Journal of Engineering, Technology & Applied Science
p-ISSN: 2721-7949, e-ISSN: 2721-8090

Original Research Paper

Comparative Analysis of Chatbot Development Methods on Flexibility and Control

Achmad Yani[1*], Andi Almeira Zocha Ismail[2], Andi Regina Acacia Ismail[2], Pratiwi Hendro Wahyudiono[3]

[1] *Department of Shipbuilding Engineering, Faculty of Engineering, Universitas Hasanuddin. Makassar, Indonesia.*
[2] *Program of Robotics Engineering, Department of Electrical Engineering, Politeknik Negeri Batam (Polibatam). Batam, Indonesia.*
[3] *Department of Industrial Engineering, Faculty of Engineering, Universitas Andalas. Padang, Indonesia.*

Article History
Received:
17.07.2025

Revised:
31.07.2025

Accepted:
16.08.2025

***Corresponding Author:**
Achmad Yani
Email:
achmad.yani@gmail.com

Abstract: Chatbots are dialogue systems driven by Natural Language Processing (NLP) and Artificial Intelligence (AI), extensively utilized in areas like customer support, education, and healthcare. Nonetheless, the variety in approaches to chatbot development, ranging from rule-based systems to generative AI, creates difficulties in harmonizing design decisions with user requirements and technical limitations. This research seeks to examine and contrast the primary approaches employed in chatbot creation: rule-based, retrieval-based, and generative-based systems. Employing a descriptive-qualitative methodology, the study is carried out in the first quarter of 2025 and utilizes scholarly literature, technical documents, and case studies of Mitsuku, Google Assistant, and ChatGPT, concentrating on applications in Indonesia and Malaysia. A comparative analysis assesses every method based on development complexity, accuracy, flexibility, user experience, interpretability, cost, and ethical risks. The results indicate that rule-based systems provide low expenses and significant transparency, but they fall short in scalability and flexibility. Retrieval-based systems excel in accuracy for domain-specific queries but struggle with new interactions. Chatbots based on generative models provide the most natural and contextually aware interactions, but they require significant resources and present issues related to interpretability and ethics. The research suggests that hybrid models integrating control and adaptability could be the most efficient approach. Further studies are required to improve transparency in generative systems, reduce bias, and create adaptive hybrid architectures appropriate for Southeast Asian use.

Keywords: Chatbot Architecture, Conversational AI, Hybrid Chatbot, NLP Applications, Rule-Based vs. Generative.

Journal of Engineering, Technology & Applied Science, vol. 7, no. 2, pp. 80-92, August 2025. **DOI: 10.36079/lamintang.jetas-0702.881**

1. Introduction

A chatbot is a software system designed to simulate human conversation through a text or voice interface, thus enabling seamless interaction between user and machine [1]. Chatbots function as conversational agents powered by natural language processing (NLP) and artificial intelligence (AI) to interpret user input and provide coherent responses [1] - [4]. These systems vary from simple rule-based designs to generative models capable of generating dialogs resembling human conversations [2] [5] [6].

In today's digital ecosystem, chatbots play an important role in various sectors, customer service, education, healthcare, e-business, and others [7] - [9]. In customer service, chatbots help lower costs and provide 24-hour support by handling repetitive queries efficiently [6]. In education and healthcare, chatbots support personalized learning and patient interaction through guided dialogue as well as information gathering [2] [9]. The growing reliance on these systems demonstrates their benefits in improving accessibility, efficiency, and continuity of services [6] [9].

Despite the obvious benefits, designing a reliable chatbot remains a challenge. Rule-based approaches with artificial if-then logic provide high interpretability yet are inflexible to input variations [6] [10] [11]. In contrast, AI-based methods such as supervised learning, deep neural networks (e.g., LSTM, transformers), and reinforcement learning, produce richer and contextualized responses, but require large data and computational resources [1] [2] [4]. Hybrid systems that combine rule logic and AI modules are beginning to emerge as a promising solution [11].

This paper aims to review the main methods used in chatbot development, ranging from rule-based, retrieval-based, to generative AI and hybrid approaches. We review how each approach works, its technical advantages, as well as its limitations. In particular, we highlight contemporary trends such as the use of transformers (e.g., BERT, GPT-4), reinforcement learning strategies, and semantic search models [1] [2] [7]. By synthesizing the latest developments (2021-2025), this review maps how various methodologies shape modern chatbots.

The significance of this study lies in its thorough review that is relevant to both researchers and practitioners. For developers, understanding the strengths and weaknesses of the methods is critical to aligning the chatbot design with project goals such as domain specifications, resource limitations, and user experience. For researchers, the review identified gaps, such as the need to balance interpretability with dialog fluency and integrate conversation management with ethical considerations. It also highlights the importance of hybrid approaches that combine rule precision and AI adaptability.

This study aims to establish a clear taxonomy of chatbot development methods, explaining how rule-based, retrieval-based, generative and hybrid systems are interrelated and complementary. In doing so, this paper provides insights that can be applied to future designs, encouraging transparent and ethical AI development, and increasing the effectiveness of chatbots in a variety of real-world applications.

2. Literature Review

2.1. Popular Chatbot Development

Popular chatbot development methods currently used include:

1) Rule-Based Chatbot

Chatbots that follow rules depend on predetermined IF–THEN logic to analyze user input and generate responses [8] [12]. They establish set patterns; for instance, when a user inputs “opening hours,” the bot replies with a predefined message like “Our hours are 09:00 to 17:00.” This consistent behavior guarantees reliable interactions.

These chatbots are ideal for straightforward FAQ and transactional scenarios such as scheduling appointments or accessing static information [9] [11]. Their design is simple, needing neither extensive datasets nor complicated training. Developers create decision trees that correlate user input with responses, making deployment easier.

The systems based on rules have difficulty with the variability of natural language. They need precise keyword or phrase matches and do not work with paraphrases or typing errors [9] [11]. This constraint results in regular breakdowns when faced with unforeseen input, diminishing user contentment.

As applications expand, upholding rule sets grows more challenging [10] [11]. Every new query type requires manual modifications to the logic. This scaling problem limits their use in changing

environments with shifting user behavior.

Rule-based chatbots, despite their constraints, continue to be an economical option for small enterprises or specific processes [9] [12]. They offer continuous support with limited resources and can work with current systems. This renders them useful in situations where the field is limited and foreseeable.

In the healthcare sector, rule-based bots are effectively utilized for initial screening and triage. An implementation for youth mental health in under-resourced areas, attaining over 88% accuracy in 60 test cases [12]. This shows that thoughtful rule construction enables these bots to assist with initial diagnostics.

In general, rule-based chatbots provide advantages including dependability, transparency, and straightforward implementation; however, they are inflexible and necessitate ongoing upkeep [9] [10] [12]. They stay applicable in areas with narrow focus and high predictability.

2) Retrieval-Based Chatbot

Retrieval-based chatbots choose replies from a predetermined database by comparing user input to existing responses [13] - [17]. They may employ basic keyword matching or more sophisticated vector similarity methods like TF-IDF and cosine similarity.

TF-IDF along with cosine similarity continues to be a popular method for ranking potential responses [13] [14]. Setiawan and Adnyana express a 75% accuracy in recognition for helpdesk chatbots by employing TF-IDF and cosine similarity within an Indonesian setting [13]. Their research shows the effectiveness of retrieval methods when resources are scarce.

Current studies investigate hybrid retrieval techniques that combine lexical retrieval (e.g., BM25) with semantic retrieval (e.g., FAISS embeddings) [14] [17]. Juvekar and Purwar present COS-Mix, integrating cosine similarity with distance metrics to enhance retrieval effectiveness, especially in sparse data contexts [14].

Reddit users also talk about enhancing retrieval-based bots with RAG (Retrieval-Augmented Generation) systems, where retrieval obtains clear contexts and generative models articulate or expand on answers [17] [18]. These hybrid systems combine the dependability of retrieval with the adaptability of generation.

While retrieval-based bots lack the ability to generate original answers, they achieve high accuracy and extensive coverage when the databases are thorough [13] [15]. They are considered more secure in areas like customer support, where precision and adherence to regulations are vital.

They experience difficulties when user queries exceed the database limits or when there are variations in paraphrasing. Coverage gaps result in unsuccessful matches or revert to default error messages [15]. Continuous upkeep of response databases is crucial.

Current trends view retrieval-based systems as supplementary components within hybrid architectures [17] [18]. By merging retrieval with generative components, developers obtain responses that are more contextually relevant and secure for intricate tasks.

3) Generative Chatbot

Generative chatbots utilize Large Language Models (LLMs) to create original replies based on user input and context. Initial architectures utilize sequence-to-sequence (Seq2Seq) models alongside RNNs or LSTM encoder-decoder setups [19] [20] [21].

Google's Meena (2020) and Amazon's AlexaTM 20B (2022) serve as prime examples of Seq2Seq models that produce conversational text with high coherence [19] [20]. These models exceed conventional retrieval bots in fluidity and flexibility.

The emergence of transformer models such as GPT and BERT architectures signifies a fundamental change towards generative agents that are instruction-tuned and trained with RLHF [20] [21]. OpenAI’s GPT-3.5 and GPT-4, Microsoft’s Bing Chat, and Google’s Gemini all employ reinforcement learning from human feedback (RLHF) to improve dialogue quality [20] [21].

RLHF aligns the outputs of models with human preferences, resulting in more appropriate and secure responses [21]. Wikipedia mentions that RLHF integrates supervised finetuning with reinforcement learning driven by models of human preferences [21]. Nonetheless, RLHF demands intricate feedback gathering and is resource-intensive.

Generative bots provide adaptability and awareness of context, yet they can create hallucinations and potentially perpetuate biases [20] [21]. The TIME interview featuring ChatGPT emphasizes that

although these models seem articulate, they do not possess genuine comprehension and can produce deceptive information [20].

Even with limitations, generative chatbots are quickly being embraced in fields that need open dialogue, writing support, and customized tutoring. Their capacity to generate text that resembles human writing provides a unique benefit in engaging users.

Current studies concentrate on enhancing performance with smaller fine-tuned models, improved alignment through RLHF, and hybrid systems that manage generation by employing retrieval or rule constraints to reduce incorrect outputs.

2.2. Technologies and Supporting Algorithms

The creation of smart chatbots depends not just on conversational frameworks but also on various sophisticated technologies and algorithms that facilitate natural and efficient communication. These technologies provide the groundwork that allows chatbots to comprehend user input, deduce intent, and produce appropriate replies. Natural Language Processing (NLP), Machine Learning (ML), Deep Learning (DL), and Reinforcement Learning (RL) are essential elements that drive contemporary chatbot systems.

Techniques in Natural Language Processing, including tokenization, lemmatization, and named entity recognition, enable chatbots to understand and derive structured meaning from unstructured text. Machine learning, particularly supervised and unsupervised techniques, is employed for activities such as intent classification and topic clustering, enhancing the precision and relevance of chatbot replies. Additionally, deep learning frameworks such as LSTM, GRU, and Transformer models provide advanced contextual comprehension and facilitate the creation of smooth, human-like responses.

Reinforcement Learning, especially when used with human feedback (RLHF), is progressively utilized to refine chatbot responses according to user engagement, improving their capacity to adjust to actual conversations. Collectively, these technologies collaborate to create chatbots that are responsive, context-sensitive, and adaptable across numerous applications.

1) Natural Language Processing

Tokenization is fundamentally connected to natural language processing (NLP), serving as the initial step that allows machines to comprehensively grasp human language. In NLP, unprocessed text data needs to be converted into a format that can be mathematically handled by a computer. The process of tokenization divides text into smaller elements known as tokens, usually consisting of words, brief phrases, or characters, to produce a more structured format that can be effectively analyzed by a natural language processing system [24].

Tokenization and lemmatization are essential preprocessing steps that help standardize user input into a uniform format. This procedure normalizes terminology, minimizes sparsity, and enhances the precision of NLP models [25]. NLP systems can be more effective in recognizing patterns, comprehending context, and producing more accurate responses or classifications by simplifying various word forms into one fundamental form (lemmatization). Tokenization allows NLP to recognize word boundaries, which is especially crucial in languages that do not have clear spacing, like Japanese or Thai.

Moreover, methods like Named Entity Recognition (NER) have become essential in numerous chatbot systems and other NLP applications. NER allows chatbots to identify key entities in text, like names of individuals, places, and dates, which are essential for grasping user intent in task-focused conversations [24] [26]. For instance, when someone requests, "Reserve a hotel in Batam for July 15th," the system can identify the entities “Batam” as the place and “July 15th” as the date, which is vital for carrying out the request or replying correctly

Recent studies have shown that the use of transformer architectures has significantly improved NLP performance, particularly in tasks such as NER and syntactic understanding for resource-constrained languages. A modified transformer encoder with relative positional embeddings outperformed a baseline model in an Arabic NER task, with a significant increase in F1 score [25]. This research suggests that more sensitive processing of the positional structure of tokens within a sentence can improve the model's ability to understand context.

The combining pretrained embeddings with syntactic parsing can improve coherence in responses to user queries [26]. This proves that strengthening linguistic structure through syntax and semantics, starting from the tokenization stage, can enrich the model's understanding of complex text. Thus, it

can be concluded that tokenization is not only the initial step, but also a foundation that influences the entire NLP pipeline,from preprocessing to achieving intelligent and contextual responses in natural language-based systems.

2) Machine Learning

Machine learning plays a central role in modern natural language understanding systems, particularly in intent classification for chatbots and virtual assistants. Among the most widely adopted approaches is supervised learning, which relies on labeled datasets to train models to recognize user intents. Recent advancements in transformer-based architectures, such as BERT and RoBERTa, have significantly enhanced performance in this area. These models are fine-tuned with task-specific data to learn context-aware representations, enabling them to classify user queries with high precision and adaptability across a wide range of language variations [29] [30].

Despite the high accuracy achieved through supervised learning, one of its major challenges is the requirement for large amounts of annotated data. Obtaining such data is often resource-intensive and time-consuming. To address this issue, a label-efficient method has been introduced that combines MixUp augmentation with consistency training. MixUp involves generating synthetic training examples by interpolating input samples and their labels, while consistency training encourages the model to produce stable predictions even when minor noise is introduced into the inputs. This combination has proven effective in enhancing intent classification accuracy while reducing the need for extensive labeled datasets [29].

These label-efficient strategies have demonstrated particular value in educational chatbot systems. In one study, they were applied to chatbots used in learning environments, leading to substantial improvements in performance when compared to traditional rule-based systems. As shown in [30], transformer models trained using MixUp and consistency training achieved greater accuracy and were better able to generalize across the diverse ways students express their queries. These results underscore the practical benefits of modern machine learning techniques in intent recognition tasks.

In addition to supervised methods, unsupervised learning offers complementary tools for intent discovery, particularly during the early stages of chatbot development. A commonly used technique involves applying the K-Means algorithm to sentence embeddings, which are dense vector representations of text obtained from pretrained language models. This approach enables the automatic grouping of semantically similar user queries, facilitating the identification of latent topics and intents within large conversational datasets [29] - [31].

One of the key advantages of clustering is its ability to simplify and accelerate the annotation process. By grouping related queries into coherent clusters, annotators can label representative examples rather than annotating each query individually. This approach significantly reduces manual workload. For instance, studies in intelligent tutoring systems have shown that such clustering-based methods can lower annotation effort by approximately 40 percent [30]. This increased efficiency also translates to broader intent coverage and better dataset quality for training downstream classifiers.

The integration of supervised and unsupervised machine learning methods results in a more effective, scalable, and resource-efficient pipeline for intent classification. Unsupervised clustering enables the rapid exploration of potential intent categories from raw data, while supervised fine-tuning ensures high-performance classification based on the labeled subset. Together, these methods support the development of intelligent conversational agents that are accurate, flexible, and better aligned with user expectations.

3) Deep Learning

Deep learning techniques have become foundational in developing intelligent conversational systems. Traditional deep sequence models such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) have long demonstrated effectiveness in modeling sequential dependencies within dialogue systems. These architectures are particularly well-suited for slot-filling tasks, intent detection, and handling multi-turn conversations due to their ability to retain contextual information over time [27] [28]. Their application in real-world scenarios continues to yield strong results, particularly in controlled or domain-specific environments.

Several studies have highlighted the ongoing relevance of LSTM-based models. Benaddi et al. propose a sequence-to-sequence (Seq2Seq) framework utilizing LSTM layers coupled with attention mechanisms, which successfully generates coherent multi-turn dialogues for domain-specific chatbot applications [27]. This framework allows the model to selectively focus on relevant input tokens

during generation, improving response accuracy and relevance. Similarly, Alaa Shaker et al. implement a hybrid LSTM-GRU architecture for Named Entity Recognition (NER) in Arabic, a morphologically complex language. Their model achieves approximately 80 percent accuracy in entity recognition, showcasing the effectiveness of recurrent networks in handling languages with rich inflectional morphology [28].

Despite the continued utility of LSTM and GRU architectures, transformer-based models have largely taken precedence in state-of-the-art conversational AI. Transformers offer parallel processing capabilities and superior performance on long-range dependencies, making them highly effective for various dialogue-related tasks. Recent trends favor hybrid models that integrate the strengths of different transformer components. For example, Jia, Liang, and Liang survey models that combine BERT-based encoders with GPT-style decoders. These hybrid systems achieve impressive results across a variety of tasks, including intent classification, slot filling, and natural language generation [27].

Such transformer-based ensembles not only improve performance metrics but also enhance system scalability and adaptability. By leveraging pretrained language models, developers can fine-tune chatbot systems to perform well across multiple domains and languages with relatively minimal task-specific data. This multilingual and domain-adaptive flexibility is particularly valuable for creating globally deployable conversational agents. The growing dominance of transformer architectures signifies a paradigm shift in how deep learning is applied to natural language dialogue systems, merging language understanding and generation into a unified, end-to-end trainable framework [27].

4) Reinforcement Learning

Reinforcement Learning (RL) has become increasingly influential in the development of adaptive and context-aware dialogue systems. Unlike supervised learning, which relies on fixed labeled data, RL enables models to learn optimal dialogue strategies by interacting with their environment and receiving feedback in the form of rewards or penalties. In the context of chatbots, this means the system can continuously improve its responses based on user satisfaction, task success, or other performance indicators. This approach is particularly valuable in dynamic conversational environments where user preferences and interaction patterns evolve over time [24, [31].

One of the most promising developments in this area is Reinforcement Learning with Human Feedback (RLHF), which integrates human evaluations directly into the learning loop. Rather than depending solely on predefined reward functions, RLHF systems incorporate qualitative judgments from human users or annotators to guide policy updates. This makes the chatbot's behavior more aligned with human values and expectations. Kaufmann et al. provide a comprehensive survey of RLHF techniques, highlighting their ability to improve contextual relevance and ensure safer outputs from large language model (LLM)-based dialogue systems [31].

The benefits of RLHF are evident in systems that prioritize long-term engagement, task success, and ethical response generation. By training on human preference data, RLHF enhances the chatbot’s ability to generate coherent, helpful, and socially acceptable replies. This is particularly important in high-stakes domains such as education, healthcare, and customer service, where conversational errors can negatively impact user experience or trust. Moreover, RLHF addresses limitations of static training datasets by enabling models to adapt in real-time to complex, nuanced user behavior [31].

Despite its advantages, RLHF also poses significant challenges. One of the primary difficulties lies in designing accurate and robust reward models that reflect desired conversational behavior. Human feedback is inherently subjective, variable, and costly to collect at scale. Additionally, aligning reward models with long-term conversation quality rather than short-term heuristics remains an ongoing research concern. Kaufmann et al. note that poor reward model design can lead to unintended behaviors, such as over-optimization of generic responses or reward hacking, where the agent exploits flaws in the reward function [31].

To address some of these challenges and improve the efficiency of reinforcement-based training, researchers have explored modular fine-tuning approaches. One such approach is adapter-based reinforcement learning, which focuses on updating only small, task-specific modules (adapters) inserted into a larger pretrained model. This method avoids the need for full model retraining, offering a lightweight and computationally efficient strategy for adapting dialogue systems in post-deployment settings [24] [32].

Adapter-based RL has shown promising results in both controlled experiments and real-world chatbot deployments. Casper et al. evaluate the effectiveness of adapter tuning in conversational

systems, demonstrating that limited parameter updates can significantly improve response quality, coherence, and user satisfaction. Importantly, this adaptation does not compromise the integrity of the base model, preserving its general language capabilities while tailoring specific conversational behaviors [32].

This modular fine-tuning strategy also supports faster iteration cycles, making it easier to refine chatbot performance based on newly collected interaction data. In practical terms, organizations can apply adapter-based RL to incrementally improve chatbot functionality without incurring the high cost of retraining large models from scratch. As reinforcement learning techniques continue to mature, particularly those incorporating human feedback and lightweight adaptation, they are likely to play a central role in building scalable, responsive, and user-aligned conversational agents [24]. [32].

3. Methodology

This research employs a descriptive-qualitative design to examine and contrast the primary methods utilized in the development of chatbots. The study takes place in the initial quarter of 2025 and seeks to deliver a comprehensive insight into how various approaches influence the effectiveness and flexibility of contemporary chatbot systems.

The main data sources consist of peer-reviewed journal articles, official technical documentation from chatbot development platforms, and case studies of popular chatbot systems. This study focuses on Mitsuku, Google Assistant, and ChatGPT, as they exemplify popular and varied chatbot applications currently utilized in Indonesia and Malaysia.

The evaluation utilizes a comparative framework to assess different chatbot development methods, including rule-based, retrieval-based, generative AI, and hybrid models. Every method is evaluated on four key factors: complexity of development, accuracy of responses, flexibility of dialogue, and overall user experience. These criteria enable a thorough assessment of the strengths and weaknesses of each method in practical applications.

Using this analytical perspective, the research uncovers patterns, compromises, and new trends that guide optimal practices in chatbot development. Emphasizing both technical and experiential aspects guarantees that the results are applicable not only for developers and researchers but also for decision-makers aiming to deploy effective conversational AI systems in Southeast Asian settings.

4. Finding and Discussion

4.1. Rule-Based Chatbot

Chatbots based on rules function using established logic via IF-THEN statements. These systems are simple and easy to set up, making them ideal for consistent, repetitive exchanges like Frequently Asked Questions (FAQs). The straightforward nature of rule-based systems allows for quick development requiring little technical knowledge. Their primary constraint is adaptability and scalability. They cannot understand differences in user input that deviate from pre-defined pathways.

In the case of Mitsuku, a rule-based chatbot with certain AI improvements, it demonstrates sufficient performance in closed-domain situations but struggles to handle ambiguous or open-ended questions. In practical use in Indonesia and Malaysia, such systems are still common in customer service portals for managing structured inquiries such as transaction status or operating hours.

Rule-based chatbots struggle to manage the different ways a user can convey the same intention. This can be clarified by the idea that rule-based chatbots function on rigid IF-THEN principles. This indicates that the chatbot responds accurately only when the user's input precisely aligns with a specified pattern. If the user phrases a question differently, makes a spelling error, or employs synonyms, the chatbot might not respond correctly.

Example:
If the system is programmed to respond to:
User: "*What are your business hours*?"
Then it will correctly reply: "*We are open from 8 AM to 5 PM*."

But if the user says:
User: "*When do you open*?", Or: "*What time does your office start*?"
The chatbot may not recognize the intent unless specific rules are created for these variants.

In other words, rule-based chatbots lack the flexibility needed to manage natural and unpredictable human language, making them suitable only for static scenarios like FAQs.

Journal of Engineering, Technology & Applied Science, vol. 7, no. 2, pp. 80-92, August 2025. **DOI: 10.36079/lamintang.jetas-0702.881**

4.2. Retrieval-Based Chatbot

Retrieval-based chatbots choose the best response from a fixed collection by comparing user inputs through keyword matching or semantic embedding methods. These systems offer superior context awareness relative to rule-based models, particularly when improved with neural embeddings like Sentence-BERT or Universal Sentence Encoder.

For instance, Google Assistant utilizes retrieval-oriented features for organized activities like verifying calendar appointments or responding to factual questions. These chatbots show great precision for specialized tasks, especially in customer service and helpdesk settings. Nevertheless, because answers are chosen rather than produced, the model cannot formulate original responses beyond its collection of replies.

Retrieval-based chatbots are highly accurate since they choose from prewritten replies, but they do not have the capacity to create unique responses.

1) High accuracy
   Retrieval-based systems are highly accurate since they provide answers by choosing the most pertinent response from a defined collection. If the user input corresponds to an existing entry in the response bank (either via keyword or semantic similarity), the chatbot delivers an accurate and contextually suitable response.
2) Zero novelty
   As these systems merely fetch responses rather than create them, they are incapable of generating new, original, or imaginative outputs. If a user poses a question that doesn't closely match any saved input, the system will either malfunction or provide an irrelevant response.

Example:
If the user asks:
"*How do I reset my password*?"

And the chatbot has this in its response bank:
"*To reset your password, go to the Settings page and click 'Forgot Password*."

This yields high accuracy because the response directly matches a known query.

But if the user asks:
"*Can you explain the difference between password reset and account recovery*?"

Since there is likely no predefined response exactly matching this more complex question, the chatbot:

- Cannot combine or synthesize information
- Cannot generate new phrasing
- May return an irrelevant or generic answer

Therefore, the retrieval-based chatbots are suitable for structured, fact-based, or domain-specific applications, like those found in customer service or helpdesk tools (e.g., Google Assistant). However, they are not suitable for open-ended conversations or educational tutoring, where novel and adaptive responses are essential.

4.3. Generative-Based Chatbot

Chatbots based on generative methods utilize deep learning structures like Transformers (e.g., GPT-4, BERT) to create responses one word at a time. These systems can generate coherent, contextually rich, and humanlike dialogues. ChatGPT, an illustrative case, shines in open-domain conversations, abstract reasoning, and language creation in both English and Bahasa Indonesia.

Generative chatbots necessitate significant computing power and extensive training data. Implementing solutions in practical environments such as customer service in Indonesia and Malaysia frequently requires a hybrid approach, merging generative models with rule-based systems to achieve a balance of innovation and regulation. In educational environments, generative chatbots assist with customized tutoring, tailoring replies to student comprehension levels and input.

Chatbots utilizing generative models (such as GPT) outperform rule-based and retrieval-based chatbots in generating responses that seem logically coherent and resemble natural human conversation.

1) Coherence refers to how well a response stays on topic and logically follows the user's input.
2) Naturalness refers to how human-like or fluent the chatbot's response sounds in everyday language.

Generative chatbots like ChatGPT are trained on massive amounts of real-world text and learn patterns in how people write and speak. This allows them to:

- Maintain conversation flow over multiple turns
- Understand context and nuance
- Use varied and natural sentence structures

Generative chatbots excel in the quality of conversations, particularly for intricate or emotionally nuanced exchanges. This renders them perfect for open-domain tasks, educational purposes, or counseling-type engagements.

4.4. Cross-Method Comparison

Here is a summary of how the three chatbot methods (Rule-Based, Retrieval-Based, and Generative-Based) respond to an emotional question:

User Input:
"*I'm feeling stressed about my exams*."

1) Rule-Based Chatbot Response:
"*Sorry, I do not understand your request*."

- Fast response if predefined.
- Fails to respond meaningfully to emotional or unstructured input.
- No empathy or understanding.

2) Retrieval-Based Chatbot Response:
"*Please visit our student support page for exam information*."

- Useful if linked to a relevant FAQ or resource.
- Generic and emotionally disconnected.
- Does not acknowledge user's feelings.

3) Generative-Based Chatbot Response:
"*I understand how stressful exams can be. Would you like some tips on how to manage your time and reduce anxiety*?"

- Shows empathy and emotional intelligence.
- Follows up with helpful suggestions.
- Highly adaptive to various emotional tones and phrasing.

Table 1 summarizes the findings based on case studies and technical documentation of Mitsuku, Google Assistant, and ChatGPT. The comparative analysis focuses on seven critical dimensions: flexibility, accuracy, scalability, development cost, development time, interpretability, and risk (bias/ethics).

Table 1. Comparative Evaluation of Chatbot Development

Criterion	Rule-Based	Retrieval-Based	Generative-Based
Flexibility	Low	Medium	High
Accuracy	Medium	High	High
Scalability	Low	Medium	High
Development Cost	Low	Medium	Very High
Development Time	Short	Medium	Long
Interpretability	High	Medium	Low
Risk (Bias/Ethics)	Low	Medium	High

Journal of Engineering, Technology & Applied Science, vol. 7, no. 2, pp. 80-92, August 2025. **DOI: 10.36079/lamintang.jetas-0702.881**

1) Flexibility
   Rule-based chatbots offer low flexibility because they rely strictly on preprogrammed IF-THEN rules and cannot handle variations in user input. Retrieval-based models perform moderately better, using semantic matching to recognize similar intents. Generative-based chatbots demonstrate the highest flexibility, as they generate responses dynamically and adaptively, accommodating diverse phrasing and open-ended conversations.
2) Accuracy
   In terms of accuracy, rule-based systems are reliable only when the user input exactly matches predefined patterns, making their accuracy inconsistent. Retrieval-based systems typically achieve high accuracy by selecting the best-fit response from a curated database, especially in structured tasks. Generative-based systems also score high, producing contextually accurate responses, though they can occasionally generate factually incorrect or misleading answers.
3) Scalability
   Rule-based chatbots are difficult to scale because each new interaction scenario requires manually adding more rules, which becomes unmanageable over time. Retrieval-based chatbots scale more easily by expanding their response database, though performance may degrade with larger datasets unless optimized. Generative-based chatbots are highly scalable, as a single model can handle a wide range of tasks and domains without the need for manual updates.
4) Development Cost:
   Rule-based chatbots incur the lowest development costs due to their simplicity and minimal technical requirements. Retrieval-based systems involve moderate costs for building and maintaining a high-quality response bank and integrating embedding techniques. Generative-based models demand the highest investment, requiring powerful hardware, large-scale datasets, and advanced expertise for training, deployment, and monitoring.
5) Development Time:
   The development time for rule-based systems is short since they can be built quickly for simple use cases. Retrieval-based models take more time due to the need for response curation and integration of semantic search algorithms. Generative-based chatbots require the longest time to develop because of their complex architectures, data preparation, fine-tuning, and evaluation cycles.
6) Interpretability:
   Rule-based systems are highly interpretable because each response is directly tied to a specific rule, making behavior transparent and easy to debug. Retrieval-based chatbots offer moderate interpretability, developers can trace selected responses, but semantic similarity decisions can be opaque. Generative models have low interpretability due to their deep neural architectures, making it difficult to understand why a particular response was generated.
7) Risk (Bias/Ethics):
   Rule-based chatbots present minimal ethical or bias risks because all outputs are manually controlled and predetermined. Retrieval-based systems carry medium risk depending on the quality and neutrality of the stored responses. Generative-based chatbots pose the highest risk, as they can produce inappropriate, biased, or harmful content unless rigorously monitored and filtered, especially in sensitive domains like healthcare or education.

The comparative analysis of rule-based, retrieval-based, and generative-based chatbots reveals distinct strengths and limitations across key development dimensions. Rule-based systems offer simplicity, low cost, and high interpretability but lack flexibility and scalability, making them best suited for static, structured scenarios. Retrieval-based chatbots strike a balance between accuracy and development efficiency, performing well in domain-specific contexts, yet limited by their inability to generate novel responses. In contrast, generative-based chatbots deliver the highest levels of conversational coherence, adaptability, and emotional intelligence, as exemplified by ChatGPT, but at the cost of significant computational demands, low interpretability, and increased ethical risks. These findings underscore the need to align chatbot architecture with the specific demands of real-world applications, where hybrid approaches may offer an optimal blend of control, performance, and contextual sensitivity.

5. Conclusion

This study provides a comparative analysis of major chatbot development approaches, rule-based, retrieval-based, and generative-based, based on case studies from Mitsuku, Google Assistant, and ChatGPT, as applied in real-world contexts in Indonesia and Malaysia. Through a descriptive-qualitative methodology and comparative framework, the findings show that each method carries distinct strengths and trade-offs across key criteria such as flexibility, accuracy, scalability, development cost, development time, interpretability, and ethical risks.

Rule-based chatbots demonstrate high interpretability and low development cost, making them suitable for static, well-defined tasks such as FAQs. However, their rigid IF-THEN logic severely limits their adaptability and scalability, especially in contexts requiring nuanced language understanding. Retrieval-based systems offer a balanced approach, achieving high accuracy in structured tasks by selecting responses from a predefined database. Nevertheless, their inability to generate novel or contextually adaptive responses restricts their use in open-ended dialogue scenarios. In contrast, generative-based chatbots provide the highest levels of coherence, flexibility, and conversational depth, making them ideal for education, counseling, and open-domain applications. However, they involve substantial computational resources, lower interpretability, and heightened ethical risks.

The analysis answers the central research question by demonstrating that no single approach is universally superior. Instead, effective chatbot development depends on aligning technical strategies with the specific demands of the application domain. The growing relevance of hybrid models, which combine the rule-based system's control, the retrieval model's efficiency, and the generative model's adaptability, emerges as a promising direction for balancing performance and responsibility.

This study also identifies critical gaps that merit future research. These include: (1) improving the interpretability of generative models without compromising fluency, (2) enhancing ethical safeguards to minimize biased or harmful outputs, and (3) exploring adaptive hybrid architectures that can switch dynamically between rule-based, retrieval, and generative modes. Further empirical validation through user experience studies and region-specific testing particularly in Southeast Asia, can strengthen the practical deployment of conversational AI. Addressing these gaps will be essential for building transparent, scalable, and ethically responsible chatbot systems that serve diverse user needs.

References

[1] G. S. Sekhon, "Chatbot Development: Techniques and Technologies," *Medium*, 2024. [Online]. Available: https://medium.com/@gianetan/chatbot-development-techniques-and-technologies-fb2fd60a37b2. [Accessed: Jan. 10, 2025]

[2] A. Martins, A. Londral, I. L. Nunes, and L. V. Lapão, "Unlocking human-like conversations: Scoping review of automation techniques for personalized healthcare interventions using conversational agents," *International Journal of Medical Informatics*, vol. 185, 105385, May 2024.

[3] A. Casheekar, A. Lahiri, K. Rath, K. S. Prabhakar, and K. Srinivasan, "A contemporary review on chatbots, AI-powered virtual conversational agents, ChatGPT: Applications, open challenges and future research directions," *Computer Science Review*, vol. 52, 100632, May 2024.

[4] M. M. Alam, A. A. Khan, A. Ali, dan M. Imran, "A contemporary review on chatbots, AI-powered virtual conversational agents, ChatGPT: Applications, open challenges and future research directions," *Journal of Network and Computer Applications*, vol. 220, 2024.

[5] O. Dogan and O. F. Gurcan, "Enhancing E-Business Communication with a Hybrid Rule-Based and Extractive-Based Chatbot," *J. Theor. Appl. Electron. Commer. Res.*, vol. 19, no. 3, pp. 1984–1999, 2024.

[6] S. Hou, S. Zhang, and C. Fei, "Rhetorical structure theory: A comprehensive review of theory, parsing methods and applications," *Expert Syst. Appl.*, vol. 157, art. no. 113421, Nov. 2020.

[7] A. C. Cullen, B. I. P. Rubinstein, K. Sithamparanathan, B. Flower, and P. H. W. Leong, "Predicting dynamic spectrum allocation: a review covering simulation, modelling, and prediction," *Artif. Intell. Rev.*, vol. 56, no. 10, pp. 10921–10959, Mar. 2023.

[8] M. A. Kuhail, N. Alturki, S. Alramlawi, and K. Alhejori, "Interacting with educational chatbots: A systematic review," *Education and Information Technologies*, vol. 28, no. 1, pp. 973–1018, 2023,

[9] Messenger Bot, "Understanding Rule-Based Chatbots: Key Differences, Types, and Limitations," Messenger Bot, 2024.
[10] F. Hafeez, "Conversational AI vs Traditional Rule-Based Chatbots: A Comparative Analysis," *Artificial Intelligence, Technology*, May 3, 2024.
[11] Isabella, "Rule-Based vs. AI Chatbot: Which One is Better?," *GoInsight.AI-AI Insights*, Oct. 30, 2023.
[12] E. Solomon and S. L. Tilahun, "Rule based chatbot design methods: A review," *J. Comput. Sci. Data Anal.*, vol. 1, no. 1, pp. 75–84, Sep. 2024.
[13] G. H. Setiawan and I. B. Adnyana, "Improving Helpdesk Chatbot Performance with Term Frequency-Inverse Document Frequency (TF-IDF) and Cosine Similarity Models," *Journal of Applied Informatics and computing*, vol. 7, no. 2, 2023.
[14] Y. Wu, Z. Li, W. Wu, and M. Zhou, "Response selection with topic clues for retrieval-based chatbots," *Neurocomputing*, vol. 316, pp. 251–261, Nov. 2018,
[15] S. Pandey and S. Sharma, "A comparative study of retrieval-based and generative-based chatbots using Deep Learning and Machine Learning," *Healthcare Anal.*, vol. 3, Nov. 2023.
[16] H. T. Y. Achsan, D. Kurniawan, D. G. Purnama, Q. K. D. Barcah, and Y. Y. Astoria, "Application of Natural Language Processing Using Cosine-Similarity Algorithm in Making Chatbot Information on the New Capital City of the Republic of Indonesia," in *Proc. 7th Int. Workshop Comput. Sci. Eng. (WCSE)*, 2022.
[17] A. Bandi, P. V. S. R. Adapa, and Y. E. V. P. K. Kuchi, "The Power of Generative AI: A Review of Requirements, Models, Input–Output Formats, Evaluation Metrics, and Challenges," *Future Internet*, vol. 15, no. 8, p. 260, 2023.
[18] L. Benaddi, C. Ouaddi, I. Khriss, and B. Ouchao, "Analysis of tools for the development of conversational agents," *Computer Sciences and Mathematics Forum*, vol. 6, no. 5, 2023.
[19] G. Bilquise, S. Ibrahim, and K. Shaalan, "Emotionally intelligent chatbots: A systematic literature review, *Human Behavior and Emerging Technologies*, vol. 2022.
[20] I. Ortiz-Garces, J. Govea, R. O. Andrade, and W. Villegas-Ch, "Optimizing Chatbot Effectiveness through Advanced Syntactic Analysis: A Comprehensive Study in Natural Language Processing," *Applied Sciences*, vol. 14, no. 5, 2024.
[21] M. Abbasian, E. Khatibi, I. Azimi, D. Oniani, Z. S. Hossein Abad, A. Thieme, R. Sriram, Z. Yang, Y. Wang, B. Lin, O. Gevaert, L.-J. Li, R. Jain, and A. M. Rahmani, "Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI," *npj Digital Medicine*, vol. 7, no. 82, 2024.
[22] L. Labadze, M. Grigolia, and L. Machaidze, "Role of AI chatbots in education: systematic literature review," *International Journal of Educational Technology in Higher Education*, vol. 20, no. 56, 2023.
[23] M. Laymouna, Y. Ma, D. Lessard, T. Schuster, K. Engler, and B. Lebouché, "Roles, Users, Benefits, and Limitations of Chatbots in Health Care: Rapid Review," *Journal of Medical Internet Research*, vol. 26, Jul. 2024.
[24] H. S. M. Alsultani and A. H. Aliwy, "Improving Arabic Named Entity Recognition with a Modified Transformer Encoder," *Journal of Computer Science*, vol. 19, no. 5, pp. 599–609, 2023.
[25] M. A. Ali, A. B. Alwahhab, and Y. Farjami, "An Integrated Deep Learning Framework Combining LSTM-CRF, GRU-CRF, and CNN-CRF with Word Embedding Techniques for Arabic Named Entity Recognition," *Int. J. Robot. Control Syst.*, vol. 5, no. 2, pp. 937–952, Mar. 2025.
[26] G. Absalamova and D. Absalamova, "Optimizing an AI-driven chatbot through natural language processing and real-time feedback for personalized recommendations," in *Proc. 1st Int. Sci.-Pract. Conf. on Digital Transformation and Artificial Intelligence: Problems, Innovations and Trends*, vol. 1, no. DTAI, Section 5, 2024.
[27] A. Schenk, V. Klockmann, and N. Köbis, "Social preferences toward humans and machines: A systematic experiment on the role of machine payoffs," *Perspectives on Psychological Science*, vol. 20, no. 1, pp. 165–181, 2023.
[28] S. Yang, X. Yu, and Y. Zhou, "LSTM and GRU Neural Network Performance Comparison Study: Taking Yelp Review Dataset as an Example," in *Proc. 2020 Int. Workshop on Electronic Communication and Artificial Intelligence (IWECAI)*, Taizhou, China, 2020.

[29] M. Kulkarni, K. Kim, N. Garera, and A. Trivedi, "Label efficient semi-supervised conversational intent classification," in *Proc. 61st Annu. Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, Toronto, Canada, Jul. 2023.

[30] Z. Chen, L. Xu, H. Zheng, L. Chen, A. Tolba, L. Zhao, K. Yu, and H. Feng, "Evolution and prospects of foundation models: From large language models to large multimodal models," *Computers, Materials & Continua*, vol. 80, no. 2, pp. 1753–1808, Aug. 2024.

[31] G. Bourahouat, M. Abourezq, and N. Daoudi, "Word Embedding as a Semantic Feature Extraction Technique in Arabic Natural Language Processing: An Overview," *The International Arab Journal of Information Technology*, vol. 21, no. 2, pp. 313–325, 2024.

[32] E. Barbierato and A. Gatti, "The challenges of machine learning: A critical review," *Electronics*, vol. 13, no. 2, p. 416, 2024.