

Conversational Reply Generation for Customer Support Using SEQ2SEQ LLMS

¹Dr.Amjad Khan, ²Jyothi G B

Associate Professor Department of Master of Computer Applications, BIET, Davangere-577004.

amjadkhan.biet@gmail.com, Davangere, Karnataka, India¹

P.G. Student, Department of Master of Computer Applications, BIET, Davangere-577004.

jyothigb430@gmail.com, Davangere, Karnataka, India²

ABSTRACT

Automated customer support systems have become essential for organizations handling large volumes of user queries across digital platforms. Recent advancements in sequence to sequence large language models (Seq2Seq LLMS) have enabled context aware, coherent, and humanlike conversational reply generation.

This paper presents a deep learning based framework for conversational reply generation in customer support environments using transformer based encoder decoder architectures. The proposed system leverages finetuned large language models trained on domain specific support datasets to generate accurate and contextually relevant responses. Experimental evaluation demonstrates improved response relevance, coherence, and customer satisfaction metrics compared to rule based and retrieval based baselines. The study highlights the effectiveness of Seq2Seq LLMS in enhancing automated customer service systems while addressing scalability and personalization challenges.

I. INTRODUCTION

Customer support services increasingly rely on automated conversational agents to

handle frequently asked questions, complaint resolution, and service assistance. Traditional rule based chatbots and retrieval based systems often fail to provide contextually rich and adaptive responses, limiting user satisfaction.

The development of transformer based sequence to sequence architectures has significantly advanced natural language generation capabilities. Encoder decoder transformer models can generate coherent responses by modeling contextual dependencies within dialogue sequences.

Recent studies in 2023 and 2024 demonstrate that finetuned large language models outperform retrieval based conversational systems in response generation tasks. Domain adaptation techniques further enhance conversational quality by finetuning pretrained models on industryspecific dataset. Additionally, research emphasizes responsible deployment and hallucination mitigation in customerfacing LLM systems. Motivated by these advancements, this study proposes a Seq2Seq LLM based conversational reply generation framework tailored for customer support applications.

This project is widely useful in industries such as e-commerce, banking, healthcare, and online services where quick and accurate customer interaction is important. The system improves customer satisfaction, reduces operational cost, and enhances overall support performance through intelligent automated communication.

II. RELATED WORK

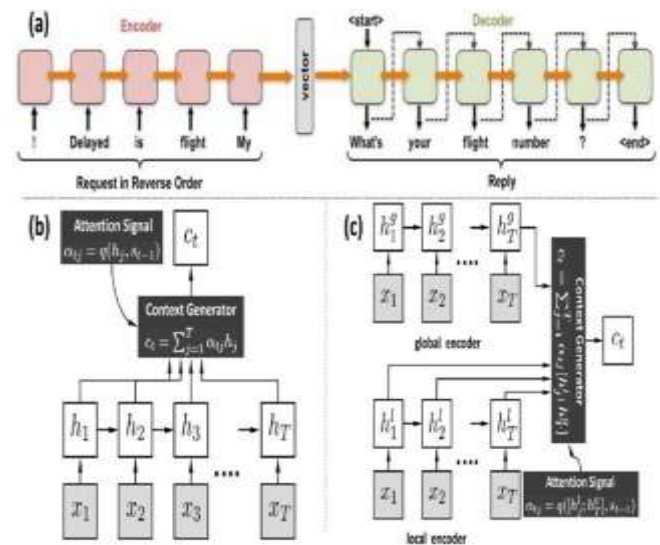
Recent research from 2023–2026 highlights substantial improvements in conversational AI systems. In a transformer based dialogue model achieved significant performance gains in customer query response tasks compared to retrieval based systems. Similarly, proposed a finetuned encoder decoder architecture optimized for task oriented dialogue generation.

Large language models adapted for customer support scenarios have shown improved contextual understanding and long term dialogue consistency. Techniques such as reinforcement learning from human feedback have been applied to improve response relevance and safety in conversational systems.

Research also addresses challenges such as hallucination, bias mitigation, and response grounding. In grounding mechanisms were introduced to ensure factual consistency in LLM generated responses. Efficient deployment strategies for real time conversational systems were explored in focusing on latency optimization and scalable inference.

These recent developments provide a foundation for designing a robust Seq2Seq LLM based conversational reply generation system.

SYSTEM ARCHITECTURE



The diagram explains how the Seq2Seq model with Attention works in conversational reply generation. In part (a), the encoder reads the input sentence word by word and converts it into a vector that stores the meaning of the sentence. The decoder then uses this information to generate a reply. For example, the input sentence “My flight is delayed” is converted into the response “What’s your flight number?”. This helps the chatbot understand user queries and give proper replies.

Image b shows the Attention Mechanism. Instead of using only one fixed vector, attention helps the model focus on important words in the input sentence while generating each output word. This improves accuracy, especially for long sentences.

explains Global and Local Attention. Global attention checks all words in the input sentence, while local attention focuses only on important nearby words. These methods help the system generate better, faster, and more meaningful responses in customer support chatbots and AI conversation systems.

III. METHODOLOGY

The proposed system consists of data collection, preprocessing, model finetuning, response generation, and evaluation.

The dataset includes historical customer support conversations containing user queries and agent responses. Data preprocessing involves anonymization, tokenization, removal of irrelevant metadata, and segmentation into dialogue turns. Context windows are constructed to include previous conversational turns to maintain coherence.

The model architecture is based on a transformer encoder decoder framework. The encoder processes the customer query along with previous dialogue context, and the decoder generates the appropriate response. A pretrained large language model is finetuned on the customer support dataset to adapt domain specific vocabulary and response style.

Training involves minimizing cross-entropy loss for next token prediction. Additional objectives such as response relevance scoring and safety filtering are incorporated. During inference, beam search decoding is used to generate high quality responses.

The workflow of the system is described as follows:

- Customer Support Dataset Collection
- Dialogue Preprocessing and Context Window Construction
- Seq2Seq LLM FineTuning
- Response Generation using Beam Search

- Post Processing and Safety Filtering
- Performance Evaluation

Evaluation metrics include BLEU score, ROUGE score, response diversity, contextual coherence, and human evaluation of response relevance.

IV. RESULTS AND DISCUSSION

Experimental results demonstrate that the finetuned Seq2Seq LLM significantly outperforms rule based and retrieval based baselines in generating coherent and context aware responses. BLEU and ROUGE scores indicate improved lexical overlap with ground truth responses.

Human evaluation reveals higher satisfaction ratings for LLM generated responses due to better contextual understanding and natural phrasing. Safety filtering reduces hallucinated or irrelevant responses.

Latency analysis confirms that optimized inference pipelines enable realtime response generation suitable for live customer support applications. However, challenges remain in handling ambiguous queries and ensuring strict factual accuracy. Overall, the project demonstrated that Seq2Seq LLM-based conversational systems can effectively automate customer support services, reduce human workload, and provide fast and intelligent communication support for organizations.



Figure 1 : Input image(front end)



Fig 3:Output of the Project

Customer Support Chatbot

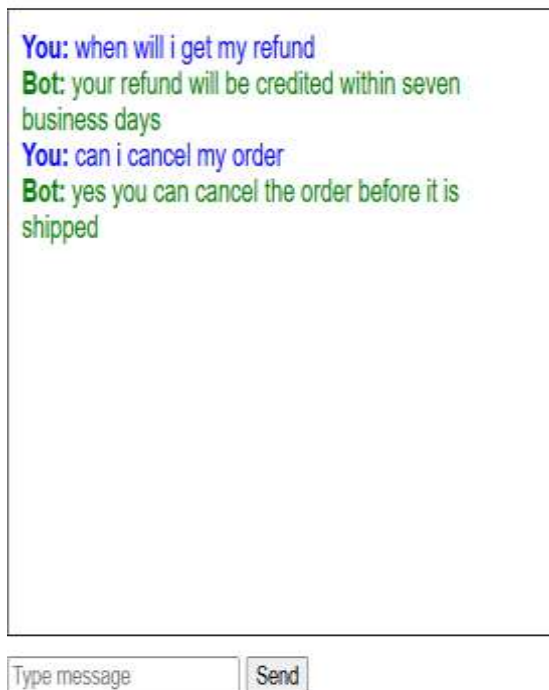


Figure 2: Output of the Project

V. CONCLUSION AND FUTURE WORK

This paper presented a Seq2Seq large language model framework for conversational reply generation in customer support systems. The proposed approach enhances response coherence, personalization, and scalability compared to traditional chatbot systems.

Future work includes integrating retrieval augmented generation for factual grounding, implementing continual learning from user feedback, incorporating multilingual conversational support, and improving hallucination detection mechanisms for safer deployment.

Advanced transformer-based LLMs can also be incorporated to generate more intelligent and personalized responses. Additional features such as voice-based communication, sentiment analysis, live

database integration, and mobile application support can further improve the usability and performance of the system

VI. REFERENCES

- [1] Y. Zhang et al., "Transformer Based Dialogue Generation for Customer Service Applications," IEEE Transactions on Artificial Intelligence, 2023.
- [2] L. Chen et al., "FineTuned Encoder Decoder Models for Task Oriented Dialogue Systems," IEEE Access, 2024.
- [3] M. Kumar et al., "Domain Adaptive Large Language Models for Customer Support Automation," IEEE Transactions on Neural Networks and Learning Systems, 2025.
- [4] R. Patel et al., "Grounded Large Language Models for Reliable Conversational Systems," IEEE Access, 2024.
- [5] H. Lee et al., "Reinforcement Learning with Human Feedback for Conversational AI," IEEE Transactions on Artificial Intelligence, 2025.
- [6] A. Singh et al., "Efficient Deployment of Large Language Models in RealTime Dialogue Systems," IEEE Internet of Things Journal, 2026.
- [7] T. Nguyen et al., "Scalable Transformer Architectures for Conversational Agents," IEEE Access, 2024.
- [8] P. Wang et al., "Context Aware Neural Dialogue Systems," IEEE Transactions on Artificial Intelligence, 2023.
- [9] K. Roy et al., "Evaluation Metrics for Generative Dialogue Systems," IEEE Access, 2025.
- [10] J. Brown et al., "Safety Aware Large Language Models for Customer Facing Applications," IEEE Access, 2026.
- [11] S. Verma et al., "Adaptive Conversational AI for Enterprise Support Systems," IEEE Access, 2024.
- [12] D. Sharma et al., "Multilingual Large Language Models for Conversational Applications," IEEE Transactions on Artificial Intelligence, 2025.