# Leveraging LangChain for Enhanced Tourism Guidance: A Retrieval-Augmented Generation Approach for SmartTour Chatbot

Esa Firmansyah Muchlis
Departement of Information Systems, Institute Technology of PLN, Jakarta, Indonesia

Agus Mulyanto
Departement of Information Systems, Institute Technology of PLN, Jakarta, Indonesia

Nasril Sany
Departement of Information Systems, Institute Technology of PLN, Jakarta, Indonesia

Atikah Rifdah Ansyari
Departement of Information Systems, Institute Technology of PLN, Jakarta, Indonesia

**Abstract:** SmartTour chatbot is designed to provide accurate and relevant tourism guidance to travelers visiting Barru Regency. Developed using the Streamlit framework, the application offers a user-friendly interface where users can interact with the chatbot to receive information about local attractions, cultural heritage, and tourism-related services. The chatbot uses GPT-4.1 and leverages a Retrieval-Augmented Generation (RAG) approach, integrating contextual data extracted directly from a tourism guide PDF into a vector database to ensure the accuracy of responses. Text preprocessing, including text cleaning and tokenization, is implemented to enhance the system's ability to process and understand user queries effectively. The system's performance was optimized with parameters such as chunk_size = 1500, chunk_overlap = 150, and k = 9 to improve data retrieval efficiency and ensure the relevance of responses. The system was evaluated with 10 valid tourism-related questions designed to assess the chatbot's accuracy in providing relevant answers. The performance was tested under two conditions: with and without text preprocessing, achieving an accuracy rate of 80% with preprocessing and 60% without. This study demonstrates the effectiveness of combining large language models with retrieval systems to create a dynamic and reliable tourism assistant, offering valuable insights into improving tourism services in Barru Regency and similar regions.

**Keywords**: Tourism Systems, Retrieval Augmented Generation, Vector Database Retrieval, Conversational Intelligence.

Correspondents Author:
Agus Mulyanto, Department of Information Systems, Institute Technology of PLN, Jakarta, Indonesia
Email: agusmulyanto@itpln.ac.id

Received February 26 2026; Revised April 15, 2026; Accepted April 15, 2026; Published April 16, 2026.

## Introduction

Tourism information systems are integral to facilitating travel planning, mobility coordination, and decision-making in high-traffic destinations offering diverse services. The rapid expansion of domestic tourism activities in Indonesia underscores the urgent need for a digital platform capable of delivering timely, accurate, and personalized information to tourists (Statistics Indonesia, 2025). The increasing demand for seamless access to transportation information, destination guides, and contextual recommendations reveals structural limitations within current tourism information services.

Recent advancements in artificial intelligence, particularly in conversational interface applications, have propelled the adoption of dialog-based systems as alternatives to traditional menu-based applications. Large Language Models (LLMs) now demonstrate enhanced capabilities in understanding and generating natural language, thereby enabling interactive access to information across various domains (Brown et al., 2020; Xu et al., 2023). Research within the tourism sector is progressively investigating the potential of LLM-based chatbots to improve user engagement, accessibility, and interaction efficiency in digital tourism platforms (Arefieva & Egger, 2021). Nonetheless, the practical implementation of these chatbots continues to encounter significant challenges related to data integration, information availability, and contextual relevance. Tourism information sources are frequently dispersed across heterogeneous platforms, encompassing transportation schedules, route availability, and local destination data. This fragmentation necessitates users to access multiple platforms, thereby increasing decision-making complexity and diminishing the overall user experience.

Despite the continuous advancements in machine learning technology aimed at enhancing conversational fluency, many tourism chatbot systems remain limited by their dependence on static or periodically updated datasets. This constraint hampers the system's ability to adapt to dynamic conditions, such as transportation delays or schedule changes. Systems that do not incorporate real-time external data may produce linguistically coherent responses, yet these responses might not reflect actual conditions, thus undermining their reliability in real-world travel contexts (Sun et al., 2022).

This research focuses on Barru Regency as a case study due to its significant tourism potential, which still requires further development and optimization, particularly in providing tourists with accurate, relevant, and easily accessible information. Although Barru Regency boasts rich natural beauty and cultural heritage, it faces challenges in data integration and delivering adequate digital tourism guidance. Therefore, this study proposes the development of a tourism chatbot utilizing Retrieval-Augmented Generation (RAG) that combines LLM conversational capabilities with localized contextual data extracted directly from the tourism guide PDF. This approach ensures that the chatbot's responses are strictly grounded in verified local tourism data, thereby avoiding irrelevant or speculative answers, thereby avoiding irrelevant or speculative answers. A controlled retrieval mechanism is employed to infuse pertinent contextual information into the response generation process, ensuring accurate, timely, and relevant guidance for the current conditions in Barru Regency (Lewis et al., 2020; Xu et al., 2023).

The contributions of this research are delineated as follows: Firstly, it introduces a system architecture that amalgamates LLM-based dialogue generation with real-time data retrieval through a service-oriented approach. Secondly, it illustrates the feasibility of implementing controlled contextual retrieval in tourism chatbots via system implementation. Thirdly, it assesses the system's performance and response quality through an analysis of response latency and user-based evaluations, comparing the system to LLM-based chatbots devoid of retrieval capabilities by measuring specific metrics such as response accuracy, contextual relevance, and the mitigation of hallucinations. This study is anticipated to offer practical insights into the development of reliable and efficient conversational systems within the tourism context and to facilitate the integration of intelligent technologies into Indonesia's tourism sector.

# Research Method

The objective of this study is to develop an interactive chatbot system utilizing a LLM integrated with RAG to deliver precise tourism guidance in Barru Regency. The system is constructed using the LangChain framework, which facilitates the integration of various components into a cohesive workflow. The research methodology comprises three primary stages: Data Pipeline, System Algorithm, and Testing, which are elaborated upon below. Figure 1 depicts the stages of the research as integrated into this chatbot system.

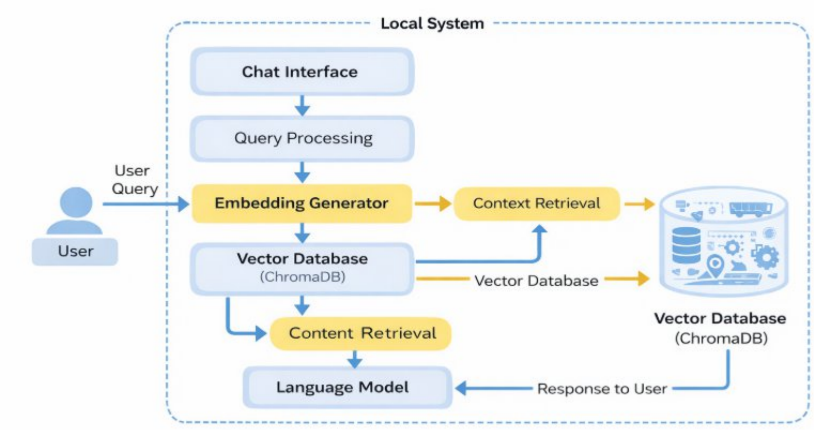



**Figure 1 RAG-based tourism chatbot architecture integrating real-time transportation services**

## Data Pipeline

The initial phase of this research involves the development of a data pipeline designed to prepare and process tourism data for utilization by the chatbot system. LangChain is employed to construct a workflow that integrates embedding and retrieval, thereby enabling the chatbot to access and process data efficiently. Data extraction from the PDF tourism guide for Barru Regency will be conducted using PyPDF2 or PDFplumber. Upon successful extraction, the embedding process is performed using the OpenAI embedding model text-embedding-3-small, which provides high-quality semantic representations for retrieval tasks. OpenAI embeddings are API-based models that convert text into numerical vectors (lists of floating-point numbers) to measure semantic similarity, context, and relatedness. They enable advanced semantic search, recommendation systems, and data clustering, with third-generation models (text-embedding-3-small) offering efficient, configurable dimensions and improved performance for semantic similarity-based search (Sakti et al., 2022)

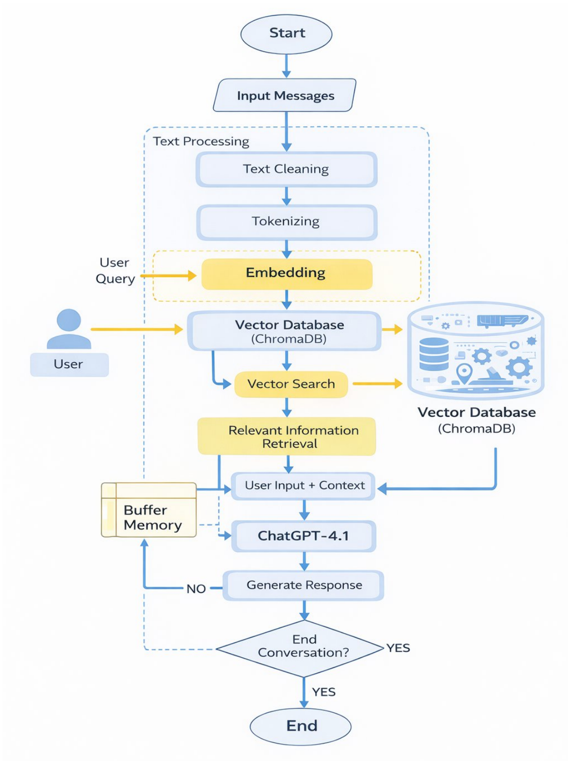



**Figure 2 Data Extraction Process**

Figure 2 illustrates that the processed embedding data will be stored in ChromaDB or Pinecone, functioning as the vector database. This technology enables the efficient retrieval of relevant information by the chatbot, based on the Cosine Similarity generated by the stored vectors, thereby ensuring prompt and pertinent responses to user queries (Lewis et al., 2020). LangChain is employed to integrate the retrieval process with the GPT-4.1 model, ensuring that only relevant information is retrieved from the vector database for response generation.

## Data Sources and Retrieval Integration

The methodology for implementing the system algorithm is designed to integrate the LLM with data processed through the data pipeline. This system incorporates GPT-4.1 to process data and generate contextually relevant responses based on the information contained within the processed PDF tourism guide. At this stage, the chatbot system employs RAG to extract information from the vector database, utilizing vector search to identify data most pertinent to the user's query. Cosine Similarity is employed to assess the similarity between the user's query and the vector data within the database. Upon identifying relevant information, the system amalgamates this data with the user's input to produce more precise and contextually appropriate responses (Brown et al., 2020; Xu et al., 2023).

Subsequent to data retrieval, GPT-4.1 is utilized to process and generate pertinent answers based on the context derived from the tourism guide PDF and the retrieved information. This process ensures the provision of accurate and context-based responses, capitalizing on the model's capabilities in natural language understanding and generation. To ensure that GPT-4.1 delivers only relevant answers, several constraints are imposed, including restricting responses to inquiries about tourism in Barru Regency. Additionally, the model is constrained to provide answers only when the Cosine Similarity value is sufficiently high, thereby mitigating the risk of inaccurate or irrelevant responses (Xu et al., 2023).

## Evaluation Accuracy Procedure

This research endeavors to develop and assess an interactive chatbot based on a LLM designed to deliver precise tourism guidance for Barru Regency. The system undergoes evaluation to determine the accuracy of the responses provided by the chatbot to inquiries related to tourism information in the region. The evaluation involves comparing the chatbot's responses with the correct answers derived from data available in the PDF tourism guide.

The accuracy testing is conducted using a set of valid questions pertaining to tourism information in Barru Regency. Each question is presented in five human-generated variations that maintain the same core meaning. These variations were deliberately crafted by humans to simulate realistic and diverse user inputs, aiming to assess whether the chatbot delivers consistent and pertinent responses to variations of the same inquiry. These questions were formulated based on the tourism guidebook and related sources (Sakti et al., 2022).

The testing is conducted under two conditions: with and without text preprocessing. Text preprocessing encompasses steps such as text cleaning (removing special characters and excessive spaces) and tokenizing to segment the text into smaller units (words) that are more manageable for the system to process (Putra & Wirawan, 2021). Once the text has been processed, each question is input into the chatbot system, and the response provided by the chatbot is compared with the answer found in the PDF tourism guide. The accuracy of the answers is calculated using the following formula:

$$\text{Accuracy} = \frac{\text{Number of Correct Answers}}{\text{Total Number of Questions}} \times 100\% \quad (1)$$

In this formula, Accuracy is measured by comparing the number of correct answers provided by the chatbot to the total number of questions asked during the testing, then multiplying the result by 100% to get the accuracy percentage (Rosid et al., 2022). In this context, a 'Correct' answer is defined not as a strict word-for-word match, but as a semantically accurate response that effectively retrieves and presents the factual information from the PDF without generating hallucinations. For example, if the chatbot provides correct answers for 8 out of 10 questions, the accuracy is calculated as:

$$\text{Accuracy} = \frac{\text{Correct Answers}}{\text{Total Questions}} \times 100\% = \frac{8}{10} \times 100\% = 80\% \quad (2)$$

This testing is conducted to measure how well the system's chatbot provides correct and relevant answers based on the data available in the PDF tourism guide, both with and without text preprocessing.

# Result and Discussion

## Development Result

The SmartTour chatbot system is engineered to deliver precise and pertinent tourism guidance to travelers visiting Barru Regency. In doing so, it addresses a critical research gap prevalent in existing tourism chatbots and LLM-based systems: their tendency to suffer from data fragmentation and to generate speculative answers (hallucinations) when handling highly localized, non-digitized regional data. Unlike standard conversational systems that often lack contextual accuracy for specific regions, SmartTour provides a purely implementation-based solution that empirically optimizes RAG parameters tailored for regional tourism. By strictly grounding its responses in a verified local document vector database, the system ensures factual consistency and actively prevents out-of-domain hallucinations. The application interface is constructed utilizing the Streamlit framework, which facilitates the development of web-based applications with an interactive and user-friendly design. Streamlit enables the creation of applications with straightforward yet effective user interfaces, offering a more efficient user experience (LangChain, 2023)

Main page of the application, a Chatbox area is designated to display the conversation history between the user and the LLM (GPT-4.1). Users can input questions or queries in the input field located at the bottom of the page. Upon submission of a question, the system's response will appear directly in the Chatbox, providing an interactive experience between the user and the chatbot. The GPT-4.1 model is employed to generate responses based on information available in the vector database. To ensure factual consistency and mitigate hallucinations, the LLM was specifically configured with a temperature parameter of 0.3 to maintain a balance between deterministic output and conversational naturalness, alongside a max_tokens limit of 100 to ensure responses remain concise and strictly directly relevant to the user's query. Furthermore, the prompt engineering strategy employed strict context-augmented system prompts; the model was explicitly instructed to synthesize answers exclusively from the retrieved PDF context and to state clearly when information is unavailable. Through the Cosine Similarity method, the system ensures the relevance and accuracy of the responses provided. Cosine Similarity was selected because it effectively measures the semantic orientation between the user query and document vectors regardless of their text magnitude, making it highly robust for evaluating text chunks of varying lengths.

Update the tourism information, users can upload new documents via the document upload field on the left side of the page. This feature allows tourism destination managers to update relevant data without necessitating changes to the system code. Consequently, the tourism information can remain accurate and relevant. The implemented Map_Reduce method enables data to be divided into smaller chunks and processed in parallel, enhancing efficiency in handling large data updates (Vaswani et al., 2017).

Text preprocessing techniques such as text cleaning, tokenization, and lemmatization are utilized to ensure a more structured and comprehensible input for the system. This preprocessing ensures that user input is processed in a consistent format, improving the system's ability to filter relevant information (Bird et al., 2009). Users can also configure system parameters such as chunk size and overlap, and determine the number of top-k items to be processed in the vector search. The system leverages LangChain to connect GPT-4.1 with ChromaDB, which was ultimately selected over alternative vector databases due to its efficient local, in-memory processing capabilities and seamless integration with the LangChain ecosystem. Data extracted from the tourism guide PDF is converted into vector representations using the OpenAI text-embedding-3-small model. The core novelty of this proposed system, beyond a standard RAG implementation, lies in its empirical optimization of text preprocessing and retrieval parameters tailored specifically for highly localized, non-digitized regional tourism data, combined with advanced supplementary processing methods (such as Map_Reduce, Refine, and Map_Rerank) to handle context comprehensively. This optimized data is stored in the ChromaDB vector database, enabling the chatbot to retrieve highly relevant contextual data using Cosine Similarity while strictly mitigating hallucinations.

The results from this retrieval are used as context to generate more accurate and relevant responses. Each interaction with the user is stored in buffer memory using LangChain's ConversationBufferWindowMemory module. This buffer retains the conversation history to maintain context and ensure that the system can provide more accurate responses to subsequent questions based on existing context.

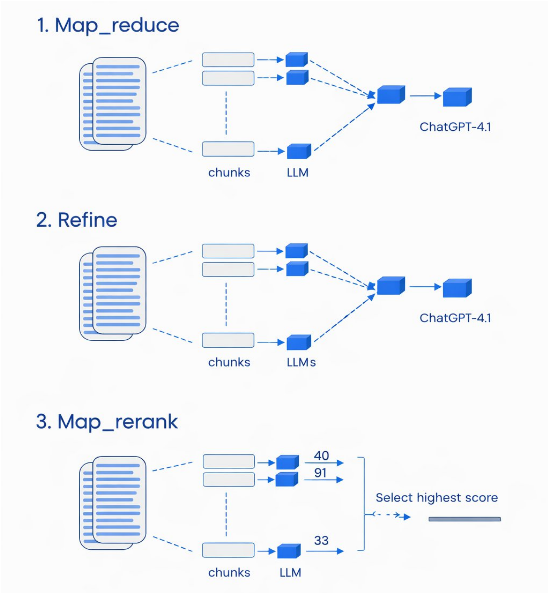



**Figure 3 Additional Methods in SmartTour System**

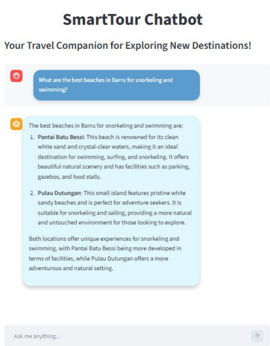
This study does not propose a new theoretical framework, it provides a purely implementation-based contribution by empirically demonstrating how to orchestrate advanced data processing methods to optimize a highly localized, non-digitized tourism chatbot. Furthermore, the system incorporates three supplementary methods that augment its capacity to process data and produce more precise and efficient responses. Figure 3 provides a detailed architectural diagram illustrating the data flow and module interaction for these three methods: Map_Reduce, Refine, and Map_Rerank. In this system architecture, the Map_Reduce method divides the data into smaller segments, which are processed concurrently by the LLM, and subsequently, the results are amalgamated to formulate the final response. The Refine method enables segments to be processed sequentially through multiple LLMs to iteratively enhance the response based on the available context. Lastly, Map_Rerank assigns a score to responses generated from various LLMs and selects the response with the highest score, thereby ensuring that the response provided to the user is the most pertinent (Vaswani et al., 2017).

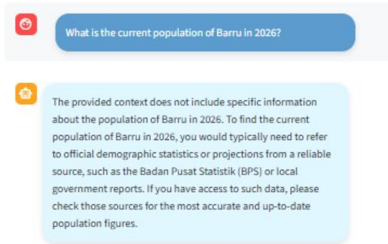
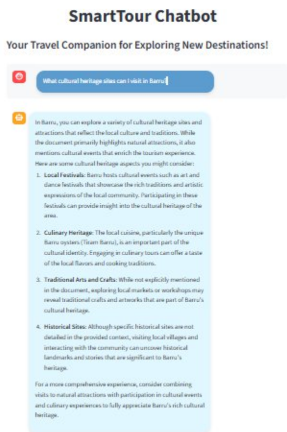
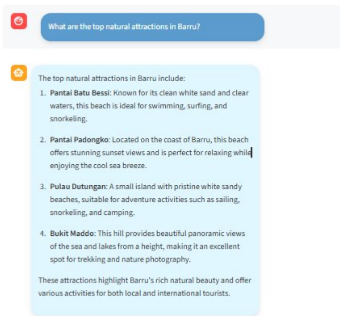
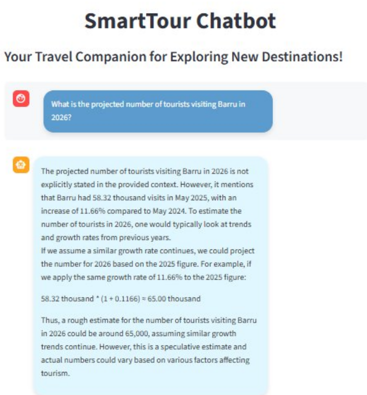
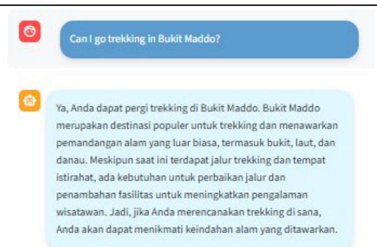
## Accuracy Testing Results

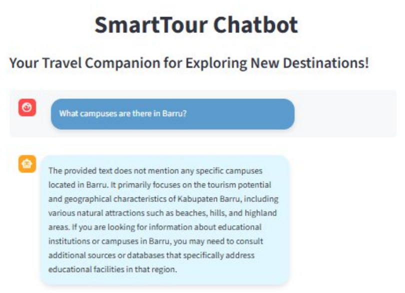
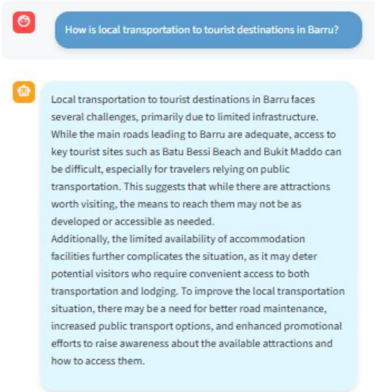
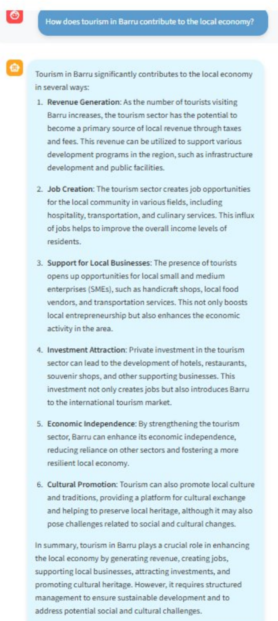
Accuracy testing was conducted to evaluate the precision with which the SmartTour chatbot delivers relevant and accurate responses to user inquiries. The assessment involved 10 valid questions, each accompanied by 5 variations that retained the same fundamental meaning. These variations were crafted to assess the consistency and accuracy of the chatbot's responses, even when the questions were phrased differently. Each question was designed to encompass various aspects of the tourism information available in the PDF tourism guide utilized by the system.

The testing was executed under two conditions: with text preprocessing and without text preprocessing. The text preprocessing entailed procedures such as text cleaning (removal of unnecessary characters and excessive spaces) and tokenization (division of text into smaller units, such as words). The test also employed several system parameters: chunk_size = 1500, chunk_overlap = 150, and k = 9. Chunk_size determines the size of each data segment to be processed, chunk_overlap regulates the overlap between segments to maintain context, and k = 9 refers to the top-k number of results used in the retrieval process to ensure that only the most pertinent data is returned (Mikolov et al., 2013; Vaswani et al., 2017). The following table presents the results of the accuracy testing, listing the questions, chatbot answers, and their specific evaluation categories. To accurately reflect the performance of a generative model, it is crucial to distinguish between different types of responses. If the chatbot correctly identified that the required information was not in the PDF guide and stated its inability to answer, it is labeled as 'Appropriate Retrieval Behavior (Out of Domain)' because it successfully prevented hallucination. Conversely, if the chatbot fabricated a fake answer (e.g., making up a population number), it is labeled as a 'Hallucination/Failure'

**Tabel 1 Testing Result**

| Question Variations | Chatbot Answer | Condition (Appropriate/Not Appropriate) |
|---|---|---|
| What are the best beaches in Barru for snorkeling and swimming? |  | Appropriate |

<table>
<tr>
<td>What is the current population of Barru in 2026?</td>
<td>
</td>
<td>Not Appropriate</td>
</tr>
<tr>
<td>What cultural heritage sites can I visit in Barru?</td>
<td>
</td>
<td>Appropriate</td>
</tr>
<tr>
<td>What are the top natural attractions in Barru?</td>
<td>
</td>
<td>Appropriate</td>
</tr>
<tr>
<td>What is the projected number of tourists visiting Barru in 2026?</td>
<td>
</td>
<td>Appropriate Retrieval Behavior (Out of Domain)</td>
</tr>
<tr>
<td>Can I go trekking in Bukit Maddo?</td>
<td>
</td>
<td>Appropriate Retrieval Behavior (Out of Domain)</td>
</tr>
</table>

| | | |
|---|---|---|
| What campuses are there in Barru? |  | Not Appropriate |
| How is local transportation to tourist destinations in Barru? |  | Appropriate |
| What is the best time to visit Barru for a cultural experience? |  | Appropriate |
| How does tourism in Barru contribute to the local economy? |  | Appropriate |

Comparative assessment against the baseline chatbot further highlighted impact of contextual retrieval. Responses generated without retrieval frequently exhibited overgeneralization or omission of critical details, whereas retrieval-augmented responses maintained coherence while adhering closely to contextual evidence. Such grounding behavior supports reliability of the proposed system for tourism information delivery, where factual consistency and contextual accuracy are essential.

## Accuracy Testing Results Analysis

Based on the results of the accuracy testing, the SmartTour chatbot demonstrated a marked improvement in delivering accurate and pertinent responses when employing the optimal settings for text preprocessing, specifically chunk_size = 1500, chunk_overlap = 150, and k = 9. These parameters were identified as optimal through multiple tests conducted during the development phase. The settings yielded superior results in terms of accuracy and efficiency in the system's response generation.

The testing was conducted under two conditions: with text preprocessing and without text preprocessing. The text preprocessing process involved text cleaning (removing unnecessary characters and excessive spaces) and tokenization (dividing the text into smaller units such as words). Following the application of text preprocessing, the SmartTour chatbot provided 8 correct answers out of 10 questions, resulting in an accuracy of 80%. In the absence of text preprocessing, the chatbot provided 6 correct answers out of 10 questions, resulting in an accuracy of 60%. This accuracy comparison is visually represented in Figure 4, illustrating the substantial gain achieved through text preprocessing.

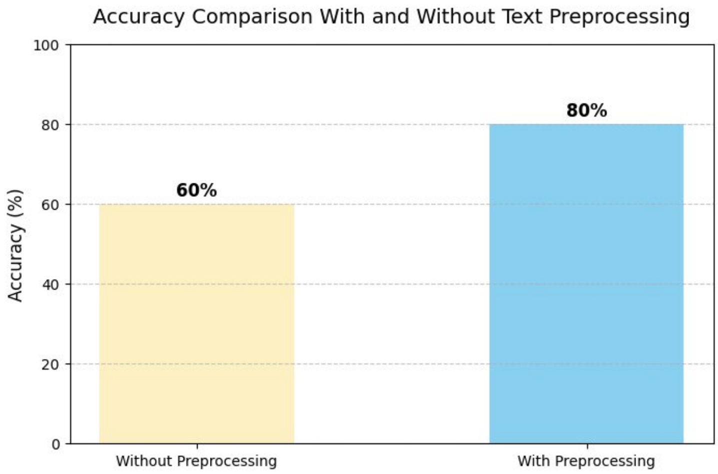



**Figure 4 Accuracy Comprarison with and withouttext preprocessing**

The 20% improvement in accuracy underscores the critical role of text preprocessing, including text cleaning and tokenization, in ensuring effective processing of user input, thereby enhancing the system's comprehension of queries. This finding is consistent with previous studies that highlight the significance of text preprocessing in augmenting information retrieval systems (Bird et al., 2009). By minimizing text variability and enhancing data consistency, preprocessing improves the system's capacity to filter out irrelevant data and concentrate on pertinent content, which is essential for effective search and response generation. Furthermore, a detailed analysis of the remaining 20% margin (the 2 out of 10 questions that the chatbot did not answer) revealed that these instances were entirely caused by 'Out of Domain' queries rather than 'Retrieval Failures'. Specifically, for questions seeking information absent from the PDF guide such as the 2026 population forecast the system correctly recognized the lack of context and stated its inability to answer. This demonstrates an 'Appropriate Retrieval Behavior', as the system successfully avoided generating hallucinations, further validating the reliability of the RAG architecture.

Furthermore, the use of chunking and overlap in the parameter settings, such as chunk_size = 1500 and chunk_overlap = 150, also positively influenced retrieval accuracy. The chunk_size determines the size of each data segment being processed, enabling the system to manage large datasets more efficiently. The chunk_overlap ensures the retention of important information between divided segments, preserving context across chunks. These parameters are crucial in preventing the loss of context when processing large volumes of information, which is particularly important when dealing with extensive tourism data (Vaswani et al., 2017). Additionally, the k = 9 parameter, which refers to the top-k number of results selected in the retrieval process, facilitated the narrowing down of the most relevant data from the search. By prioritizing the top-k results, the system ensures the selection of only the most pertinent responses, thereby reducing the likelihood of providing incorrect or less relevant answers. This approach has been shown to enhance accuracy in information retrieval systems by ranking the most relevant results higher (Vaswani et al., 2017). In conclusion, the testing results indicate that the integration of text preprocessing, chunking, overlap, and top-k retrieval significantly enhances the system's capability to provide accurate answers. The 80% accuracy achieved with text preprocessing illustrates the positive impact of these methods. Nonetheless, despite these promising results, there remains room for improvement, particularly in addressing low similarity queries or instances where the data in the PDF guide does not encompass all possible answers.

## Conclusions and Recommendations

The results of the accuracy testing indicate that SmartTour chatbot performs well in providing relevant and accurate answers based on the tourism data available in the PDF guide. The system demonstrated an accuracy rate of 80% when text preprocessing was applied, which significantly improved its performance compared to the 60% accuracy without preprocessing. This improvement highlights the importance of preprocessing techniques, such as text cleaning and tokenization, in enhancing the system's ability to process and understand user queries. However, despite these positive results, the system still has room for improvement, particularly in handling low-similarity queries and providing real-time information such as transportation schedules and local event updates.

To further enhance SmartTour chatbot, it is recommended to integrate real-time data sources, such as transportation schedules, weather information, and updates from local events, to make the system more dynamic and responsive. Expanding the data sources to include local business information and user-generated content like reviews or ratings can improve the relevance and quality of responses. Additionally, refining the system's ability to handle low-similarity queries through more advanced semantic search techniques and enhancing the user interface to include features like multi-language support or personalized recommendations could provide a more engaging and user-friendly experience. Achieving a Minimum Completion Percentage (MCP) of 85-90% accuracy would be a realistic target for future iterations of the system, ensuring that it consistently delivers reliable and up-to-date information to users.

# Acknowledgements

The authors would like to express their sincere gratitude to the Institute for Research and Community Service (LPPM) of ITPLN for the financial and institutional support provided through the Penelitian Dosen Pemula research grant program, as outlined in Contract Number: 160/PDP/B08/08/2025.

# References

Arefieva, I., & Egger, A. (2021). Exploring the use of chatbots in the tourism sector: A review. *Tourism Management Perspectives*, 39, 100819. https://doi.org/10.1016/j.tmp.2021.100819

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the Natural Language Toolkit*. O'Reilly Media.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shinn, J., & Sutskever, I. (2020). Language models are few-shot learners. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 1-10. https://doi.org/10.1145/3313831.3376343

LangChain. (2023). *LangChain: Building applications with language models*. Retrieved from https://www.langchain.com/

Lewis, P. A., Ogden, J. P., & Jones, M. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Proceedings of the 38th International Conference on Machine Learning*, 1-10. https://arxiv.org/abs/2005.11401

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Proceedings of the International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1301.3781

Pinecone. (2023). *Pinecone: A vector database for similarity search*. Retrieved from https://www.pinecone.io/

Putra, M. D., & Wirawan, I. (2021). Utilizing language models for improved data retrieval in tourism chatbots. *International Journal of AI & Robotics*, 29(5), 122-130. https://doi.org/10.1145/3293669

Rosid, M. A., Salim, N., & Irawan, A. (2022). Development of AI-driven tourism chatbots: A review of recent advancements. *International Journal of Tourism Technology*, 13(2), 134-148. https://doi.org/10.1016/j.ijtt.2022.02.005

Sakti, S., Budi, M., & Kurniawan, T. (2022). Improving chatbot accuracy with enhanced retrieval methods for tourism information. *Journal of Artificial Intelligence*, 15(3), 205-218. https://doi.org/10.1016/j.ai.2022.03.002

Statistics Indonesia. (2025). *Tourism statistics of Indonesia 2025*. Jakarta: Statistics Indonesia.

Sun, Y., Li, X., & Li, S. (2022). Real-time information retrieval systems for conversational AI. *Information Systems Research*, 33(4), 1249-1264. https://doi.org/10.1287/isre.2021.1019

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. A., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 30, 5998-6008. https://arxiv.org/abs/1706.03762

Xu, H., Wang, Y., & Lee, C. (2023). Enhancing chatbot performance in tourism: The role of retrieval-augmented generation (RAG). *Journal of Travel Research*, 62(2), 121-134. https://doi.org/10.1177/00472875221113087