1/2020.emnlp-main.550

20. Albert GD, Voutama A. Pengembangan chatbot berbasis pdf menggunakan local retrieval-augmented generation (rag) dan ollama. Jitet. 2025;13(2). doi: https://doi.org/10.23960/jitet.v13i2.6361

21. Es S, James J, Espinosa Anke L, Schockaert S. RAGAs: Automated Evaluation of Retrieval Augmented Generation. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. 2024:150-158. doi: https://doi.org/10.18653/v1/2024.eacl-demo.16

Additional Information

How to Cite

APA 7th Edition: Prayanto, F. A. & Perwira, R. I. (2026). An LLaMA 3.1-Based Chatbot with Retrieval-Augmented Generation (RAG) for Academic Services at UPN "Veteran" Yogyakarta. Digital System and Computing, 2(1), 19-26. https://doi.org/10.58920/dsc0201633

Vancouver: Prayanto FA, Perwira RI. An LLaMA 3.1-Based Chatbot with Retrieval-Augmented Generation (RAG) for Academic Services at UPN "Veteran" Yogyakarta. Digital System and Computing. 2026;2(1):19-26. https://doi.org/10.58920/dsc0201633

Harvard: Prayanto, F. A. & Perwira, R. I. (2026) 'An LLaMA 3.1-Based Chatbot with Retrieval-Augmented Generation (RAG) for Academic Services at UPN "Veteran" Yogyakarta', Digital System and Computing, 2(1), pp. 19-26. doi: 10.58920/dsc0201633

Publisher Note

All claims expressed in this article are solely those of the authors and do not necessarily reflect the views of the publisher, the editors, or the reviewers. Any product that may be evaluated in this article, or claim made by its manufacturer, is not guaranteed or endorsed by the publisher. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access

This article is licensed under a Creative Commons Attribution 4.0 International License. You may share and adapt the material with proper credit to the original author(s) and source, include a link to the license, and indicate if changes were made.