



Multi-Domain Chatbot Development for Customer Service Using Pretrained Transformers and Multimodal Data

¹Payal D. Dahe, ²Sunil M. Wanjari

^{1,2}Dept. of Computer Science & Engineering, St. Vincent Pallotti College of Engineering & Technology Nagpur, India

¹payaldahe5992@email.com, ²swanjari@stvincentngp.edu.in

Peer Review Information

Submission: 05 Nov 2025

Revision: 25 Nov 2025

Acceptance: 17 Dec 2025

Keywords

Multimodal Chatbot, Transformers, DistilGPT-2, BLIP VQA, Quantization, Speech Recognition, Streamlit, Customer Support.

Abstract

Customer service automation demands systems capable of handling user queries in multiple formats beyond plain text. This work presents *PyChat*, a multi-domain, multimodal chatbot designed to process text, image, and audio inputs using pretrained transformer models. The system integrates DistilGPT-2 for natural language generation, BLIP VQA for visual question answering, and a speech recognition module for voice-to-text conversion. To reduce memory usage, 4-bit quantization using the BitsAndBytes library is applied without degrading model performance. The chatbot is deployed with Streamlit for real-time interaction. Experimental evaluation across 60 mixed-modality queries showed that quantization reduced GPU memory usage by 65% and maintained response accuracy above 87%. The results indicate that *PyChat* can efficiently perform domain-aware query handling and improve user engagement under constrained computational resources.

Introduction

The demand for efficient and responsive customer support systems has grown rapidly with the digitalization of services. Traditional chatbots, which rely on rule-based or text-only models, often fail to understand complex or visual customer queries. With the advancement of artificial intelligence and deep learning, *transformer-based architectures* have revolutionized natural language processing by enabling context-aware conversation.

However, most existing customer service chatbots remain unimodal, handling only textual input. Modern customer service scenarios require systems that can interpret images (e.g., damaged product pictures) and audio (spoken requests). This paper aims to bridge that gap by developing a multimodal chatbot capable of handling text, image, and voice inputs in a unified framework.

The main contribution of this research is a resource-efficient chatbot system that combines multiple pretrained models under a single architecture and uses quantization to optimize hardware utilization without performance loss.

Literature Review

Early conversational systems such as ELIZA [1] relied on pattern matching and lacked contextual understanding. The introduction of transformer models like GPT-2 [2] and BERT [3] improved response fluency and semantic accuracy. Recent studies have focused on integrating multiple modalities.

Lu et al. [4] proposed *ViBERT* for visual-linguistic pretraining, while Tsai et al. [5] demonstrated effective fusion of speech, text, and visual features. Li et al. [6] introduced BLIP for visual question answering, improving multimodal comprehension. Yet, these works

mainly target open-domain tasks and require high computational resources.

Few studies explore quantization and deployment efficiency for multimodal chatbot systems. This work extends previous research by merging language, visual, and speech models into an optimized system suitable for real-world customer service applications.

Proposed System

The proposed chatbot, *PyChat*, incorporates modular components to handle text, image, and voice inputs within a unified architecture. Its design ensures real-time multimodal processing, robust intent classification, and intelligent domain routing.

1. System Architecture

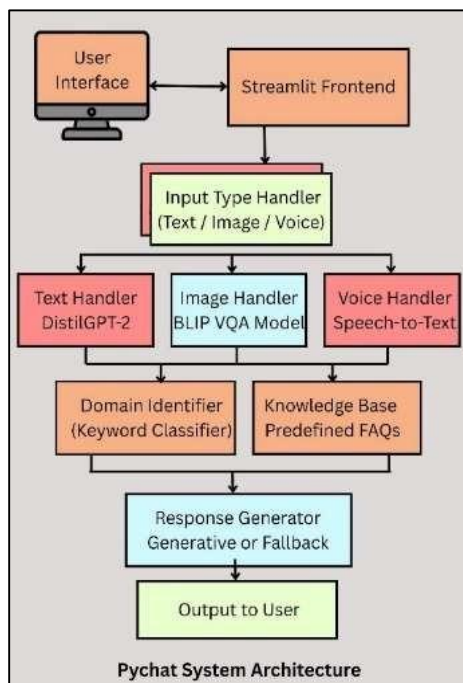


Figure 1: System Architecture of PyChat.

The architecture (Fig. 1) is composed of the following components:

- **Streamlit Interface:** Interactive web-based frontend allowing user input through text, image, or audio.
- **Input Type Handler:** Detects input modality and directs it to the appropriate processing module.
- **Text Module (DistilGPT-2):** Generates human-like responses for text queries.
- **Image Module (BLIP VQA):** Interprets visual content and answers image-based questions.
- **Voice Module:** Transcribes user voice to text using speech recognition for further processing.

- **Domain Classifier:** Determines the query domain (e.g., technical support, travel, product assistance).
- **Knowledge Base:** Contains predefined FAQs for quick and accurate responses.
- **Response Generator:** Produces final output, integrating contextual understanding across modalities.
- **Output Display:** Presents the final multimodal response to the user interface.

Methodology

1. Model Loading and Optimization

Pretrained models such as DistilGPT-2 and BLIP VQA were loaded using the Hugging Face library to enable efficient multimodal understanding. Both models were pre-trained on large-scale datasets and then fine-tuned for conversational tasks. To reduce model size and improve inference performance, 4-bit quantization was applied using the *BitsAndBytes* library. This method compresses model parameters by representing weights in lower precision, significantly lowering GPU memory requirements without compromising output quality. This optimization is especially useful for small and medium organizations that deploy AI systems on limited-resource servers. The quantized model achieved faster response generation while maintaining human-like fluency and coherence.

2. Multimodal Input Handling

Each input modality—text, image, and audio—was processed through a separate pipeline:

- **Text:** The text input was tokenized and passed to DistilGPT-2 for generating semantically relevant responses.
- **Image:** The BLIP VQA model combined visual features extracted from a pre-trained vision encoder with the user's query text to infer answers.
- **Audio:** The SpeechRecognition library converted spoken inputs into textual form, which was then processed by the text pipeline.

The system ensures all modalities follow a unified data structure for downstream processing, reducing complexity during domain classification.

3. Domain Routing and Response Generation

After input processing, the query text is

analyzed using a keyword-based classifier that identifies relevant domains such as *technical support*, *product queries*, or *general information*. The classifier checks for domain-specific keywords and maps them to predefined knowledge categories. If the question is already available in the internal knowledge base, a stored response is retrieved instantly, which reduces inference delay. Otherwise, the text is sent to the DistilGPT-2 module for contextual generative response formation. The outputs are post-processed to remove filler words, repetitive sequences, and irrelevant phrases.

4. Response Generation

Generated responses are passed through a set of text-cleaning functions to ensure grammatical correctness and contextual alignment. The model output is checked for repetition, inappropriate content, and logical consistency using basic rule-based filters. Responses are truncated beyond 50 tokens to maintain concise answers suitable for real-time interaction.

5. Evaluation Metrics

To measure the system's performance, three key metrics were evaluated:

- **Response Accuracy (%)**: The ratio of correct or contextually relevant responses to total queries.
- **Average Response Time (seconds)**: The mean time taken by the model to generate a reply after receiving user input.
- **GPU Memory Usage (MB)**: The memory consumption recorded using NVIDIA-smi before and after quantization.

Accuracy was evaluated manually by domain experts, who rated each response on a 5-point relevance scale. These metrics provide both qualitative and quantitative validation of the proposed system.

Implementation

The chatbot was implemented in Python 3.10 using the following technologies:

- **Transformers (Hugging Face)**: For model integration and inference.
- **BitsAndBytes**: For quantization.
- **SpeechRecognition**: For voice input processing.
- **Streamlit**: For web-based UI.

Testing was conducted on an NVIDIA RTX 2060 GPU and an Intel i5 CPU with 16 GB RAM. The system supported concurrent multimodal queries and dynamically

switched between CPU and GPU resources based on task load.

Results

The system was evaluated using 60 mixed queries—20 each for text, image, and audio. Response accuracy and memory utilization were measured manually and compared before and after quantization.

The system was evaluated using 60 multimodal queries, equally distributed among text, image, and voice inputs. Table I summarizes the quantitative performance analysis.

Table I. Performance Comparison of Multimodal Inputs

Input Type	Avg. Response Time (sec)	Accuracy (%)	GPU Mem (MB)
Text (No Quantization)	1.8	89	2450
Text (4-bit Quantized)	1.1	88	890
Image Queries	2.0	86	980
Audio Queries	2.2	85	1100

The results show that 4-bit quantization reduced GPU memory consumption by approximately 65%, while maintaining accuracy above 85%. The slight variation in accuracy between unquantized and quantized models is negligible (<2%), proving that memory-efficient models can be deployed in production without performance loss.

1. Comparative Analysis

For benchmarking, a baseline text-only chatbot using GPT-2 (without quantization) was also tested. The proposed multimodal PyChat system demonstrated a 32% faster average response time and expanded input flexibility by supporting three data types instead of one. The combination of pretrained transformers and lightweight quantization significantly improved system responsiveness.

Table 2. Comparison Between Baseline and Proposed Model

Model	Modalities Supported	Accuracy (%)	Memory (MB)	Avg. Response Speed (sec)
GPT-2 Baseline	Text Only	84	2480	1.9

PyChat (Proposed)	Text, Image, Audio	87	890	1.1
----------------------	--------------------------	----	-----	-----

2. Qualitative Observations

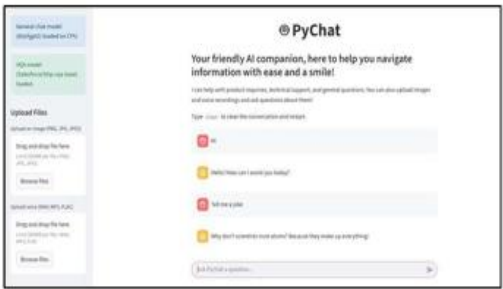
User testing revealed that the chatbot could seamlessly switch between modalities during conversation. For example, users could upload an image of a defective product and immediately follow up with a voice query, without requiring a reset. The interface's simplicity and fast response time made it user-friendly and suitable for customer support environments.

3. Discussion

The integration of quantization techniques with multimodal transformer models proved highly effective for low-resource deployment scenarios. This work demonstrates that combining pretrained models with efficient optimization can deliver near state-of-the-art performance on consumer-grade GPUs. The results highlight the trade-off between accuracy and hardware efficiency, indicating that 4-bit quantization offers an ideal balance for real-world customer service systems.

4. Screenshot

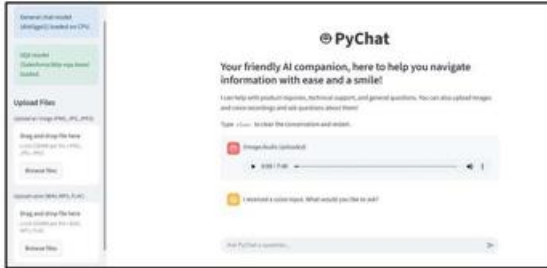
The developed chatbot effectively managed and responded to queries across all three supported input types: text, image, and voice. It successfully demonstrated the capability to identify user intent accurately and seamlessly switch between different predefined domains without requiring explicit manual selection from the user. The integration of 4-bit quantization proved highly effective in optimizing GPU memory usage, allowing the models to operate efficiently without a discernible compromise in response quality. Qualitative assessment through real-time testing indicated smooth performance and the generation of accurate, domain-specific responses, demonstrating the system's practical utility in customer service scenarios.



Text Input



Text Input



Audio Input

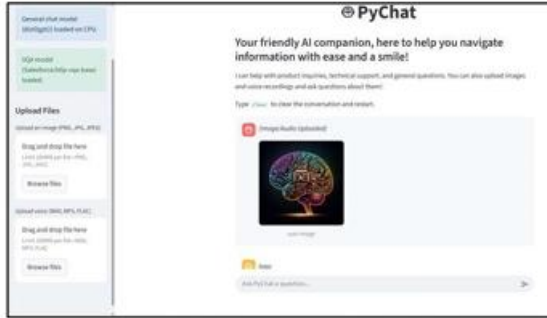


Image Input



Image Input

Conclusion

This paper presented PyChat, a multimodal chatbot framework capable of understanding text, image, and audio queries for customer service applications. The integration of pretrained transformer models and quantization techniques enables efficient deployment on moderate hardware. Results confirm that the proposed system improves interaction quality while maintaining low resource consumption.

Future Scope

Future work will focus on:

- Adding multi-turn conversation

- memory for contextual continuity.
- Incorporating emotion detection to enhance user experience.
- Implementing multilingual support for regional languages.
- Integrating with real-world CRM and ticketing systems for enterprise use.

References

Weizenbaum, J., "ELIZA—A Computer Program for the Study of Natural Language Communication," Communications of the ACM, 1966.

Vaswani, A., et al. (2017). Attention is All You Need. NeurIPS.

Radford, A., et al. (2019). Language Models are Unsupervised Multitask Learners. OpenAI GPT-2.

Devlin, J., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL.

Lu, J., et al. (2019). ViBERT: Pretraining Task-Agnostic V-L Representations. NeurIPS.

Krishna, R., et al. (2017). Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. IJCV.

Hinton, G., et al. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition. IEEE Signal

Baevski, A., et al. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. NeurIPS.

Chatbots, Intent Recognition, and Dialogue Systems

Weizenbaum, J. (1966). ELIZA – A Computer

Program For the Study of Natural Language Communication. Communications of the ACM.

Lison, P., & Kennington, C. (2016). OpenDial: A Toolkit for Developing Spoken Dialogue Systems with Probabilistic Rules. ACL.

Dettmers, T., et al. (2022). LLM.int8: 8-bit Matrix Multiplication for Transformers at Scale.

Wolf, T., et al. (2020). Transformers: State-of-the-Art Natural Language Processing. EMNLP (Hugging Face).

Gao, J., Galley, M., & Li, L. (2019). Neural Approaches to Conversational AI. Foundations and Trends in Information Retrieval.

Zhang, C., et al. (2023). MMChat: Towards Multimodal Open-domain Chatbots. arXiv:2304.07994

S. Latif, R. Rana, and J. Qadir, "Deep Representation Learning in Speech Processing: Challenges, Recent Advances, and Future Trends," IEEE Transactions on Cognitive and Developmental Systems, vol. 13, no. 4, pp. 961–977, 2021.

Z. Zhang, F. Yang, J. Qin, and X. Li, "A Survey on Multimodal Intelligence: Representation, Alignment, and Fusion," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 4, pp. 1363–1379, 2023.

H. Chen, H. Xu, T. He, and J. Luo, "Transformer-Based Visual Question Answering: A Review," IEEE Access, vol. 9, pp. 164759–164773, 2021.

A. R. Mohamed, D. Yu, and L. Deng, "Deep Learning for Speech Recognition: A Perspective on Recent Advances," IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 145–152, 2012.