Assessment of Retrieval and Generative Chatbots in Tourism Information Service

Sharfina Febbi Handayani[1], Dairoh[2], Dwi Intan Af'idah[3*]

[1,2,3] *Informatics Engineering, Politeknik Harapan Bersama, Indonesia*

*corr-author: dwiintanafidah@poltektegal.ac.id

Abstract - Chatbots are essential for improving the customer experience on tourism websites, especially when it comes to arranging travel and offering precise information. The purpose of this study is to evaluate the effectiveness of generative and retrieval-based chatbots in the tourism information service. Two retrieval-based models are MLP-based single QA and multi QA and two generative-based models namely LLaMA 2 and GEMMA were evaluated using confusion matrix, BLEU score, response correctness and response naturalness. The study found that LLaMA 2 outperformed other models, with the highest response Accuracy of 0.89, naturalness of 0.75, and BLEU score of 0.33. GEMMA received the lowest score, suggesting that it has trouble coming up with precise and organic answers. The retrieval-based models showed strong accuracy but were less natural in their responses. The ease of dataset creation for generative models, which only requires narrative text, further positions LLaMA 2 as the most suitable option for improving user experience in the tourism services.

Keyword: Chatbot; GEMMA; LLaMA2; Tourism.

I. INTRODUCTION

Technological advancements have led to the emergence of a new paradigm in tourism known as Tourism 4.0, distinguished by its rapid growth and widespread adoption. In this context, smart technologies play a central role, seamlessly integrating with online information sources to enhance the overall visitor experience and streamline the travel planning process [1-3]. Recognizing the tourism potential of Tegal Regency, a website-based virtual tour Tegal Tourism application was developed. Tegal, an Indonesian city in Central Java, has a lot to offer travelers in terms of historical landmarks, scenic landscapes, and distinctive cultural events. This website serves as an encouragement for Tegal tourism [4]. Utilizing the newest technologies is essential to improving Tegal tourism, especially considering the destination's growing popularity.

Artificial intelligence (AI) has significantly transformed daily human activities through the development and evaluation of advanced systems and applications, commonly referred to as intelligent agents, which are designed to perform diverse functions [5]. Among these advancements, chatbots, as AI-driven conversational agents, have gained widespread adoption across various industries [6]. Previous study by [7] demonstrates chatbot performance for student admission services by integrating RNN and Decision Tree algorithms. This approach underscores the effectiveness of combining machine learning and natural language processing techniques to improve the quality of user interactions in structured and educational contexts. The smart tourism sector, in particular, has experienced profound advancements, driven by the needs of increasingly connected and demanding travelers. Modern tourists seek not only memorable travel experiences but also highly personalized, responsive, and convenient services, reflecting the evolving expectations of the digital era [8]. In the ever-evolving digital era, chatbot technology has become an essential tool in the tourism industry. Chatbots, as AI-based applications, offer a range of functions, from providing general information to answering specific questions in real-time [9]. They can enhance the user experience by providing instant assistance and support that is accessible at any time.

Tourism websites are increasingly adopting chatbots to enhance user experience and provide instant information to visitors [10]. The previous research explores BiLSTM models to analyze sentiment in Indonesian tourist attraction reviews, the findings of this study are particularly relevant for enhancing chatbot interactions in the tourism sector by integrating sentiment-aware responses, supporting the broader goal of improving user experience in tourism information systems [11]. The article conducts a comparative analysis of the existing online application and the anticipated functionality of the website following the integration of the chatbot. By examining the performance measurements, the web application and chatbot demonstrated a 20% boost in performance and an improvement of 5% in accessibility [12]. This is an important reason for implementing chatbots on Tegal

tourism websites. Tegal tourism website is a virtual tour application that allows users to explore tourist attractions virtually[4]. Besides that, previous studies have shown that chatbots have great potential to enhance user experience in the tourism sector [13].

Rule-based, retrieval-based, and generative-based paradigms are three kinds of models that are applied by the chatbot response module to generate responses [5]. Rule-based chatbots utilize pattern-matching algorithms to analyze user input and match it to predefined rule patterns, enabling the selection an appropriate response from a set of predetermined options. The selection of the rule and the corresponding response format can be influenced by the specific context or circumstances of the interaction. The Rule-based methodology chooses a response from a list of guidelines. The pattern-matching technique lacks the spontaneity of human reactions, as it is automated and repetitive[14]. Artificial Neural Networks (ANNs), a subset of machine learning algorithms, have gained significant prominence in recent years, driven by advancements in processor technology. These networks are increasingly utilized in the development of conversational agents, particularly for tasks involving input and output sequences of variable length [15].

A retrieval-based chatbot application related to healthcare was successfully developed in this research [16]. The system utilizes neural networks, specifically the Multilayer Perceptron (MLP) model, to predict diseases and recommend appropriate medications to users. However, it does not engage in additional queries to provide a comprehensive understanding of the illness. In contrast, the proposed chatbot is designed to overcome these limitations, offering a more effective solution for illness diagnosis. A key distinction lies in the capabilities of retrieval-based and generative-based chatbots: while retrieval-based models are constrained to responding with pre-existing outputs that closely match user inputs, generative-based models can dynamically generate new and contextually relevant content [15].

Previous research [17] indicates that generative-based approaches are constrained in supporting free-form dialogue due to their reliance on predetermined outputs, whereas retrieval-based approaches enable chatbots to provide more contextually meaningful responses. Chatbots employing decision tree mechanisms, often used in choice-based formats, are less suitable for handling multi-linear conversations, as these structures limit flexibility. Furthermore, improving system usability is challenging in such configurations, as the system fails to complete tasks when user inputs do not match the information in its database. Generative-based chatbots, on the other hand, are trained on extensive datasets and utilize this knowledge to generate responses dynamically, rather than depending on predefined outputs. Studies have demonstrated that encoder-decoder architectures used in generative-based chatbots achieve an accuracy of 94.45%, representing a notable advancement in chatbot performance.

Another study [18] approaches generative-based chatbot depending on how they generate responses. The encoder-decoder framework, which is built on GRU cells, was used in the development of the generative-based chatbot, which produces better results with LSTM in terms of accuracy and loss. This study also finds that the suggested framework can give consumers more information about the perceived cultural things they require. With the use of a conversational agent built on the Seq2Seq paradigm, this research architecture aims to offer a number of services to travelers. However, further research needs to be conducted using more powerful methods.

Large Language Models (LLMs) like the Large Language AI Meta Model (LLaMA) [19] developed by Meta, the GPT Family [20] developed by OpenAI, and Gemini [21] developed by Google, among others, are examples of how quickly the field of NLP is revolutionizing. Renowned for their skill in identifying the semantic connections between words and sentences, which contributes to the advancement of chatbot technology [22]. Numerous studies on generative chatbots utilizing the GEMMA (Google Exploratory Models for Multitask AI), and LLaMA model have been spurred by this [21][23][24].

Large language models (LLMs) have gained significant popularity, with the emergence of open-source variants providing users with more flexible and secure alternatives. This study utilized several open-source models, including Mixtral, LLaMA 2, LLaMA 3, GEMMA, and Qwen2. Experimental findings indicate that while LLaMA 2 and GEMMA demonstrate potential, they still require substantial improvements to fully realize their capabilities [25]. These observations have sparked increased interest in customizing LLMs for specific domains, such as biomedical applications. The technique of instruction tuning has been extensively employed to align LLMs with natural language instructions, enabling their effective application to real-world tasks [24].

The issue with this study is a gap in the existing literature regarding the indirect comparison between generative and retrieval chatbots in the context of Tegal tourism information service. There is an urgent need to compare the two types of chatbots to determine which is

more effective in enhancing the Tegal tourism information service and providing relevant information. Although chatbots have already proven useful in tourism, the main challenge is selecting the most suitable type of chatbot.

This study aims to analyze the performance differences, user satisfaction, and information accuracy between generative and retrieval chatbots. By analyzing the interactions and performance of both types of chatbots, this study hopes to provide in-depth insights into the strengths and weaknesses of each, and offer data-driven recommendations to improve user experience on the Tegal tourism information service.

This study provides a comprehensive comparison of generative and retrieval chatbots, contributing to the broader understanding of their implementation and applications in tourism information services. While the research focuses on Tegal tourism as a case study, the findings offer generalizable insights applicable to other regions and sectors within the tourism industry. These insights are expected to assist technology developers and destination managers in designing chatbot solutions that improve information quality and enhance user experiences across diverse contexts. By incorporating a global perspective in the discussion and conclusion, this study advances the understanding of chatbot technologies and establishes a framework for their adoption and development in the global tourism sector, increasing its relevance and scientific impact.

II. METHOD

The methodology of this research is designed to evaluate and compare two chatbot approaches includes Multilayer Perceptron (MLP)-based retrieval model and a generative model such as LLaMA 2 and GEMMA. MLP-based, namely MLP using Single Question Answer (QA) in dataset and MLP using Multi Question Answer (QA) in dataset. This study encompasses four main stages: dataset creation, preprocessing, model development, and testing as illustrated by Fig. 1.

A. Dataset Creation

The creation of a dataset is a fundamental initial step in chatbot development. The dataset consists of two types, one for the MLP-based retrieval chatbot and the other for the generative chatbot based on LLaMA 2 and GEMMA. The dataset for the retrieval-based chatbot contains question-and-answer text, while the dataset for the generative-based chatbot contains descriptive narrative text.

This dataset was collected through interviews with tourism managers in Tegal, Indonesia, covering 11 destinations. The information gathered includes descriptions of tourist attractions, facilities, operating hours, ticket prices, transportation, activities, and history. The data is organized in JSON format, structured with tags, patterns, and responses, where each tourist attraction is represented by categories such as description, facilities, and others.

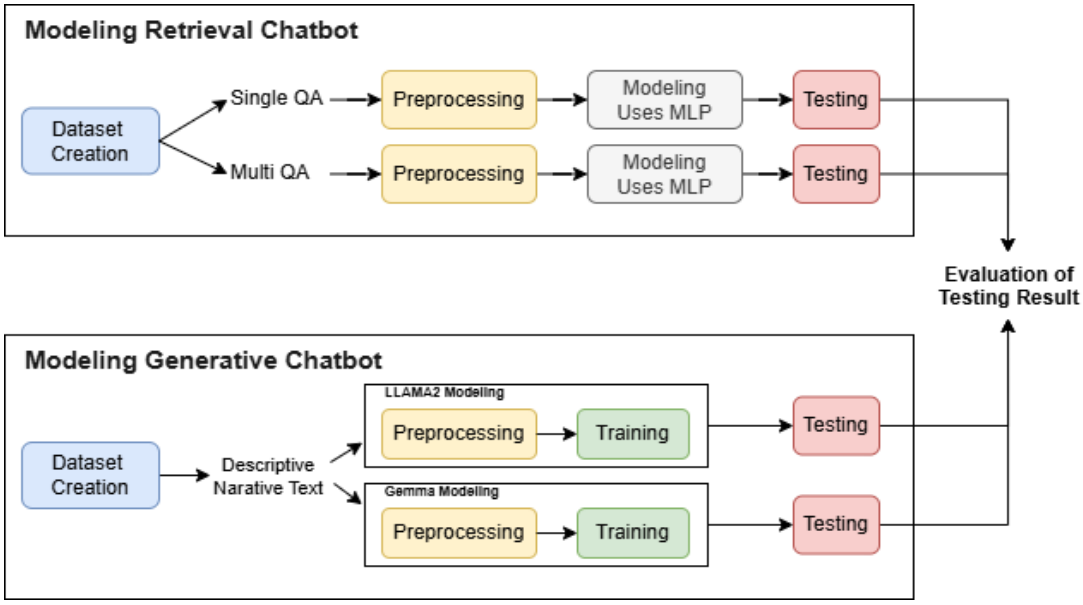

Fig. 1 Research procedure

Tags serve as labels or categories that classify specific types of information, such as 'description,' 'facilities,' 'operating hours,' 'ticket prices,' and other relevant categories that help organize data. Patterns refer to examples of questions or statements that users might pose related to each tag, guiding the chatbot in recognizing and responding appropriately. Responses are the answers prepared by the chatbot to address the corresponding patterns, ensuring that the information provided is relevant and accurate based on the user's input. Datasets for MLP-based chatbots are divided into single QA (Question Answer) and multi QA (Question Answer). In the single QA dataset, each question topic (as tag) has one question (as pattern) and one answer (as responses). Meanwhile, in the multi QA dataset, each topic is represented by several tags and several responses.

The generative dataset contains more extensive textual information about tourist attractions in Tegal. This information includes detailed narratives about the tourist sites, which are necessary for training the generative model to produce more contextual and informative responses.

B. *Preprocessing*

Preprocessing for the retrieval chatbot (MLP-based) involves several essential stages. The process begins with case folding, where all text is converted to lowercase to standardize the format and minimize variations in word forms. Following this, filtering is performed to remove non-alphabetic characters, which helps reduce noise and enhance data quality. Next, tokenization is carried out to break the text into tokens or units, typically separating the text into individual words. Finally, lemmatization is applied to convert words to their base forms, thereby reducing variations of words that have the same meaning.

Preprocessing for generative chatbot using LLaMA 2 and GEMMA begins with text normalization, which involves removing punctuation, symbols, numbers, and other irrelevant characters. This is followed by tokenization and the creation of a specialized vocabulary for LLaMA 2 using Byte-Pair Encoding (BPE), which groups the most frequently co-occurring bytes to generate the most efficient tokens. In contrast, GEMMA employs a tokenizer tailored to its model architecture, capable of preserving semantic meaning [21] [26]. Subsequently, the dataset is organized into sequences that reflect the narrative context. Finally, unnecessary information such as tags or non-verbal data, as well as any duplicate entries, is removed.

C. *Model Development*

At this stage, the architecture for both chatbot models, namely the retrieval-based chatbot and the generative chatbot, will be designed and implemented. This explanation covers the various key components that interact to generate relevant and informative responses to users.

The architecture of the retrieval MLP-based chatbot is designed to retrieve the most relevant response from a predefined response archive based on user input. This architecture consists of several components as illustrated by Fig. 2. The user input is the initial point where users submit questions or statements. This input can be either short or long text containing requests for information about tourist attractions in Tegal. Context plays a crucial role in understanding the meaning and intent of the user input. The output from the preprocessing stage is used to form a context representation, which helps the model select the most relevant response. This context representation typically involves linguistic features extracted from the user input, such as key terms, sentence structure, and recurring patterns in the conversation [16].

The response archive is a collection of all pre-prepared answers for various question patterns. These responses are arranged in a way that allows the model to access them easily based on the context derived from the user input. The retrieval-based model uses an Artificial Neural Network (ANN) architecture to match user input with the most appropriate response from the response archive. This matching process involves several layers of the artificial neural network, where user input is converted into feature vectors that are then compared with the feature vectors of the available responses. The best match based on vector similarity is selected as the output. Finally, once the most relevant response is identified by the retrieval model, it is sent back to the user. This response is tailored to the given context and is expected to meet the user's informational needs accurately [5].

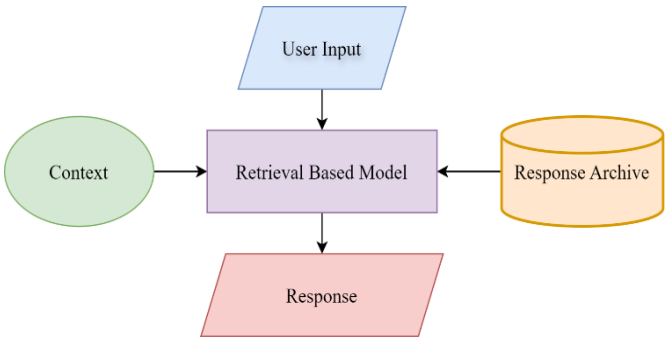

Fig. 2 Architecture of retrieval based chatbot

The architecture of the generative chatbot as Fig. 3 is designed to dynamically create new responses based on user input, taking into account the context of previous conversations. At the core of this architecture is the user input, where users submit questions, statements, or requests for information. This input can vary from simple to complex text related to specific topics, such as tourist attractions in Tegal. The Previous Input (contextual memory) component plays a crucial role in storing and managing the history of previous interactions. This component includes data from one or more prior exchanges, which is essential for ensuring that the chatbot can provide relevant and coherent responses. This system enables the model to retain the context of the ongoing conversation, so that the responses generated are not only based on the current input but also consider the context from previous interactions.

The generative model (LLaMA 2 / GEMMA) is the core of this architecture and is responsible for generating coherent and informative text. In this model, user input and previous input (contextual memory) are combined to create a comprehensive context representation. LLaMA 2 or GEMMA then processes this representation to generate a response. Generative models like LLaMA 2 utilize Transformer architectures trained to predict sequences of words in text, allowing the model to produce responses that are rich in detail and contextually relevant. After the generative model processed the input and contextual memory, the generated text response is delivered back to the user. This response is typically more contextual, reflecting the ongoing conversation and providing relevant and precise information. The model can generate responses that not only address specific questions but also offer broader elaboration in accordance with the accumulated context throughout the interaction [25].

D. Testing

The testing phase is a critical step in ensuring that the developed model functions as expected and can be applied in real-world scenarios. The testing for the retrieval-based MLP chatbot model was conducted to evaluate how effectively the model can recognize patterns and provide accurate responses. The model was tested for accuracy by measuring its ability to match input patterns with the appropriate tags and deliver correct responses. This was done using a test set that was not utilized during training.

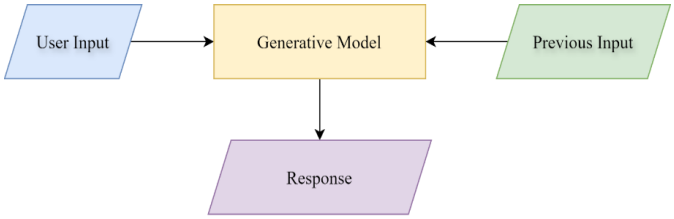

Fig. 3 Architecture of generative based chatbot

For the generative chatbot models, including LLaMA 2 and GEMMA, the testing process was more complex as it involved assessing the quality of the generated text. Beside BLUE (Best Linear Unbiased Estimation), this evaluation also was conducted by human evaluators to provide a more subjective and in-depth assessment. Human-in-the-loop testing was also conducted, involving direct user participation to evaluate how well the model could provide responses that meet user needs and expectations. The generative models' responses were tested in realistic conversational scenarios to observe how they handle various requests and contexts. Finally, the performance of the two approaches retrieval-based and generative chatbots was compared to assess the strengths and weaknesses of each.

III. RESULT AND DISCUSSION

A. Dataset Result

The initial stage of this research involved creating a dataset. For the retrieval-based chatbot, the dataset was organized in JSON format, consisting of tags, patterns, and responses, where each tourist attraction was represented by categories such as descriptions, facilities, and others. The result shows that there are 71 tags as labels or categories, and 268 patterns as sample questions. Meanwhile, the dataset for the generative-based chatbot approach was structured as long, informative narrative texts and stored in PDF format containing 11.621 words. The sample of the dataset can be seen in Table I.

The process of building a dataset for generative chatbots is less complicated because it mostly entails gathering narrative text without requiring extensive classification. Generative models produce replies in a flexible manner by learning from enormous text datasets. Retrieval-based chatbots, on the other hand, need structured databases with tags that classify data, like "facilities" or "ticket prices." Furthermore, prepared responses are made to match the patterns that the chatbot uses to identify particular user inquiries. These patterns must be constructed. Building datasets for retrieval chatbots is more complicated than for generative models because of this organized approach.

TABLE I
DATASET SAMPLE

Dataset of MLP Single QA		
Tag	**Pattern**	**Responses**
"Guci_ activities"	"*Aktivitas apa saja yang bisa dilakukan di Guci*?" ("What activities can you do at Guci?")	"*Aktivitas di Guci meliputi mengunjungi air terjun yang menawan, berendam di pemandian air panas, …….*" ("Activities in Guci include visiting charming waterfalls, soaking in hot springs, …….")
Dataset of MLP Multi QA		
Tag	**Pattern**	**Responses**
"Guci_ activities"	*"Aktivitas apa saja yang bisa dilakukan di Guci?", "Apa saja kegiatan di Guci?", "Apa yang bisa dilakukan di Guci?"* ("What activities can be done in Guci?", "What are the activities in Guci?", "What is there to do in Guci?")	*"Di Guci, Anda dapat menikmati berbagai aktivitas alam yang menarik seperti mengunjungi air terjun, berendam di pemandian air panas, ...", "Menjelajahi keindahan air terjun, merasakan hangatnya pemandian air panas, menikmati suasana hutan pinus, dan memetik stroberi segar, ...", "Aktivitas di Guci meliputi mengunjungi air terjun yang menawan, ..."* ("In Guci, you can enjoy various interesting natural activities such as visiting waterfalls, soaking in hot springs, ...", "Exploring the beauty of waterfalls, feeling the warmth of hot springs, enjoying the atmosphere of a pine forest, and picking fresh strawberries, ...", "Activities in Guci include visiting charming waterfalls, ...")
Dataset of Generative Based Chatbot (LLaMA 2, GEMMA)		
Daya tarik utama Guci adalah pemandian air panas alami dan pemandangan alam pegunungan yang sejuk. Anda bisa menikmati pemandian air panas alami dan pemandangan alam yang indah di Guci. Guci menawarkan pengalaman relaksasi dengan pemandian air panas alami dan pemandangan alam yang menenangkan. Di Guci terdapat fasilitas seperti kolam pemandian air panas, tempat makan, dan penginapan. (The main attraction of Guci is the natural hot springs and cool mountain views. You can enjoy natural hot springs and beautiful natural views in Guci. Guci offers a relaxing experience with natural hot springs and calming natural views. In Guci there are facilities such as hot springs, places to eat and accommodation.)		

B. Model Deployment Result

During the model deployment phase, we constructed two types of chatbots: an MLP-based retrieval chatbot and generative chatbots built with LLaMA 2and GEMMA. These models were deployed and incorporated into a user-friendly interface via Streamlit. The models were initially rigorously tested to confirm their performance and applicability for real-world applications. After passing these tests, the models were deployed to Streamlit for real-time interaction, with human evaluations obtained to judge the accuracy and naturalness of their responses. The deployed chatbots were subsequently subjected to human rating trials. This configuration enabled a thorough examination of both retrieval and generative approaches in real time via direct user feedback on the platform.

C. Testing Result

To assess the performance of the chatbots, we used confusion matrix analysis for the retrieval-based chatbot and BLEU score for the generative chatbot. In addition, both models were evaluated using human ratings to determine overall user happiness and the quality of responses delivered by each chatbot. Before the testing process on the model for a retrieval-based chatbot, a data split process is carried out first. This split data produces 80% training data and 20% testing data. Table II present the confusion matrix results for the MLP-based retrieval chatbot, both in single-question-answer (QA) and multi-QA setups. These visualizations provide information about the chatbot's performance.

TABLE II
CONFUSION MATIX OF MLP-BASED RETRIEVAL

MLP-based retrieval	Accuracy	Precision	Recall	F1-Score
Single QA	1.00	1.00	1.00	1.00
Multi QA	0.85	0.73	0.75	0.73

The results in the Table II indicate that the MLP-based retrieval model's single QA configuration is most likely overfitting. The perfect results across all metrics are accuracy of 1.00, precision of 1.00, recall 1.00, and f1-score of 1.00 which is indicate that the model works flawlessly on the test data, which is usually a sign that it has learnt too many specific features from the training data. This suggests that the model may have memorized the training instances rather than generalizing patterns that can be applied to new, untested data. As a result, the perfect results in single QA raise worries about overfitting, which occurs when a model is well-tuned for training data nevertheless struggles to perform effectively on fresh data. In contrast, the multi-QA configuration produces lower, more realistic values 0.85 as accuracy, 0.73 as precision, 0.75 as recall, 0.73 as f1-ccore, indicating higher generalization.

Overfitting arises when a model becomes excessively tailored to the patterns and features present in the training data, resulting in exceptional performance on the training set while performing poorly on unseen data. Single QA datasets, which typically exhibit straightforward structures such as a single question-answer pair per topic, are particularly susceptible to overfitting. This susceptibility is amplified in Multilayer Perceptron (MLP) models due to their capacity to learn complex mappings. Additionally, the MLP model might possess an excessive number of parameters relative to the dataset's size or complexity, further increasing its likelihood of memorizing the training data rather than capturing generalizable patterns.

Table III shows the results of the BLEU score evaluation for the generative chatbots GEMMA and LLaMA 2. These scores provide a quantifiable measure of the quality of each model's responses, judging their capacity to generate text that closely resembles human-like interaction within the context of Tegal tourism.

TABLE III
BLEU SCORE OF GENERATIVE BASED CHATBOT

Generative-Based Chatbot	BLEU Score
LLaMA 2	**0.33**
GEMMA	0.26

The Table III displays the BLEU scores for the generative chatbots LLaMA 2 and GEMMA. The BLEU score is a metric utilized to assess the quality of machine-generated text by measuring its similarity to predefined reference responses, with a higher score suggesting a greater similarity to human responses. In this situation, LLaMA 2 has a BLEU score of 0.33, while GEMMA has 0.26. These findings indicate that LLaMA 2 provided greater precision and fluent responses than GEMMA, as evidenced by the higher BLEU score. Both models, however, have space for development, as BLEU scores approaching 1.00 suggest near-perfect alignment with human-like language in terms of accuracy and fluency.

Following the confusion matrix and BLEU score evaluations, we conducted a human rating assessment using Streamlit's online chatbot interface. The human rating was carried out with 36 participants who submitted the same set of questions to all four chatbot models. This testing focused on two main aspects: response accuracy (rating the correctness of the chatbot's answers to user queries) and naturalness/fluency (grading how human-like and fluid the chatbot's responses were throughout interaction). The results of this evaluation offer valuable insights into the practical performance of the system in real-world scenarios and user experience of both the MLP-based retrieval chatbot and the generative chatbots, GEMMA and LLaMA2. Table IV contains a full summary of the human rating results for all four chatbot models.

TABLE IV
RESULT OF CORRECTNESS AND NATURALNESS

Methods	Correctness (Response Accuracy)	Naturalness	Overall Rating
MLP-Based Retrieval (Single QA)	0.78	0.69	0.74
MLP-Based Retrieval (Multi QA)	0.83	0.64	0.74
LLaMA 2 (Generative)	**0.89**	**0.75**	**0.82**
GEMMA (Generative)	0.56	0.33	0.44

LLaMA 2 has the highest response accuracy of 0.89 and naturalness of 0.75, resulting in a strong overall rating of 0.82. This demonstrates that LLaMA 2 excelled in both providing accurate responses and producing natural, human-like text. GEMMA scored the lowest across all criteria, with a response accuracy of 0.56 and a naturalness of 0.33, resulting in an unsatisfactory overall rating of 0.44. This implies that GEMMA struggled to offer both correct and fluid responses, making it less effective overall.

GEMMA, as a generative model, performs inferior to retrieval models. GEMMA might struggle to comprehend the context of questions or might not have adequate training on various data, resulting in less accurate and contextually relevant responses. Generative models, especially those with little training data, frequently produce responses that are less aligned with user expectations, resulting in lower accuracy. Furthermore, GEMMA's responses might be hampered since the model is unable to properly capture human-like conversational patterns, resulting in more rigid or unnatural language. The results indicate that while LLaMA 2 is the best performing model in terms of both accuracy and fluency as well as a higher BLEU score.

In support of these findings, a study [27] highlights that LLaMA 2 outperforms other models, particularly in code optimization tasks. This research underscores LLaMA 2's superior ability to deliver precise and optimized responses, which is consistent with the high accuracy observed in our evaluation. The enhanced performance of LLaMA 2, as noted that work, further validates its effectiveness and reliability in complex tasks, reinforcing the conclusion that LLaMA 2 is the most capable model among those tested in this study.

The research [28] suggests that LLaMA 2 is considered the optimal choice for NLP tasks. This is because the model offers an excellent balance between computational efficiency and high performance. Additionally, LLaMA 2 can be fine-tuned using readily accessible GPUs, making it a more practical solution for various NLP applications.

IV. CONCLUSION

In the final analysis, the ease of creating datasets for generative models like LLaMA 2, which primarily relies on narrative text data, further strengthens its advantage over retrieval-based models. The comparison research reveals that LLaMA 2, as a generative-based chatbot, surpasses the other models in terms of response accuracy is 0.89 and naturalness is 0.75, earning the highest overall rating of 0.82, along with a BLEU score of 0.33. The MLP-based retrieval models, both single QA and multi QA, show great accuracy but fall short of naturalness, illustrating the trade-off between precision and fluency in retrieval-based systems. Meanwhile, GEMMA falls short in terms of accuracy and naturalness, indicating limitations in its ability to perceive context and deliver human-like responses. These findings show that, while generative models such as LLaMA 2 provide higher overall conversational quality, models such as GEMMA require additional enhancements to achieve comparable performance levels. In the context of Tegal tourism information service, installing a generative model such as LLaMA 2 could improve user experience by giving more accurate and fluent responses, hence boosting the quality of tourist information service.

ACKNOWLEDGEMENT

This research was completely supported by the funding provided by the Ministry of Research, Technology, and Higher Education of the Republic of Indonesia under contract number 103/SPK/D.D4/PPK.01.APTV/III/2024.

REFERENCES

[1] T. Pencarelli, “The digital revolution in the travel and tourism industry,” *Inf. Technol. Tour.*, vol. 22, no. 3, pp. 455–476, 2020, doi: 10.1007/s40558-019-00160-3.

[2] C. C. Lin, W. Y. Liu, and Y. W. Lu, “Three-dimensional internet-of-things deployment with optimal management service benefits for smart tourism services in forest recreation parks,” *IEEE Access*, vol. 7, pp. 182366–182380, 2019, doi: 10.1109/ACCESS.2019.2960212.

[3] J. R. Chang, M. Y. Chen, L. S. Chen, and S. C. Tseng, “Why Customers Don’t Revisit in Tourism and Hospitality Industry?,” *IEEE Access*, vol. 7, pp. 146588–146606, 2019, doi: 10.1109/ACCESS.2019.2946168.

[4] D. I. Af’Idah, Dairoh, and S. F. Handayani, “Virtual Tour Application as A Tourism Information Media Using Multimedia Development Life Cycle,” in *2024 1st International Conference on Robotics, Engineering, Science, and Technology (RESTCON)*, 2024, pp. 207–213. doi: 10.1109/RESTCON60981.2024.10463581.

[5] E. Adamopoulou and L. Moussiades, “Chatbots: History, technology, and applications,” *Mach. Learn. with Appl.*, vol. 2, no. July, p. 100006, 2020, doi: 10.1016/j.mlwa.2020.100006.

[6] R. Dsouza, S. Sahu, R. Patil, and D. R. Kalbande, “Chat with Bots Intelligently: A Critical Review Analysis,” *2019 6th IEEE Int. Conf. Adv. Comput. Commun. Control. ICAC3 2019*, 2019, doi:

10.1109/ICAC347590.2019.9036844.

[7] D. R. Darmawan and R. Arifudin, "Enhancing Durrotalk Chatbot Accuracy Utilizing a Hybrid Model Based on Recurrent Neural Network (RNN) Algorithm and Decision Tree," *JUITA J. Inform.*, vol. 12, no. 1, p. 81, 2024, doi: 10.30595/juita.v12i1.20868.

[8] L. Benaddi, C. Ouaddi, A. Jakimi, and B. Ouchao, "Towards A Software Factory for Developing the Chatbots in Smart Tourism Mobile Applications," *Procedia Comput. Sci.*, vol. 231, no. 2023, pp. 275–280, 2024, doi: 10.1016/j.procs.2023.12.203.

[9] O. M. Alyasiri, K. Selvaraj, H. A. Younis, T. Mueen Sahib, M. F. Almasoodi, and I. M. Hayder, "A Survey on the Potential of Artificial Intelligence Tools in Tourism Information Services Keywords Artificial Intelligence ChatGPT Hospitality and Tourism Industry Tourism Education and Research Tourism and Management," *Babylonian J. Artif. Intell.*, vol. 2024, pp. 1–8, 2024.

[10] P. J. Antony, R. Kannan, and A. Professor, "Revolutionizing the Tourism Industry through Artificial Intelligence: A Comprehensive Review of AI Integration, Impact on Customer Experience, Operational Efficiency, and Future Trends," *www.chandigarhphilosophers.com International Journal for Multidimensional Research Perspectives (IJMRP*, vol. ISSN, no. 1. pp. 2584–2613, 2024.

[11] D. I. Af'idah, P. D. Anggraeni, M. Rizki, A. B. Setiawan, and S. F. Handayani, "Aspect-Based Sentiment Analysis for Indonesian Tourist Attraction Reviews Using Bidirectional Long Short-Term Memory," *JUITA J. Inform.*, vol. 11, no. 1, p. 27, 2023, doi: 10.30595/juita.v11i1.15341.

[12] H. Akkineni, P. V. S. Lakshmi, and L. Sarada, "Design and Development of Retrieval-Based Chatbot Using Sentence Similarity BT - IoT and Analytics for Sensor Networks," P. Nayak, S. Pal, and S.-L. Peng, Eds., Singapore: Springer Singapore, 2022, pp. 477–487.

[13] B. El Bakkouri, S. Raki, and T. Belgnaoui, "The Role of Chatbots in Enhancing Customer Experience: Literature Review," *Procedia Comput. Sci.*, vol. 203, pp. 432–437, 2022, doi: 10.1016/j.procs.2022.07.057.

[14] T. B. Brown *et al.*, "Language models are few-shot learners," *Adv. Neural Inf. Process. Syst.*, vol. 2020-Decem, 2020.

[15] N. Mathur, T. Baldwin, and T. Cohn, "Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 4984–4997, 2020, doi: 10.18653/v1/2020.acl-main.448.

[16] M. K. Ogirala, R. Tallapaneni, S. M. Chalamcharla, and A. Chinta, "A Medical Diagnosis and Treatment Recommendation Chatbot using MLP," in *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 2023, pp. 495–500. doi: 10.1109/ICAAIC56838.2023.10141211.

[17] S. Pandey and S. Sharma, "A comparative study of retrieval-based and generative-based chatbots using Deep Learning and Machine Learning," *Healthc. Anal.*, vol. 3, no. May, p. 100198, 2023, doi: 10.1016/j.health.2023.100198.

[18] G. Sperlí, "A cultural heritage framework using a Deep Learning based Chatbot for supporting tourist journey," *Expert Syst. Appl.*, vol. 183, no. May, p. 115277, 2021, doi: 10.1016/j.eswa.2021.115277.

[19] H. Touvron *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models," 2023.

[20] "GPT-4 Technical Report," vol. 4, pp. 1–100, 2023.

[21] Gemma Team *et al.*, "Gemma: Open Models Based on Gemini Research and Technology," 2024.

[22] K. Adnan and R. Akbar, "An analytical study of information extraction from unstructured and multidimensional big data," *J. Big Data*, vol. 6, Oct. 2019, doi: 10.1186/s40537-019-0254-8.

[23] D. Cabezas, R. Fonseca-Delgado, I. Reyes-Chacón, P. Vizcaino-Imacaña, and M. Morocho-Cayamcela, "Integrating a LLaMa-based Chatbot with Augmented Retrieval Generation as a Complementary Educational Tool for High School and College Students," no. July, pp. 395–402, 2024, doi: 10.5220/0012763000003753.

[24] C. Wu, W. Lin, X. Zhang, Y. Zhang, W. Xie, and Y. Wang, "PMC-LLaMA: toward building open-source language models for medicine," *J. Am. Med. Informatics Assoc.*, 2024, doi: 10.1093/jamia/ocae045.

[25] H. Yin, A. Aryani, and N. Nambiar, "Evaluating the Performance of Large Language Models for SDG Mapping (Technical Report)," pp. 1–6, 2024.

[26] M. Hinck, M. L. Olson, D. Cobbley, S.-Y. Tseng, and V. Lal, "LLaVA-Gemma: Accelerating Multimodal Foundation Models with a Compact Language Model," 2024.

[27] O. Ridge and O. Ridge, "Comparing Llama-2 and GPT-3 LLMs for HPC kernels generation," pp. 1–13.

[28] A. Kumar, "Towards Optimal NLP Solutions : Analyzing GPT and LLaMA-2 Models Across Model Scale , Dataset Size , and Task Diversity," vol. 14, no. 3, pp. 14219–14224, 2024.