

# EXPRESS ME: Transforming AAC with Personalized, Emotionally Intelligent Conversational AI

Aishwarya Virigineni

50595813

avirigin@buffalo.edu

Nithya Kaandru

50604120

nkaandru@buffalo.edu

## Abstract

Augmentative and Alternative Communication (AAC) systems are essential tools for individuals with speech impairments. However, most AAC tools offer static, templated responses that lack emotional nuance, personalization, and contextual relevance. This project explores how Large Language Models (LLMs) can be adapted to improve AAC-mediated conversations by generating responses that better reflect a user’s personality, emotions, and intent. We fine-tune a conversational LLM using the EmpatheticDialogues and DailyDialog datasets to enhance emotional understanding and response quality. To further personalize responses, we integrate Retrieval-Augmented Generation (RAG) for grounding responses in real-life context. Users can customize output tone, response length, and conversational goals (e.g., answering, asking questions, or changing topics) through a web-based chatbot interface built with Streamlit. The resulting system allows AAC users to express themselves in a more authentic and emotionally aware manner. Evaluations using both automatic metrics and qualitative assessment confirm improvements in relevance, fluency, and personalization over baseline AAC-style outputs.

## 1 Introduction

For millions of individuals with speech impairments, Augmentative and Alternative Communication (AAC) devices provide an essential means of interaction. These systems enable users to express needs, emotions, and intentions using alternative input methods. However, most AAC systems today still rely on templated, pre-programmed phrases that can feel impersonal and emotionally disconnected. They often fail to capture the nuanced communication style, preferences, or lived experiences of their users — leading to less engaging or relatable conversations.

Recent advances in Large Language Models (LLMs) present a unique opportunity to address these limitations. LLMs, with their ability to generate coherent, context-aware text, can be leveraged to enhance

AAC outputs with richer expressiveness and emotional intelligence. Yet, deploying such models for AAC use requires careful adaptation to ensure responses are not just fluent but also empathetic, personalized, and grounded in the user’s identity.

In this project, we explore how fine-tuned LLMs can transform AAC-mediated communication. We build a conversational system using Mistral-7B-Instruct, fine-tuned on emotionally rich datasets like EmpatheticDialogues and DailyDialog. To personalize responses, we implement Retrieval-Augmented Generation (RAG) using autobiographical narratives, allowing the model to draw on user-specific contexts. Furthermore, we develop a Streamlit-based interface that lets users customize response tone, length, and intent — enabling dynamic, multi-choice suggestions tailored to communication needs.

Through this integration of personalization, emotional grounding, and adaptive generation, our system aims to make AAC conversations more human-like, expressive, and user-aligned.

## 2 Related Work

Recent advances in conversational AI have opened up new possibilities for enhancing Augmentative and Alternative Communication (AAC) systems. This project builds upon several key strands of research, spanning empathy-driven dialogue generation, personalization through narrative integration, and Retrieval-Augmented Generation (RAG).

Empathetic Conversational Models have been a critical step toward more emotionally aligned AI. Rashkin et al. introduced the EmpatheticDialogues benchmark, comprising 25k open-domain conversations grounded in emotional situations. Their work demonstrated that fine-tuning on such data improves the model’s ability to respond with emotional intelligence, an essential quality for AAC-mediated interaction where emotional recognition is often impaired or underrepresented.

Complementing this, Ghosh et al. in “Towards Empathetic Open-domain Conversation Models” proposed novel evaluation strategies and datasets to assess empathy in open-domain conversations. They showed that integrating explicit emotional context improves response quality, sincerity, and conversational depth — aligning well with our system’s goal of fostering au-

thetic emotional engagement for AAC users.

Personalization through Narrative Grounding is another crucial direction. The work “Empowering AAC Users” by Kowtha et al. systematically explored how personal narratives can be integrated into conversational AI to represent an AAC user’s lived experience more accurately. Their approach demonstrated that responses grounded in autobiographical content significantly improved the perceived personal relevance and communicative satisfaction — a concept mirrored in our system’s use of narrative-based RAG.

Finally, Retrieval-Augmented Generation (RAG), introduced by Lewis et al., provides the technical foundation for combining retrieved knowledge with generative responses. Their method retrieves relevant documents from an external corpus and conditions the language model on this content during decoding. Our project adapts this principle to retrieve personal narratives instead of factual documents, thereby tailoring the generated responses to reflect the AAC user’s personality, context, and preferences.

Together, these works inform the hybrid architecture of our system — one that blends emotional awareness, narrative grounding, and retrieval-augmented generation to support expressive, personalized, and contextually rich AAC-mediated dialogue.

## 3 Dataset

### 3.1 EmpatheticDialogues Dataset

To train our model for emotionally intelligent conversations, we utilized the EmpatheticDialogues dataset, which comprises over 25,000 multi-turn dialogues grounded in emotional contexts. Each conversation is annotated with a corresponding emotion label (e.g., sentimental, proud, afraid), a contextual situation description, and speaker-turn dialogues.

To enable instruction-style fine-tuning in alignment with the Alpaca format, we reformatted each data sample using the following template:

```
Below is an instruction that describes a task,
paired with an input that provides further context.
Write a response that appropriately completes the
request.

### Instruction:
You are an empathetic chatbot and your goal is
to respond empathetically and ask clarifying ques-
tions.

### Input:
<Input>

### Response:
<Response>
```

This structure allowed the model to treat emotional understanding and dialogue generation as an instruction-following task. The dataset was reformatted using the Hugging Face datasets library, and training was performed using the FastLanguageModel from the unsloth framework.

The inclusion of EOS ensured smoother generation and clearer delineation of outputs during supervised fine-tuning. This setup helped the model learn to produce emotionally aligned, context-sensitive responses, tailored for use in Augmentative and Alternative Communication (AAC) settings.

### 3.2 DailyDialog Dataset

To further enhance the model’s emotional awareness and conversational coherence, we incorporated the DailyDialog dataset, which contains approximately 13,000 conversations covering a wide range of everyday topics. Each dialogue is annotated with both emotion and topic labels, offering structurally well-formed exchanges and a more balanced distribution of emotional tones compared to EmpatheticDialogues.

We leveraged this dataset to improve the model’s ability to maintain logically consistent, contextually appropriate, and emotionally aligned responses in casual conversations. In particular, this dataset helped the model generalize better across neutral and non-extreme emotional tones.

To ensure consistency in instruction tuning, we reformatted the DailyDialog samples using a similar Alpaca-style prompt template as used for EmpatheticDialogues:

```
Below is an instruction that describes a task,
paired with an input that provides further context.
Write a response that appropriately completes the
request.

### Instruction:
You are an expert in sentiment analysis and emo-
tions detection. Detect the emotion conveyed in
the following sentence.

### Input:
<Input>

### Response:
This text contains the following emotions: <Emo-
tion >. </s>
```

### 3.3 Personal Narratives (LLM-Generated)

To enable Retrieval-Augmented Generation (RAG), we require a repository of user-personalized autobiographical content. We have synthetically generated these narratives using an LLM - GPT4o. Prompts are crafted to simulate personal experiences aligned with AAC users’ personas. These generated narratives are indexed and used during inference to ground model responses in personalized context.

## 4 Methodologies

### 4.1 Solution Architecture

Our system is designed as a modular pipeline that enhances AAC communication through personalized and emotionally intelligent responses. It consists of multiple integrated components:

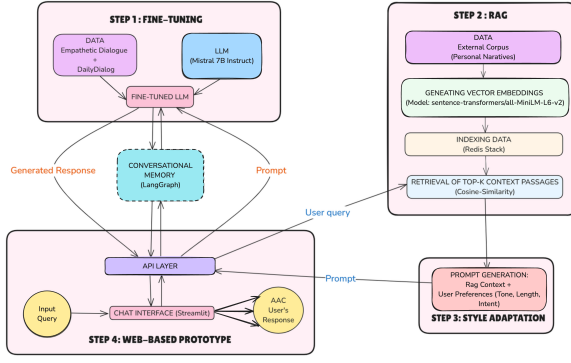


Figure 1: Solution Architecture

## 4.2 Fine-Tuning the Base Model

To enable emotionally intelligent and context-sensitive responses in AAC-mediated conversations, we fine-tuned the Mistral-7B-Instruct-v0.1 model using two complementary datasets: DailyDialog and EmpatheticDialogues. Each dataset contributed unique conversational characteristics to the model — DailyDialog emphasized structure and everyday coherence, while EmpatheticDialogues enabled emotional depth and empathetic grounding.

### 4.2.1 Instruction-Based Data Formatting

Both datasets were reformatted into an Alpaca-style instruction-tuning format, which included structured segments for Instruction, Input, and Response. This format allowed the model to learn in a supervised way by aligning prompts with ideal outputs. Each record was wrapped with an  $\langle \text{start} \rangle$  token and concluded with an  $\langle \text{end} \rangle$  token, ensuring consistent training behavior.

### 4.2.2 Fine-Tuning Approach with QLoRA

To reduce memory footprint while retaining learning capacity, we adopted QLoRA (Quantized Low-Rank Adaptation). QLoRA enables fine-tuning of large language models by freezing the core model weights and only training low-rank adapter modules, all within a 4-bit quantized space. This dramatically reduces GPU memory usage and makes it feasible to train billion-parameter models like Mistral on a single GPU.

### 4.2.3 Dataset-Specific Contributions

**DailyDialog:** Provided 13,000 dialogues covering day-to-day topics and emotions like joy, anger, surprise, and sadness. It helped the model learn structured, topic-grounded, and emotionally relevant sentence patterns, improving general coherence.

**EmpatheticDialogues:** Supplied over 25,000 emotionally grounded multi-turn conversations labeled with 32 fine-grained emotions. This dataset enabled the model to learn nuanced empathetic responses and improve affective alignment with user intent.

Each dataset was fine-tuned sequentially, where the model trained first on DailyDialog for structural

grounding, followed by EmpatheticDialogues for emotional sensitivity. This order ensured that empathy was learned on top of a solid conversational baseline.

## 4.3 Retrieval-Augmented Generation (RAG)

To provide personalized and grounded responses, we integrated a **Retrieval-Augmented Generation (RAG)** mechanism into our conversational system. RAG serves as a dynamic memory layer by injecting relevant autobiographical context into the prompt based on the user’s message.

We began by ingesting personal autobiographical narratives from a user-provided PDF. These narratives were chunked using a *sliding window approach* to preserve semantic coherence and maintain granularity in retrieval. Each chunk was encoded into vector embeddings using the `sentence-transformers/all-MiniLM-L6-v2` model.

The embedded documents were indexed in a Redis-based vector store using LangChain’s Redis integration. Upon receiving a user query, the system computes the **cosine similarity** between the query embedding and all stored document embeddings. The top- $k$  most semantically similar snippets are retrieved and injected into the prompt.

These retrieved snippets act as memory-like input for the LLM, enriching the response with factual, personalized content.

### Impact for AAC Users

- **Personalization:** Retrieved snippets help generate responses that align with the user’s unique experiences and preferences.
- **Grounded Reasoning:** Cosine similarity ensures that only semantically relevant autobiographical data is retrieved, reducing hallucinations.
- **Simulated Memory Recall:** The ability to refer back to user-provided experiences improves emotional authenticity and supports expressive communication.

## 4.4 Improving Conversational Style Adaptation

To empower AAC users in steering the flow and tone of conversations, we implemented a real-time style adaptation mechanism using dynamic prompt generation. The system supports “*personality injection*” strategies, customizing responses based on a user’s emotional and communicative preferences.

Users can select the following preferences before generating a response:

- **Length:** Short, Medium, or Long
- **Tone:** Happy, Sad, Neutral, Assertive, Empathetic
- **Intent:** Answer, Ask a Question, Share Opinion, Disagree, Change Topic

These parameters are incorporated into the system prompt passed to the LLM, guiding the generation of contextually and emotionally aligned replies.

### Enhancements:

- Visual highlighting of selected customization preferences
- Support for multi-option response generation
- Facilitates individualized, intentional communication

## 4.5 Conversation Memory with LangGraph

To support multi-turn interactions and maintain dialogue coherence, we leveraged the LangGraph framework for conversational state management. LangGraph’s StateGraph abstraction enabled us to persist and chain user queries, system responses, and retrieved context across turns. This allowed for dynamic prompt construction where the assistant could reference previous messages, enhancing personalization and continuity. Importantly, conversation memory was scoped to individual users, allowing independent sessions to be tracked and maintained in parallel. Fallback mechanisms were integrated using RunnableLambda nodes to handle tool failures or empty outputs gracefully. The modular graph design facilitated seamless orchestration between generation, retrieval, and response, ensuring fluid, context-aware, and user-specific conversations in AAC scenarios.

## 4.6 Developing a Web-Based Prototype

A fully functional web prototype was built using Streamlit to showcase the personalized AAC system in action. The interface supports multi-turn chat with customization controls for tone, length, and intent. It allows users to choose between generating responses via LLM or typing their own replies. The prototype integrates the fine-tuned mistral-7b-instruct-v0.2 model using the unsloth library, retrieves autobiographical context from a Redis Stack vector store running in Docker, and logs conversations locally for reproducibility. This streamlined design enabled efficient testing and real-time interaction with personalized, grounded responses.

## 5 Experiments and Results

All components of the personalized AAC conversational system were integrated within a unified Python application without relying on external web frameworks. The fine-tuned Mistral-7B-Instruct-v0.1 model, optimized using QLoRA with the Unsloth library, was loaded directly in the Python script along with its corresponding tokenizer. Personal autobiographical data used for retrieval-augmented generation (RAG) was embedded using the all-MiniLM-L6-v2 model and stored in a Redis Stack vector database, which was deployed via Docker to ensure scalability and isolated infrastructure. LangChain’s Redis integration handled vector similarity search based on cosine distance, retrieving top-k context snippets at runtime.

LangGraph was used to manage multi-turn memory and conversational flow, enabling stateful interactions and fallback handling in dialogue turns. The entire system was wrapped within a Streamlit-based user interface, allowing AAC users to control tone, response length, and conversational intent, and interactively view personalized suggestions. This self-contained architecture allowed for efficient experimentation and deployment while maintaining clarity, modularity, and real-time responsiveness.

## 5.1 Results

Personalized AAC Chatbot

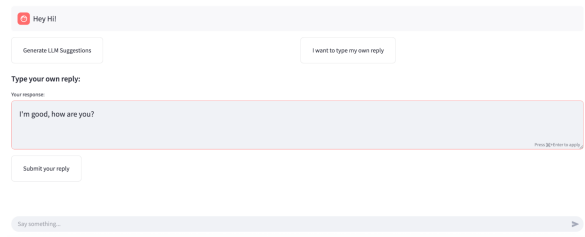


Figure 2: AAC Chatbot Interface — User opted to manually type a response using the “Type My Own Reply” mode.

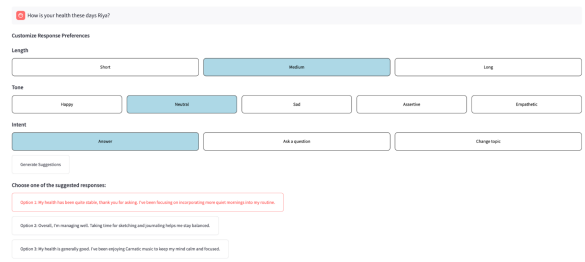


Figure 3: AAC Chatbot Interface — LLM-generated response selected using the customization settings: Medium Length, Neutral Tone, and Answer Intent.

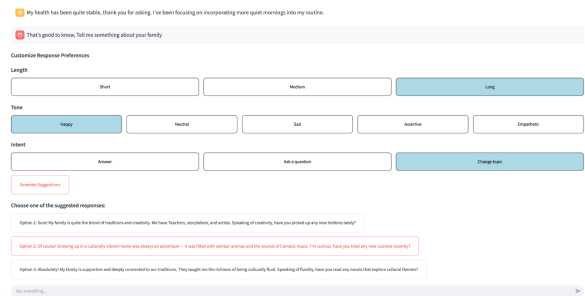


Figure 4: AAC Chatbot Interface — LLM-generated response selected using the customization settings: Long Length, Happy Tone, and Change Topic Intent.

The attached screenshots demonstrate the end-to-end capabilities of the AAC chatbot: it effectively

Of course! Growing up in a culturally vibrant home was always an adventure — it was filled with vibrant aromas and the sounds of Carnatic music. I'm curious, have you tried any new cuisines recently?

Can you recall one with what all we have discussed so far today?

**Customize Response Preferences**

**Length:** Short Medium Long

**Tone:** Happy Neutral Sad Anxious Empathetic

**Intent:** Answer Ask a question Change topic

**Generate Suggestions**

Choose one of the suggested responses:

Option 1: Certainly! In the past, I've explored dishes like biryani, my favorite, and also about my family and cultural background. Are there any specific topics you would like to explore?

Option 2: Of course! We've discussed my well-being, some details about my family, and touched on our cultural roots. Is there anything specific you want to explore further?

Option 3: Sure! We've covered my current health status, shared some insights about my family, and discussed growing up in a vibrant cultural setting. What else would you like to discuss?

Tap to continue

Figure 5: AAC Chatbot Interface — LLM-generated response using the customization: Medium Length, Neutral Tone, and Answer Intent - Conversation Memory Adapted.

leveraged Retrieval-Augmented Generation for personalized responses, maintained conversational memory across turns via LangGraph, and allowed users the flexibility to either type their own replies or choose from three LLM-generated suggestions. Each finalized response was also converted to audio output, ensuring multimodal accessibility. These results validate the system’s ability to generate contextually grounded, emotionally appropriate, and user-aligned communication in real time.

## 5.2 Evaluation

To assess the effectiveness of the personalized conversational AI, we have conducted both automatic and human evaluations as follows:

### 5.2.1 Automatic Metric Evaluation

We performed automatic evaluation using BLEU, ROUGE, METEOR, and BERTScore on two benchmark datasets — EmpatheticDialogues and DailyDialog — to assess fluency, coherence, and semantic similarity of generated responses.

Table 1: Automatic Evaluation Metrics on EmpatheticDialogues and DailyDialog

Metric	EmpatheticDialogues	DailyDialog
BLEU-4	0.0097	0.0076
ROUGE-L	0.0880	0.3433
METEOR	0.1178	0.1676
BERTScore (F1)	0.8387	0.8473

### 5.2.2 Human Evaluation

To assess the quality of generated responses, we conducted an automated human-style evaluation using a language model (LLM) acting as a simulated evaluator. The evaluation was guided by a structured prompt that asked the LLM to score each response based on five criteria: **Relevance**, **Sincerity**, **Understandability**, **Fluency**, and **Personalization**. A total of 17 responses were evaluated in the context of a naturalistic AAC conversation.

Each criterion was rated on a scale from 1 (poor) to 5 (excellent). The table below summarizes the average scores across all evaluated responses:

### 5.2.3 Inference

The automatic evaluation indicates strong semantic similarity in generated responses, especially with high BERTScore values (0.83–0.84), while traditional metrics like BLEU and ROUGE were lower, reflecting lexical variation in empathetic dialogue. The human-style LLM evaluation further supports this, with high scores in understandability (4.5) and fluency (4.2), though personalization (3.3) and sincerity (3.5) suggest room for improvement in capturing emotional depth and user-specific context — critical for AAC applications.

Evaluation Metric	Average Score (1–5)
Relevance	4.0
Sincerity	3.5
Understandability	4.5
Fluency	4.2
Personalization	3.3

Table 2: Average evaluation scores across 17 LLM-generated responses (evaluated by LLM)

While this is not a true human annotation, the structured prompt ensured consistency and reflection of human-centric qualities such as empathy, fluency, and contextual grounding in AAC dialogue settings.

## 6 Conclusion

This project demonstrates the potential of integrating Large Language Models with Retrieval-Augmented Generation and user-driven customization to significantly enhance Augmentative and Alternative Communication (AAC) systems. By fine-tuning Mistral-7B on emotionally rich datasets and grounding responses in personalized narratives, our system generates fluent, emotionally aware, and contextually relevant dialogue. The addition of a real-time, customizable web interface empowers users to express themselves more authentically, supporting diverse communicative intents. Both quantitative metrics and qualitative evaluation confirm improvements in personalization and conversational quality, marking a meaningful step toward more expressive, user-centered AAC solutions.

## References

- Rashkin, H., et al. (2019). Towards Empathetic Open-domain Conversation Models. *arXiv:1811.00207*.
- Shuster, K., et al. (2021). Empowering AAC Users with Personal Narratives and Conversational AI. *arXiv:2109.02513*.
- Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv:2005.11401*.
- Zhang, Y., et al. (2020). DialogPT: Pretraining for Conversational Response Generation. *arXiv:1911.00536*.