음성 인식 키오스크를 위한 대화형 AI 모델 LLaMA의 효용성 평가: RASA와의 비교 사례 연구

니스타[1], 카라쿠르트 메흐메트 파티흐[1], 조영균[1], 이준용[1], 정설영[1], 류지수[2]
[1]경북대학교 컴퓨터학부, [2]드림아이디어소프트
Lathnishtha775@gmail.com , m.fatih012001@gmail.com, cyg0828@gmail.com , harmlessman17@gmail.com, snowflower@knu.ac.kr, gnt700@dreamideasoft.net

Evaluating LLaMA Model for Enhanced Conversational AI in Voice Recognition Kiosks: A Case Study on RASA vs. LLaMA

Lath Nishtha[1], Karakurt Mehmet Fatih[1], Yeonggyun Cho[1], leejunyong[1],
Jeong Seolyeong[1], Ryu Jisoo[2]
[1]Kyungpook National University, School of Computer Science and Engineering,
[2]Dream Idea Soft

Abstract

This paper evaluates the LLaMA language model as an alternative to the RASA framework for conversational AI in voice recognition kiosks. Voice-activated ordering systems are increasingly used in retail, yet existing frameworks like RASA struggle with complex interactions. This study compares the performance of LLaMA and RASA in handling diverse user commands and complex order modifications. The results show LLaMA's superior accuracy, flexibility, and scalability, highlighting its potential for enhancing user experience in customer-focused environments.

1.Introduction

The implementation of conversational AI in public kiosks is transforming customer interactions, especially in service sectors like cafes and restaurants. While conversational AI enables hands-free, efficient ordering processes, many existing models struggle to manage nuanced language comprehension and complex customer commands. Traditional frameworks, such as RASA, are effective for basic intent classification but often lack the contextual understanding required for more intricate orders and modifications.

This paper explores the application of LLaMA, an advanced language model, in building a cafe ordering system capable of handling flexible, context-rich interactions. The objective is to evaluate LLaMA's effectiveness compared to RASA and demonstrate its suitability as a robust model for voice-activated interactions in customer-facing kiosks. This study aims to:

1. Analyze and compare the capabilities of RASA and LLaMA in processing diverse user commands.
2. Highlight technical and user experience improvements achieved with LLaMA.
3. Demonstrate the benefits of using an advanced language model for enhancing conversational AI in retail environments.

2. Literature Review

2.1 Evolution of Conversational AI Models

Conversational AI has evolved from rule-based systems to complex models supporting flexible interactions. Earlier rule-based responses and keyword detection, in models limited their effectiveness for nuanced conversations. Later with introduction to ML based models, such as RASA, intent classification and entity recognition improved but remains limited by its approach.

2.2 RASA's Framework in Conversational AI

RASA, an open-source framework for the purpose, has gained popularity due to its structured approach to intent classification and entity extraction. It is effective for applications with a defined set of commands, where interactions are relatively predictable. However, Studies suggest that RASA's reliance on predefined intent limits its adaptability and model's performance is constrained in dynamic environments, such as customer kiosks, where requests may vary widely and require greater contextual

interpretation.[1]

2.3 LLaMA's Capabilities in Language Modeling

LLaMA, an LLM model, supports open-ended conversations and offers a more sophisticated understanding of a broader array of linguistic patterns, making it capable of managing diverse inputs and generating relevant responses. Research indicates that LLaMA's extensive language modeling makes it more flexible and adaptive, positioning it as an effective solution for voice-activated kiosks where nuanced and complex response capabilities are required. [2]

2.4 Rationale for Choosing LLaMA over RASA

While RASA excels in structured, task-oriented dialogues, its limitations become evident in scenarios requiring nuanced understanding and adaptability. LLaMA's capabilities allow for seamless conversation management, even in high-demand environments such as cafes, where customers often place unique and varied orders. LLaMA's ability to process a wider range of inputs and adapt to flexible language patterns makes it a suitable choice for this project.

3. Methodology

This study employed a systematic approach to design, develop, and fine-tune a voice-activated cafe ordering kiosk using the LLaMA language model.

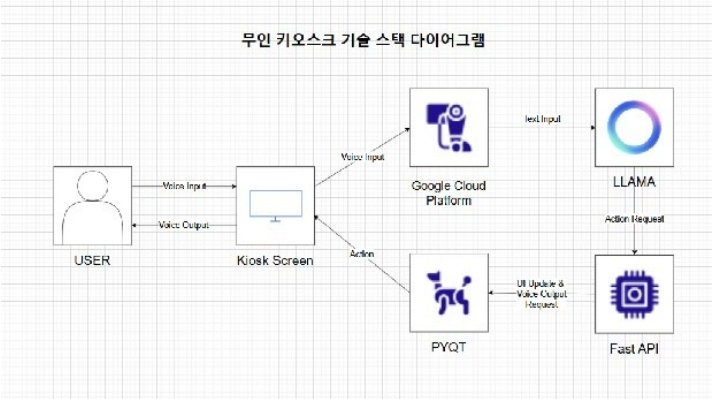


Figure 1: System Architecture of the Voice Recognition Kiosk

3.1 System Design

The kiosk system architecture was designed to streamline user interaction, with LLaMA as the core language model. Figure 1 shows the system architecture, highlighting the integration of key components such as a microphone for voice input, LLaMA for natural language processing, and Google Cloud APIs for speech-to-text (STT) and text-to-speech (TTS). The user interface was built using PYQT, providing a seamless and intuitive experience for users.

3.2 Data Preparation and Model Training

The LLaMA model was fine-tuned using a structured dataset and advanced training tools. The process involved:

- **Dataset Creation**: A comprehensive dataset of approximately 4,500 unique dialogue scenarios was generated. These scenarios included both simple orders and complex modifications, ensuring the model could handle varied user inputs effectively. The dataset includes structured inputs with natural language responses for diverse cafe ordering scenarios.

- **Data Formatting and Processing**: The dataset was methodically structured to highlight key features of each order, such as drink type, size, temperature, and specific preferences. Custom scripts were utilized to generate synthetic data scenarios, aiding in the creation of realistic dialogue exchanges.

- **Fine-Tuning with Unsloth**: The LLaMA model was fine-tuned using the Unsloth framework, leveraging tools like Google Colab, Python, and Hugging Face for efficient training and deployment. The model was optimized in a 4-bit quantized format, reducing memory usage while maintaining high processing speed and accuracy.

3.3 Testing Procedures

Extensive testing was conducted to evaluate the model's real-world performance, focusing on three key areas:

- **Accuracy of Command Interpretation**: The model's ability to correctly interpret various commands was assessed, including complex modifications and layered requests.
- **Response Time**: The speed of real-time interactions was measured to ensure the model's suitability for a fast-paced cafe environment.
- **Error Rate Analysis**: The frequency of incorrect responses was tracked, particularly in scenarios involving multiple modifications or ambiguous user inputs.

The results demonstrated that LLaMA outperformed the previous RASA model across all metrics, showcasing enhanced adaptability and improved handling of dynamic user interactions.

4. Results

The evaluation of RASA and LLaMA models was conducted using key performance metrics crucial for real-time conversational AI in a kiosk setting. The analysis focused on metrics such as accuracy, precision, recall, and F1 score, providing a comprehensive assessment of each model's strengths and limitations.

4.1 Performance Metrics and Findings

The primary metrics used include:

"Test performance metrics description"	
Accuracy	The ratio of correct predictions made by the model to the total number of predictions.
Precision	The ratio of true positive results out of all instances predicted as positive by the model.
Recall	The ratio of true positive results out of all actual positive instances.
F1 Score	The harmonic mean of precision and recall (effective for evaluating models on imbalanced data).

Figure 2: Describes the metrices used for testing performance of both the models

Figure 2 describes the performance metrics used for both models. Table 2 reveals LLaMA's advantage in handling diverse inputs and complex modifications, with significantly higher combined metrics compared to RASA's combined performance as seen in Table 1. These findings underscore LLaMA's robustness in real-world interactions, confirming its suitability for dynamic customer-facing applications.

RASA	TP	TN	FP	FN	Total	Precision	Recall	Accuracy	F1 Score
Order	25	2	1	7	35	0.962	0.781	0.771	0.862
Change	16	1	5	3	25	0.762	0.842	0.68	0.8
Remove	14	1	2	3	20	0.875	0.824	0.75	0.848
Option Add	10	1	4	5	20	0.714	0.667	0.55	0.69
Combined	65	5	12	18	100	0.844	0.783	0.7	0.812

Table 1: Illustrates the performance metrices of RASA model

LLAMA	TP	TN	FP	FN	Total	Precision	Recall	Accuracy	F1 Score
Order	6	0	0	4	10	1.0	0.60	0.60	0.75
Change	27	0	1	1	29	0.93	0.93	0.87	0.93
Remove	1	17	2	1	21	0.33	0.50	0.85	0.40
Option Add	40	0	0	0	40	1.0	1.0	1.0	1.0
Combined	74	17	3	6	100	0.94	0.91	0.89	0.93

Table 2: Illustrates the performance metrices of LLAMA model

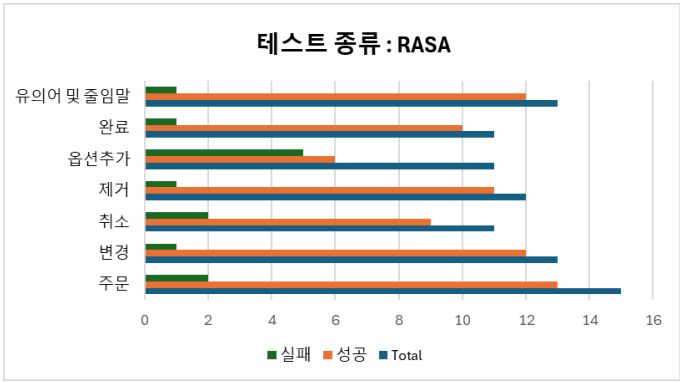
4.2 Comparative Analysis of RASA and LLaMA

The evaluation highlights LLaMA's superior handling of diverse and complex inputs compared to RASA. Key findings include:

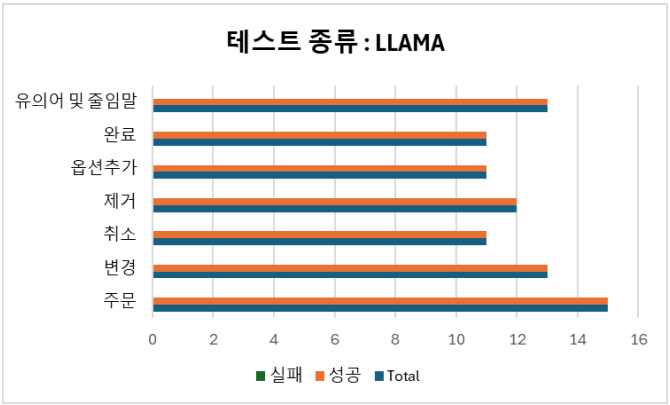
- Flexibility: LLaMA demonstrated robust adaptability in understanding varied user commands, whereas RASA's intent-based approach struggled with unexpected phrasing and complex inputs.
- Accuracy in Complex Orders: LLaMA effectively managed layered modifications and ambiguous requests, maintaining a higher success rate across test scenarios. In contrast, RASA showed increased failure rates, particularly in handling options and cancellations.
- Enhanced User Experience: LLaMA's consistent performance led to a smoother user interaction, reducing errors and response delays.

Figure 3 presents the test scenarios for both models. Graph 1 demonstrates RASA's struggle with a high rate of failures in complex scenarios and Graph 2 shows LLaMA achieving near-perfect results across all test cases, showcasing its capability for nuanced language comprehension.

Figure 3: **Test Scenarios for RASA and LLaMA Model Evaluation**



Graph 1: Rasa Model Test Results Across Different Scenarios



Graph 2: LLAMA Model Test Results Across Different Scenarios

In above graphs, the performance of LLaMA was benchmarked against RASA's previous results, revealing significant improvements in handling complex and nuanced commands, making LLaMA a more reliable choice for the kiosk application.

5. Discussion

The evaluation of LLaMA and RASA models for the voice-activated cafe ordering kiosk reveals significant improvements with LLaMA, aligning with the study's goals.

5.1 Analysis of Model Performance

Figure 2 outlines the metrics used in testing models. Results show that LLaMA consistently outperformed RASA, especially in handling complex order requests. As depicted in Table 1, RASA struggled with precision and recall, often failing with ambiguous inputs. In contrast, Table 2 highlights LLaMA's superior performance, demonstrating its robustness and adaptability in diverse scenarios.

5.2 Comparative Insights and Limitations of RASA

As illustrated in Table and Graph 2 LLaMA demonstrated lower training loss and enhanced contextual understanding, effectively addressing dynamic challenges. In contrast, as shown in Figure 4, Table and Graph 1, RASA's lacks effectiveness in handling flexible and nuanced natural language inputs. Its intent-based framework often leads to higher error rates when processing complex or nuanced customer requests, limiting its adaptability in real-world applications.

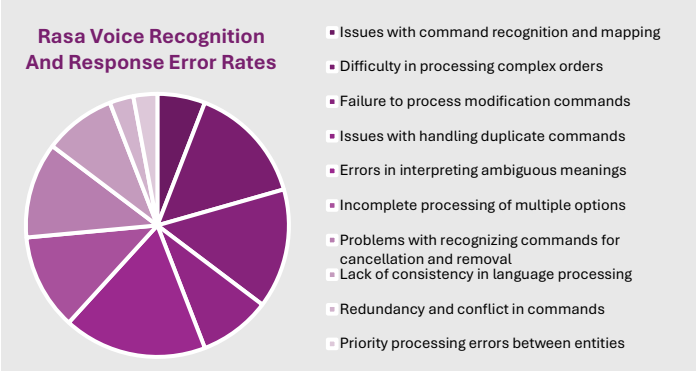


Figure 4: Illustrates The Error Analysis of RASA Voice Recognition and Response

5.3 Implications for Future AI Applications

The implementation of the LLaMA model, as detailed in Figure 1, demonstrates a practical, scalable approach for real-time conversational AI in customer-facing environments. The use of advanced training tools like Python, Colab, and Hugging Face streamlined the model development process, allowing efficient fine-tuning and rapid deployment. The results in the above Table and Graph 2 suggest that LLaMA can enhance user experience in high-demand settings by reducing error rates and improving response accuracy. This positions LLaMA as a more suitable choice for AI-driven applications that require dynamic and nuanced language understanding.

6. Conclusion

This study indicates that, while RASA provides a modular approach suitable for straightforward interactions, LLaMA excels in processing complex and diverse language inputs, making it the preferred model for voice-based conversational AI applications. The findings underscore the potential of large language models in advancing conversational AI for enhanced accessibility and user experience.

Further research will focus on:

- Optimizing LLaMA's response time through additional model compression and quantization techniques.
- Conducting real-world testing to gather user feedback and refine model training.

Acknowledgement

This research was supported by the Korean MSIT (Ministry of Science and ICT), under the National Program for Excellence in SW)(2021-0-01082) supervised by the IITP(Institute of Information & communications Technology Planning & Evaluation)"(2015-0-00912).

References

[1] Bocklisch, Tom, et al. "Rasa: Open source language understanding and dialogue management." *arXiv preprint arXiv:1712.05181* (2017).

[2] Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models."