International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Programmable NLU Powered 3-in-1 Bot :Video, Voice, Text

Ramesh Alladi[1], Nagaraju Boyapally[2],
A Arun Kumar Reddy[3], Samudrala Narain[4], Bommaraju Nikhil Sai[5]

Associative Professor[1], Students[2,3,4,5],
Department of Computer Science and Engineering,
ACE Engineering College, Ghatkesar, Hyderabad, Telangana, India

---

Abstract - This study introduces a Programmable NLU-Powered 3-in-1 ChatBot, a conversational AI framework designed to enhance response precision and contextual relevance through the integration of multiple specialized language models. The system comprises three distinct large language models (LLMs), each optimized for specific interaction domains: general conversation, problem-solving, and technical inquiry. A fourth LLM operates as a natural language understanding (NLU) controller, responsible for analyzing user input and determining the most appropriate model to handle the query. Developed using a modular architecture, the chatbot supports real-time interaction, session-level context retention, and adjustable response creativity. By employing a model selection mechanism guided by semantic analysis, the framework ensures that responses are not only contextually aligned but also domain-appropriate. This study highlights the effectiveness of combining specialized LLMs with an intelligent routing layer to improve the adaptability and accuracy of multi-domain conversational systems. The study represents a significant step toward redefining virtual assistant capabilities by offering an engaging, multi-modal interface that enhances user interaction and information retention.

Key Words: Natural Language Processing (NLP), Multi-Model ChatBot, Large Language Models (LLMs), Domain-Specific Language Models, Conversational AI.

1. INTRODUCTION

The evolution of artificial intelligence, particularly in the field of Natural Language Processing (NLP), has transformed how humans interact with machines. Conversational agents—commonly referred to as chatbots—have progressed from rigid, rule-based systems to advanced neural architectures capable of maintaining contextual understanding, generating fluent text, and performing complex tasks. This shift has been primarily driven by the development of large language models (LLMs) such as OpenAI's GPT [1], Google's BERT [2], and Meta's LLaMA [3], which are pretrained on vast datasets and demonstrate remarkable generalization across a wide range of tasks.

Despite these advancements, most modern chatbots rely on a single LLM to manage all types of user interactions. While these models perform reasonably well in general dialogue, they often exhibit limitations in domain-specific or task-oriented contexts. For example, a model optimized for casual conversation may struggle to accurately address a technical question, whereas a model fine-tuned for problem-solving may not handle open-ended or empathetic dialogues effectively. This "one-size-fits-all" limitation leads to challenges in delivering consistently relevant and context-aware responses.

To address this gap, this study explores a multi-model conversational framework that integrates three specialized LLMs—each targeting a distinct interaction domain: general conversation, analytical reasoning, and technical query resolution. Central to this architecture is a fourth LLM that functions as a semantic interpreter or NLU controller. Its role is to analyze the user input, identify the underlying intent, and dynamically select the most suitable model to generate an appropriate response. This routing mechanism allows the system to combine the strengths of each individual model, thereby improving precision, relevance, and domain adaptability.

The concept of utilizing multiple expert models for different sub-tasks has been investigated in prior studies. Approaches such as Toolformer [5] and RETRO [7] show the advantages of combining modular reasoning with retrieval or tool-based augmentation. Additionally, systems like Kosmos-1 [6] and BlenderBot [4] emphasize the need for aligning multimodal understanding with response generation. Inspired by these works, the current study adopts a modular, LLM-driven strategy with a focus on text-based interaction and intelligent query delegation.

The increasing demand for AI-driven conversational agents has led to the development of various chatbot systems designed to handle a wide range of user queries. However, most existing chatbots rely on a single language model, limiting their ability to cater to diverse conversation types, ranging from casual interactions to complex problem-solving and technical discussions. As a result, there is a growing need for chatbots that can intelligently adapt to different contexts and provide tailored responses based on the user's specific needs. A multi-model approach, where different models specialize in various domains of knowledge, offers a promising solution to this challenge. By leveraging multiple language models optimized for general conversation, technical queries, and problem-solving, the system can dynamically switch between models to provide more accurate and relevant responses, ensuring a more personalized and engaging user experience. This system proposes a novel framework for such a system, integrating advanced natural language understanding (NLU) techniques with a programmable, multi-LLM architecture to facilitate seamless transitions between specialized models. Furthermore, the proposed system supports real-time interaction, adjustable creativity or temperature settings, and conversational memory that maintains context across dialogue turns.

These features are intended to enhance the user experience, especially in scenarios where clarity, continuity, and relevance are critical—such as educational tutoring, software support, or interactive knowledge bases.

This study contributes to the growing body of system focused on adaptive, multi-domain conversational agents. By leveraging the synergy between specialization and intelligent orchestration, the proposed chatbot architecture demonstrates improved

performance over traditional monolithic models. The findings suggest that such multi-model systems may play a significant role in the next generation of scalable and flexible human-AI interfaces.

2. Related Work

The field of multimodal conversational AI has seen significant advancements in recent years, with various studies focusing on improving the capabilities and efficiency of chatbots through the integration of diverse modalities such as text, speech, and vision. Ma et al. (2025) explore the integration of multimodal machine learning in AI-driven chatbots for diagnosing ophthalmic diseases. Their system combines textual and image-based inputs to provide accurate medical diagnoses. This work exemplifies the potential for multimodal AI to significantly enhance domain-specific chatbots, particularly in fields where visual information plays a crucial role, such as medical diagnostics. The ability to integrate vision and language models offers an effective framework for handling tasks that require deep domain knowledge, similar to how the current study proposes specialized models for technical, analytical, and general-purpose queries. However, the current study extends this approach by focusing on a broader conversational domain rather than a specific medical field [8].

Zhang et al. (2024) present a comprehensive evaluation of how multi-modal interactions affect user engagement in AI-driven conversations. Their study investigates how the incorporation of different interaction modalities influences user satisfaction, trust, and engagement. This aligns with the focus of the current system, which aims to enhance the user experience by providing more adaptive and engaging responses through a multi-LLM framework. While their study primarily focuses on user engagement with varied multimodal inputs, the study here proposes a dynamic routing mechanism for directing user queries to specialized models based on intent, which can further personalize the user experience and improve engagement [9].

Chen et al. (2024) explore the performance of multimodal AI chatbots in clinical oncology cases. By integrating visual and textual data, their study emphasizes the use of AI for providing high-quality medical support. This work contributes to the understanding of how multimodal chatbots can serve as effective tools for specialized professional fields, echoing the direction of the current study in its effort to employ multiple expert models for specialized tasks. However, while their system targets medical applications, the proposed study takes a more general approach by exploring the use of task-specific models across a variety of domains, demonstrating the scalability of multimodal techniques across various contexts [10].

Bar (2024) provides a practical developer guide for creating multimodal chatbots using LangChain agents. This work highlights the importance of modularity and flexibility in building chatbots that can seamlessly integrate multiple tools and models. The framework discussed in this guide is particularly relevant to the current study's architecture, as the modularity of LangChain agents allows for smooth switching between specialized models. The proposed study builds on this idea by incorporating a fourth model that functions as a semantic controller to intelligently route user queries to the most appropriate expert model, adding an additional layer of flexibility and performance optimization [11].

Lee (2023) discusses the foundational principles behind constructing multimodal AI chatbots, emphasizing the integration of various sensory inputs, such as text, speech, and images, to improve conversational capabilities. This study provides insights into the technical and architectural challenges associated with building multimodal systems. The current study expands upon these principles by focusing on the use of multiple language models rather than just text, allowing for a more nuanced approach to conversation management. By incorporating specialized LLMs for different domains, the proposed chatbot can select the most relevant model for each user query, a feature not fully explored in Lee's work [12].

3. Problem Formulation

The problem of designing a versatile conversational AI system capable of handling a diverse range of user queries has gained significant attention in recent years. Current chatbots often rely on a single language model, which limits their ability to adapt to different conversational contexts such as general dialogue, technical problem-solving, or specialized queries. These systems may struggle when faced with complex or multi-domain requests, often providing generic or less relevant responses. While some models have been developed to handle complex queries, they typically lack the flexibility to seamlessly switch between multiple domains, leading to suboptimal user experiences in cases where a query spans multiple areas of expertise.

A key challenge in addressing this problem is the development of a system that can intelligently route user queries to the most appropriate specialized model based on the context and content of the query. To achieve this, a mechanism is needed that can analyze the user input, understand the underlying intent, and select the language model that is best suited to handle the request. This dynamic routing between models requires not only analyzing the text but also ensuring that the model transition is smooth, preserving the conversation's context and maintaining continuity. It adds a layer of complexity that is not present in single-model systems, where responses are generated based on predefined logic.

Additionally, providing personalized and adaptive responses is essential for improving user engagement and satisfaction. Current conversational AI systems often rely on a fixed set of responses or have limited adaptability, which may not adequately meet the needs of every user. In contrast, a more flexible approach is required—one that adjusts the level of creativity or specificity of the responses in real-time based on the user's input. Achieving this adaptability requires a mechanism capable of analyzing both the query's content and the broader context of the ongoing conversation.

The goal of this system is to address these challenges by developing a Programmable NLU-Powered 3-in-1 Chatbot that leverages multiple specialized language models (LLMs) for different tasks: general conversation, problem-solving, and technical queries. Additionally, a fourth model will be introduced to process the query and intelligently route it to the appropriate specialized LLM based on the query’s content and intent. This intelligent routing mechanism will enable the system to handle a wide range of queries more effectively, ensuring responses are accurate, contextually relevant, and appropriate for the task at hand.

In summary, the primary problem this study aims to solve is the development of a flexible and intelligent conversational AI system capable of dynamically switching between multiple specialized language models. The system must analyze user queries to select the most appropriate model, ensure context retention across interactions, and provide adaptive responses that enhance user engagement and satisfaction. By addressing these challenges, this study aims to create a more effective conversational AI framework capable of handling complex, domain-specific interactions with improved accuracy and user experience.

4. PROPOSED METHODOLOGY

The proposed methodology for this system focuses on the development of a dynamic, multi-model chatbot system powered by natural language understanding (NLU) techniques. This system aims to leverage multiple specialized language models (LLMs) to handle different types of user queries, such as general conversation, technical problem-solving, and domain-specific inquiries. The methodology consists of several key components that enable the system to analyze, route, and generate responses based on the context of the conversation and the nature of the query.

The first step in the methodology involves the integration of three distinct LLMs, each optimized for a specific purpose. The first LLM is focused on handling general conversation and casual interactions, capable of engaging users in light discussions. The second LLM is designed for technical queries, providing responses based on specialized knowledge and more complex problem-solving tasks. The third LLM is optimized for domain-specific queries, such as providing information on particular industries, services, or technical topics. These models are selected based on their expertise and are pre-trained on datasets tailored to their respective domains.

To manage the interaction between these models, a fourth model is introduced that acts as an intelligent query router. This model is responsible for analyzing the user's input and determining which of the three specialized LLMs is best suited to respond to the query. The query router employs NLU techniques, such as intent classification and semantic analysis, to assess the content and context of the user's request. This ensures that the response provided by the system is accurate, contextually relevant, and generated by the most appropriate model.

Once the query is routed to the correct LLM, the response is generated and returned to the user. Throughout the interaction, the system is designed to maintain the context of the conversation. This ensures that the chatbot remembers the previous interactions, allowing for a coherent flow of conversation even when switching between different models. Techniques such as session memory and context tracking are utilized to maintain this continuity, ensuring that each response builds on the previous one, regardless of which model is handling the query.

The methodology also includes the consideration of adaptability and personalization. The system dynamically adjusts the level of creativity or specificity in its responses based on the user's preferences and the nature of the query. For example, more open-ended queries may receive responses with greater creative freedom, while technical queries are expected to be more precise and factual. This adaptability is achieved by tuning the parameters of the models in real-time, ensuring that the user's experience is both engaging and relevant.

In summary, the proposed methodology integrates four key components: multiple specialized LLMs for different types of queries, an intelligent query router to dynamically route user inputs to the appropriate model, context tracking to ensure continuity across interactions, and adaptability to provide personalized responses. This combination of techniques ensures that the system can provide accurate, relevant, and engaging responses to a wide variety of user queries, creating a flexible and efficient conversational AI solution.

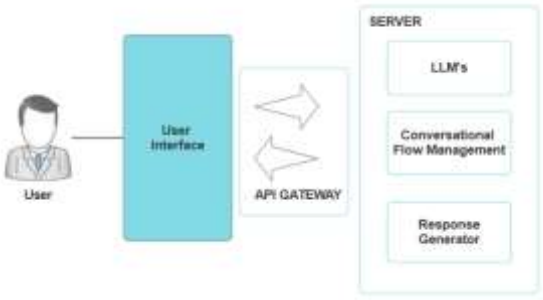

Fig-1: Architecture Diagram

To better illustrate the workflow, a architectural representation and flowchart representation has been provided in the fig 1 and fig 2 respectively. This diagram visually maps out the end-to-end process of user interaction with the AI-driven system.

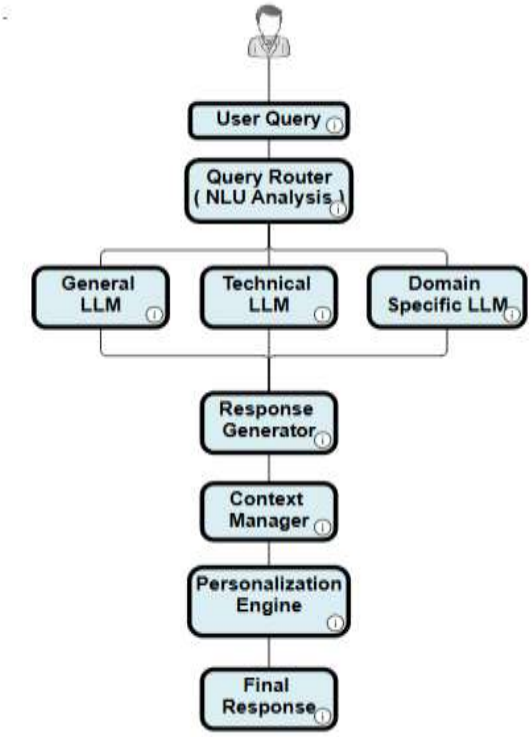

Fig-2: Flow Diagram

5. IMPLEMENTATION

The proposed chatbot system employs a modular architecture integrating multiple large language models (LLMs), each specialized for different tasks. Built with the Streamlit framework, the system offers an interactive, web-based user interface where users can seamlessly communicate with the chatbot. Through the integration of the Hugging Face Inference API, the system interacts with various pre-trained models that serve different purposes, such as general conversation, technical problem-solving, and creative content generation.

A standout feature of the system is the use of a specialized "selector" LLM. This selector model is designed to analyze user input and determine the most appropriate model to respond based on the query's nature. For instance, if a user submits a technical question, the selector model directs the query to an LLM focused on coding and logical reasoning. On the other

hand, a general conversation or casual query would be routed to a conversational model, while creative requests would be handled by a model optimized for generating creative content. This dynamic selection mechanism enhances the relevance and accuracy of the chatbot's responses. The system includes a range of specialized models that users can choose from based on their needs. These models include a general-purpose conversational model for casual interactions, a technical model tailored for coding and problem-solving, and a creative model designed to generate content. Each model is calibrated to perform optimally within its domain, ensuring that the chatbot can provide precise and useful responses across a variety of contexts.

Additionally, the chatbot allows users to adjust the "temperature" setting for the selected model. The temperature controls the creativity of the model's responses—higher values result in more diverse and creative outputs, while lower values generate more focused and deterministic answers. This customization gives users greater control over the type of response they receive, enhancing the user experience.

The chatbot interface also maintains session memory, storing the conversation history to ensure continuity across multiple interactions. This allows the system to provide more contextually relevant responses as the conversation progresses. Users can reset the chat at any time, clearing the session history for a fresh start. This feature ensures flexibility in user interactions, catering to both ongoing conversations and new queries.

By integrating multiple LLMs and using a specialized selector model, the system ensures that the user receives the most appropriate and relevant responses for their specific query. The dynamic selection of models, based on the nature of the query, provides a tailored experience that adapts to the user's needs. This modular approach, combined with real-time interactions powered by Hugging Face's API, delivers an efficient and responsive conversational AI system.

6. Performance Evaluation

The performance evaluation of the proposed 3-in-1 ChatBot system was carried out using a structured two-fold methodology aimed at offering both quantitative and qualitative insights into the system's effectiveness. This evaluation focused on response accuracy, contextual relevance, coherence, and system latency (response time).

Table-1: Comparsion of response times of three individual models across three different queries

Query	Dorado	Hercules	Lepus
Hello, how are you?	1.82ms	2.86ms	8.52ms
Write code for hello world program	2.99ms	2.06ms	2.43ms
Tell me about software development	4.29ms	2.41ms	2.67ms

The first phase of the evaluation employed automated testing across the three individual models—Dorado, Hercules, and Lepus—integrated within the system. Each model was tested using a uniform set of queries, and their respective response times were recorded. The automated tests allowed for the identification of unique performance characteristics for each model and offered a precise comparison across a standardized environment. As shown in Table 1, the response times varied across models, reflecting differences in model architecture and computational complexity.

The second phase involved manual testing of the entire 3-in-1 ChatBot system. This phase simulated real-world user interactions to evaluate the system's performance in comparison with widely used conversational agents such as ChatGPT, Gemini, and Grok. The same set of benchmark queries was submitted to all systems to ensure fair comparison. The evaluation considered multiple dimensions, including latency, contextual understanding, and the coherence of responses. The results of this comparative study are presented in Table 2.

Table-2: 3-in-1 Bot VS Other Chatbot Response Times

Query	3-in-1 Bot	ChatGPT	Gemini	Grok
Hello, how are you?	0.95ms	1.53ms	15.38ms	36.85ms
Write code for hello world program	3.50ms	4.95ms	20.85ms	13.75ms
Tell me about software development	3.75ms	5.78ms	8.95ms	15.98ms

While the 3-in-1 ChatBot demonstrated competitive accuracy and contextual relevance, it exhibited marginally higher response times than the benchmark models. This increased latency is primarily due to the dynamic model selection mechanism, where the system internally decides which of the integrated models (Dorado, Hercules, or Lepus) is best suited for a particular query. This decision-making logic, while adding a slight overhead, significantly enhances the chatbot's flexibility and specialization, allowing it to adapt better to various query types—be it casual conversations, technical prompts, or development-related requests.

Moreover, the system's architecture, which relies on API-based inference using Hugging Face's hosted models, introduces an additional layer of latency. However, this trade-off is justified by the system's modular design, which promotes scalability, reusability, and ease of maintenance. The specialized routing of queries to appropriate models enables high-quality and context-sensitive responses, supporting the hypothesis that multi-model systems can outperform single-model systems in complex, multi-domain interactions. In summary, the combined results of both evaluation methods underscore the practical strengths of the 3-in-1 ChatBot system. Its ability to deliver reliable, accurate, and context-aware responses across a wide spectrum of user inputs validates the use of a modular, multi-model framework for conversational AI. The response time trade-offs are acceptable, particularly in scenarios where accuracy, specialization, and adaptability take precedence over minimal latency.

7. Conclusion and Future Work

The development of the 3-in-1 ChatBot marks a significant advancement in the domain of conversational AI. By integrating multiple Hugging Face models—each with unique strengths—into a unified interface, the system delivers versatile and contextually accurate responses across a wide range of queries. Performance evaluations, both automated and manual, have demonstrated the ChatBot's capability to provide relevant and informative answers, making it a strong alternative to existing platforms like ChatGPT, Grok, and Gemini. The modular design also enables adaptive model selection, further enhancing the quality of user interaction.

Despite its advantages, the current system does show slightly higher response times due to API-based model invocation and dynamic selection mechanisms. However, this trade-off is offset by its flexibility and superior handling of diverse query types. The chatbot is particularly beneficial in use cases where depth, clarity, and accuracy of responses are prioritized over minimal latency. Another key advantage of this architecture lies in its scalability. As new and more powerful language models become available, the modular framework allows them to be integrated seamlessly without disrupting the core logic of the system. This future-proofs the ChatBot and ensures it remains relevant and up-to-date in the rapidly evolving field of AI-powered communication. Moreover, the idea of using one specialized LLM to orchestrate the selection of other models opens up further possibilities for intelligent model routing and context-aware decision-making.

Future enhancements to the 3-in-1 ChatBot are focused on integrating multimodal capabilities such as image generation, voice-based interaction, and video-based outputs. These additions are expected to enrich user engagement by enabling more intuitive and expressive modes of communication. Further system will also aim at optimizing the response pipeline to reduce latency, thereby improving the system's responsiveness. Additionally, incorporating an adaptive learning mechanism is envisioned to facilitate greater personalization and contextual adaptability. These developments will extend the applicability and intelligence of the 3-in-1 ChatBot, making it a more versatile and impactful solution across a wide range of real-world scenarios.

8. REFERENCES

[1]. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165.

[2]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171-4186).

[3]. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Jegou, H. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971.

[4]. Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., ... & Weston, J. (2021). Recipes for Building an Open-Domain Chatbot. arXiv preprint arXiv:2004.13637.

[5]. Schick, T., Dwivedi-Yu, J., Hosseini, S., Goyal, N., & Srivastava, A. (2023). Toolformer: Language Models Can Teach Themselves to Use Tools. arXiv preprint arXiv:2302.04761.

[6]. Huang, P. S., Bapna, A., Chen, B., Huang, Y., Lee, J., Qian, C., ... & Wang, W. (2023). Language Is Not All You Need: Aligning Perception with Language Models. arXiv preprint arXiv:2302.14045.

[7]. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., ... & Irving, G. (2022). Improving language models by retrieving from trillions of tokens. In Proceedings of the 39th International Conference on Machine Learning (ICML).

[8]. Ma, R., Cheng, Q., Yao, J., Peng, Z., Yan, M., Lu, J., Liao, J., Tian, L., Shu, W., Zhang, Y., Wang, J., Jiang, P., Xia, W., Li, X., Gan, L., Zhao, Y., Zhu, J., Qin, B., Jiang, Q., Wang, X., Lin, X., Chen, H., Zhu, W., Xiang, D., Zhao, C. (2025). Multimodal machine learning enables AI chatbot to diagnose ophthalmic diseases and provide high-quality medical responses. Nature.

[9]. Zhang, L., Yu, J., Zhang, S., Li, L., Zhong, Y., Liang, G., Yan, Y., Ma, Q., Weng, F., Pan, F., Li, J., Xu, R., Lan, Z. (2024). Unveiling the Impact of Multi-Modal Interactions on User Engagement: A Comprehensive Evaluation in AI-driven Conversations. Zhejiang University, Westlake University, Westlake Xinchen Technology Co. Ltd.

[10]. Chen, D., Huang, R. S., Jomy, J., Wong, P., Yan, M., Croke, J., Tong, D., Hope, A., Eng, L., Raman, S. (2024). Performance of Multimodal Artificial Intelligence Chatbots Evaluated on Clinical Oncology Cases. University of Toronto.

[11]. Bar, N. (2024). A Developer Guide for Creating a Multi-Modal Chatbot Using LangChain Agents. Medium.

[12]. Lee, M. Y. (2023). Building Multimodal AI Chatbots. arXiv.

[13]. Ahmad Abdellatif, Khaled Badran, Diego Costa, Emad Shihab. A Comparison of Natural Language Understanding Platforms for Chatbots in Software Engineering. IEEE.2020

[14]. E. Pratt. Artificial Intelligence and Chatbots in Technical Communication – A Primer. Iiblog'2017.

[15]. A. Abdellatif, D. E. Costa, K. Badran, R. Abdelkareem, E. Shihab. "Challenges in Chatbot Development:" A Study of Stack Overflow Posts, 2020