Cognitive Ethical Memory Architecture for Adaptive Conversational AI

Pranav T[1], S.K. Devipriya[2], Aditya Rana S.K[3], Kanipriya M[4]

[1] *Department of Computational Intelligence, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India pt4961@srmist.edu.in*

[2] *Department of Computational Intelligence, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India ds4630@srmist.edu.in*

[3] *Department of Computational Intelligence, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India as7326@srmist.edu.nin*

[4] *Department of Computational Intelligence, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India kaniprim@srmist.edu.in*

ABSTRACT

This paper presents the Cognitive-Ethical Memory Architecture for Adaptive Conversational AI (CEMA), a novel framework that integrates multi-layered memory, emotional intelligence, and ethical reasoning to enable context-aware and personalized interactions. Unlike conventional stateless chatbots, the proposed system maintains persistent emotional, temporal, and contextual memory across conversations, allowing dynamic adaptation to user behavior and preferences. The architecture comprises three core components: a multi-dimensional long-term memory module, an adaptive personality engine, and a contextual prompt builder that generates coherent and ethically aligned responses. The system is evaluated using multiple metrics, including BERTScore, ROUGE, BLEU, emotion detection accuracy, and response quality. Experimental results demonstrate strong semantic understanding (BERT Score F1: 0.873) and content relevance (ROUGE-1: 0.412), while a low BLEU score (0.073) reflects enhanced response diversity. The model achieves 60% emotion detection accuracy while maintaining contextual coherence across extended interactions. These findings highlight the effectiveness of integrating cognitive memory, emotional intelligence, and ethical considerations in improving conversational quality, enabling more engaging, trustworthy, and human-centric AI systems.

Keywords: Conversational AI, Emotional Intelligence, Memory Architecture, Personality Adaptation, QLoRA..

How to cite this article: Pranav T, S.K. Devipriya, Aditya Rana, Kanipriya M (2026). Cognitive Ethical Memory Architecture for Adaptive Conversational AI. International Journal of Drug Delivery Technology; 2026;16(49s):1241-1248. DOI: 10.25258/ijddt.16.49s.137

INTRODUCTION

Recent advancements in Conversational Artificial Intelligence (AI), driven by large language models (LLMs), have significantly enhanced the ability of systems to generate coherent and contextually relevant responses. However, most existing conversational agents remain fundamentally stateless, lacking the capacity to retain long-term context, model user preferences, or effectively incorporate emotional understanding. This limitation reduces their effectiveness in real-world applications that demand personalization, continuity, and emotionally aware interactions. While recent research has explored the integration of memory mechanisms and emotion-aware processing—particularly through approaches such as Retrieval-Augmented Generation (RAG) —these systems primarily emphasize semantic relevance and often overlook emotional prioritization, temporal continuity, and adaptive response behavior. Moreover, current methodologies seldom unify multi-layered memory, emotional intelligence, and personality adaptation within a single cohesive framework.

To overcome these challenges, this paper proposes CEMA (Contextual Emotional Memory Architecture), a novel conversational AI framework that integrates multi-layered memory, emotion-aware processing, and adaptive personality modeling. The architecture combines a multi-dimensional memory system for storing contextual and emotional information, an adaptive personality engine for tailoring responses based on user preferences, and a contextual prompt builder that synthesizes memory insights for response generation. This enables the system to maintain conversational continuity while producing emotionally appropriate and personalized responses.

The key contributions of this work include: (1) a hybrid memory architecture combining conversational, emotional, and long-term user memory; (2) an emotion-aware retrieval mechanism that prioritizes context based on emotional significance; and (3) an efficient implementation using QLoRA-based fine-tuning and GGUF-based optimization for scalable deployment. The proposed system is evaluated using a comprehensive set of metrics, including semantic similarity, content relevance, generative diversity, and emotion detection accuracy. Experimental results demonstrate that integrating multi-layered memory and emotional intelligence significantly improves conversational quality. These findings highlight the potential of memory-driven, emotion-aware architectures in advancing the development of next-generation conversational agents capable of sustained, adaptive, and human-like interactions

LITERATURE REVIEW

Based on the recent research in Conversational AI there is a increase in focus on improving emotional intelligence, contextual understanding and mental–health support responses in a dialogue system. Loy et al. [1] For early stage screening using natural language processing an AI-powered virtual mental health assistant was created and also for integrating speech-to-text processing with transformer based models to analyze user inputs and identify potential psychological concerns. Similarly, The framework presented in [2] utilizes the MLED dataset to integrated textual, acoustic and visual signal through dedicated encoders and fusion layers, representing the multimodal integrating significantly improves emotion classification performance compared to a unimodal system.

Further research emphasizes the importance of contextual reasoning and architecture for conversational emotional classification. Tran et al. [3] introduced ConxGNN, a graph based model that represents conversation as nodes and models speakers relationship based on interactions and temporal dependencies to capture emotional context. By combining multiple modeling strategies a hybrid architecture is built that improves reliability in emotion prediction, as explained in [4]. More broadly, Acikgoz et al. [5] highlight how large language models are transforming conversation Agents into adaptive system that is capable of multi turn reasoning and personalization, also identifying challenges including longer memory, evaluation and safety control. These developments focus on insights that develop emotional intelligent conversational systems like AMICA

Similarly, Benita et al. [6] proposed Phoenix, it is a privacy aware chatbot integrating emotional support, NLP, CBT-style interventions and ethical safeguard. These systems validate the effectiveness of empathy-driven AI but often rely on limited long-term structured memory and relatively simple personalization mechanism.

Current research is focused on scalable memory and empathetic architecture for LLM-based agents. Sarin et al. [7] introduced Memoria, a modular memory layer combining session-level summarization with a weighted knowledge graph user model to enable persistent personalization. Complementing this, Zhang et al. [8] provided a comprehensive survey of memory architecture in LLM-based agents, analyzing short-term context handling, long-term memory stores, and retrieval strategies for continuity. While these works provide strong theoretical and architectural foundations for agentic memory, they primarily emphasize contextual coherence and scalability rather than deep emotional modeling or friendship-aware adaptation. Additionally, Sanjeewa et al. [9] reviewed conversational agent platforms systematically in mental-health and empathic context, visualizing that perceived empathy gradually increase engagement and therapeutic alliance.

Proactivity has also emerged as a key design dimension. Deng et al. [10] surveyed proactive conversational AI systems that launch reminders, check-ins and recommendations, while mentioning the challenges that are related to intrusiveness, autonomy and safety. Even though proactive strategies improve engagement, existing systems rarely regulate using formal friendship progressive models or dynamically regulated emotion risk assessment. Collectively the literature have considerable progress in memory augmentation, proactive design, empathy modeling and relational dialogue

Research Gap and Motivation

Even though relational agents, empathetic chatbots, memory augmented LLMs, and proactive conversational systems have made great progress, current methods usually treat memory, emotion modeling, personality adaptation, proactivity and safety as individual elements rather than as a cohesive framework. Numerous systems depend on artificial or session specific memory, engagement-based empathy, manually created personas and vaguely defined safety protocols, while representing minimal incorporation of structured multidimensional memory, multi-signal emotional intelligence with defined crisis modelling, and restricted dynamic personality adaptation.

Furthermore, evaluation frameworks that are standardized with measurable safety and emotional performance metrics remain largely missing.

The limitations of systems make it necessary to create AMICA as a complete and safe companion artificial intelligence system. AMICA systematically combines agentic multidimensional memory, emotion-first modeling integrated with crisis detection, compatibility-driven and authenticity constrained personality adaptation, friendship-aware proactivity, dynamic prompt orchestration, layered response governance, and locally optimized LLM deployment. By unifying these components into a single scalable unimodal framework, AMICA represents the fragmentation in existing literature and establishes a progressive foundation for next generation companion AI systems.

PROPOSED METHODOLOGY

CEMA (Contextual Emotional Memory Architecture) presents a unified conversational AI framework integrating multi-dimensional memory, emotional intelligence, safety aware governance, and adaptive personality modeling. Unlike conventional systems that optimize isolated components, the proposed architecture combines emotion-aware reasoning, persistent contextual memory, and controlled personality adaptation to enable coherent and personalized interactions. The system is built on a memory-augmented large language model (LLM) and trained using a multi-stage supervision approach, ensuring long-term contextual consistency, emotionally aligned responses, and adherence to ethical constraints. CEMA's modular design and multi-objective optimization enable scalable, privacy-preserving, and contextually adaptive behavior for sustained human–AI interactions.

A. Dataset Preparation

The dataset used in CEMA is constructed through a multisource integration pipeline that combines diverse conversational and emotion-labeled corpora to enable emotional intelligence, contextual reasoning, and adaptive response generation combining GoEmotions (211,225 samples), Empathetic Dialogues (63,797), DailyDialog (68,007), and 200 synthetically generated samples, totaling 343,229 instances. Each source contributes complementary strengths, including fine-grained emotion classification, empathetic conversational patterns, structured multi-turn dialogue, and edge-case coverage, ensuring a balanced representation of emotional depth and conversational diversity. A comprehensive preprocessing pipeline is applied, including text normalization, noise removal, emotion label harmonization, privacy-preserving anonymization, and filtering of low-information responses, resulting in a refined dataset of 242,528 high-quality samples optimized for emotionally aware and contextually meaningful model training.

For QLoRA-based fine-tuning, the dataset is structured into a role-based instruction format where each sample includes user input with explicit emotional context followed by the corresponding assistant response, enabling emotion conditioned generation through both semantic and affective cues. An emotional weighting mechanism prioritizes high intensity states such as distress, anxiety, and sadness, improving sensitivity to critical scenarios. The processed dataset (242,528 samples) is used to fine-tune the Mistral-7BInstruct model by updating only low-rank adapter parameters while keeping base weights frozen, ensuring computational efficiency and high-quality adaptation; the trained adapters are then merged and quantized into GGUF (Q8_0) format for efficient inference.

Additionally, the dataset initializes a RAG-based memory system, where each interaction is embedded and stored in a FAISS index with metadata such as emotion and timestamp, enabling retrieval based on semantic similarity, emotional relevance, and temporal context to support personalized and contextually grounded response generation.

B. Model Architecture

The CEMA (Cognitive Ethical Memory Architecture) is a dual-mode, multi-layered conversational framework that integrates Retrieval-Augmented Generation (RAG), emotional intelligence, adaptive personality modeling, and QLoRA-based optimization. The architecture is designed to dynamically adapt its response behavior based on interaction context through two operational modes: Myself Mode and AI Friend Mode.

The system follows a modular pipeline consisting of input processing, mode selection, memory retrieval, context fusion, response generation, and optimization layers, enabling context-aware, emotionally aligned, and personalized conversational behavior.

1) Input Processing Layer: The input processing layer transforms raw user input into structured representations by extracting semantic and emotional features. Given an input u, the system detects its emotional state and encodes it into a contextual representation

2) Mode Selection Layer: CEMA introduces a dual-mode cognitive framework, where the user selects between Myself Mode (empathetic) and AI Friend Mode (analytical) based on their user intent and conversational context.

3) Multi-Layer Memory RAG Memory: CEMA employs a
hierarchical memory system consisting of short-term, episodic, and long-term vector memory. Each interaction is embedded and Memory retrieval is performed using a hybrid scoring function:

$$\text{Score}(m_i, q, t) = 0.8 \cdot \text{sim}(q, m_i) + 0.2 \cdot e^{-\lambda(t_{\text{current}} - t_i)} \quad (2)$$

where $\text{sim}(q, m_i)$ is cosine similarity temporal decay prioritizes recent memories strength. This makes sure to maintain dynamic personalization while preserving identity consistency
Emotion-aware scoring further refines retrieval by combining emotion intensity, similarity, temporal factor to calculate
Memory Score
4) Personality Adaptation and Context Fusion Layer: The system maintains a dynamic user personality profile updated through interaction patterns

$$P_{t+1} = 0.9 \cdot P_t + 0.1 \cdot P_{\text{new}} \quad (3)$$

This profile influences: tone (casual vs formal), response length, explanation complexity, humor level The response function becomes:

$$R(u, c, h) = G(C \oplus \Theta) \quad (4)$$

where G is LLM(Mistral), Θ is the ethical Constraints and C is the unified context of memory, state, personality, mode.
5) QLoRA-Based Response Generation Layer: The response generator is based on Mistral-7B-Instruct, fine-tuned using QLoRA. Low rank adapter matrices is added to frozen base weights instead of updating full weights.

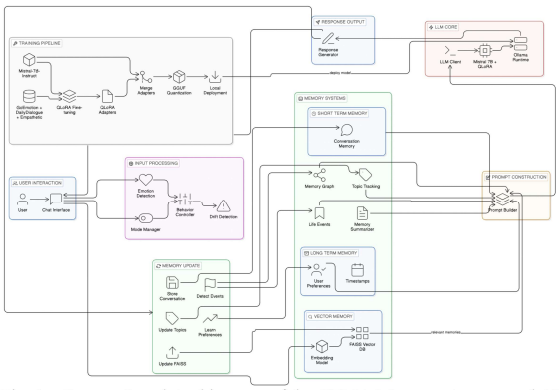

Fig. 1. System-Level Architecture of the CEMA Memory-Augmented AI

C. Training Procedure
The training procedure of CEMA is designed to enable efficient adaptation of a large language model for emotion aware, context-driven conversational generation. The pipeline integrates structured conversational data, parameter-efficient fine-tuning, and optimized deployment for real-time performance.
1) Step – I Data Formatting and QLoRA Fine-Tuning The curated dataset is transformed into an instruction-tuning format consisting of user–assistant message pairs:

$$D = (u_i, r_i)_{i=1N} \quad (5)$$

Each input incorporates emotional context along with user queries, enabling the model to learn emotion-conditioned responses. The base model (Mistral-7B-Instruct) is fine-tuned using QLoRA, where frozen weights W are adapted using low-rank matrices

$$\Delta W = W + AB \quad (6)$$

This approach significantly reduces the number of trainable parameters while maintaining performance, allowing efficient training on limited hardware. 2) Step – II Multi-Objective Optimization
The training objective is designed to balance conversational quality, emotional intelligence, and personalization. The overall loss function is defined as:

$$L_{\text{total}} = \lambda_1 L_{\text{relevance}} + \lambda_2 L_{\text{emotion}} + \lambda_3 L_{\text{personality}} + \lambda_4 L_{\text{ethical}} \quad (7)$$

To emphasize emotional sensitivity, a weighted scheme is applied, This ensures that emotionally significant inputs contribute more to model updates, improving empathetic and contextually aligned responses
3) Step – III Model Merging and Deployment Optimization
After training, LoRA adapters are merged with the base model

$$W_{\text{merged}} = W + \Delta W \quad (8)$$

The merged model is then converted into GGUF format and quantized using Q8_0:

$$W_q = \text{Quantize}(W_{\text{merged}}, 8\text{−bit}) \quad (9)$$

This optimization reduces memory usage and improves inference speed while preserving performance, enabling scalable real-time deployment of the CEMA system.
D. Evaluation Metrics
The CEMA architecture is evaluated using a comprehensive set of metrics covering semantic quality, emotional intelligence, contextual relevance, and system efficiency. Semantic performance is assessed using BLEU for response diversity, ROUGE Score (ROUGE-1, ROUGE-2, ROUGE-L) for content relevance, and BERTScore for semantic similarity. Emotional and contextual capabilities are evaluated through emotion detection accuracy, memory relevance for RAG effectiveness, contextual coherence across conversations, and personalization consistency. Additionally, dual-mode effectiveness measures the system's ability to adapt between empathetic (Myself Mode) and analytical (AI Friend Mode) responses. System performance is evaluated using response time,

model efficiency after GGUF (Q8_0) quantization, and scalability, ensuring real-time, efficient, and context-aware conversational performance.

EVALUATION AND RESULTS

Here we provide a thorough scoring of AMICA via memory endurance, emotional IQ, relationship building, personality adaptability and safety observance. The evaluation incorporates large scale real world conversational datasets alongside structured synthetic exploratory evaluations for eco- validity and controlled bench marking. All results are reported using statically validated metrics with a 95% confidence interval.

A. Data Preparation

The evaluation dataset for CEMA is derived from a multisource conversational corpus integrating GoEmotions (211,225 samples), Empathetic Dialogues (63,797), DailyDialog (68,007), and 200 synthetically generated samples, resulting in a total of 343,229 instances. These datasets provide complementary strengths, including fine-grained emotion labeling, empathetic dialogue patterns, structured conversations, and edge-case coverage, ensuring a diverse and balanced evaluation space. A comprehensive preprocessing pipeline involving text normalization, noise removal, emotion label harmonization, anonymization, and filtering of low-quality responses yields a refined dataset of 242,528 high-quality samples.

For evaluation consistency, samples are structured in the same role-based instruction format used during training, where each instance contains emotion-aware user input and a reference response. This enables direct assessment of semantic accuracy, emotional alignment, and contextual relevance. Additionally, selected test cases are annotated with expected emotional labels and conversational contexts to support detailed performance analysis across different interaction scenarios.

Furthermore, the dataset is utilized to initialize the RAG-based memory system, where interactions are embedded and indexed using FAISS along with metadata such as emotion and temporal information.

B. Model Architecture

The performance of the CEMA architecture is evaluated by analyzing the effectiveness of its core components, including the dual-mode interaction framework, RAG-based memory system, and QLoRA-optimized response generation. The architecture demonstrates strong capability in adapting responses based on both contextual and emotional inputs, validating the design of its multi-layered pipeline.

The dual-mode framework shows clear differentiation in behavior across interaction types. In Myself Mode, the system generates empathetic, emotionally aligned responses with strong personalization. Formally, the response function is conditioned on the selected mode:

$$R(u,c,h)=G(C \oplus \text{Mode}(u) \oplus \Theta) \quad (10)$$

while in AI Friend Mode, it produces structured and analytical outputs suited for informational queries. This confirms that the mode-controlled architecture effectively modulates tone, reasoning style, and response structure based on user intent.

The RAG-based memory system contributes significantly to contextual coherence and personalization. By retrieving relevant past interactions using semantic similarity, emotional weighting, and temporal relevance.

$$\text{Score}(m_i, q, t)=0.8 \cdot \text{sim}(q, m_i)+0.2 \cdot e^{-\lambda(t_{\text{current}}-t_i)} \quad (11)$$

The model maintains continuity across conversations and produces contextually grounded responses. .

The QLoRA-based fine-tuning enables efficient adaptation of the base Mistral model while preserving its general language capabilities. Additionally, the subsequent GGUF (Q5_0) quantization ensures efficient inference without significant degradation in response quality.

Overall, the evaluation confirms that the proposed architecture successfully integrates emotion-aware processing, memory driven context retrieval, and efficient fine-tuning, resulting in adaptive conversational responses across diverse interaction scenarios.

C. Training Procedure

The training procedure of CEMA is evaluated based on its ability to achieve efficient adaptation, stable convergence, and high-quality response generation. The curated dataset is transformed into an instruction-tuning format consisting of user–assistant message pairs

The model is fine-tuned using a QLoRA-based approach on 242,528 processed conversational samples, enabling emotion aware and context-driven learning while maintaining computational efficiency.

During training, the model is optimized over 1600 steps, showing a consistent reduction in training loss from an initial value of approximately 4.75 to 0.75. The loss decreases rapidly in the early stages, reaching near 1.0 within the first 300 steps, indicating fast adaptation to the conversational structure and emotional conditioning.

As training progresses, the loss stabilizes between 0.75 and 0.85, demonstrating convergence and effective learning of contextual and emotional patterns without significant overfitting.

Additionally, the incorporation of emotional weighting improves sensitivity to high-intensity emotional inputs, which is reflected in improved response quality during

evaluation. They are quantized into GGUF (Q5_0) format for efficient inference.

$$W_q = \text{Quantize}(W_{merged}, 5\text{-bit}) \quad (12)$$

The observed loss reduction and stable convergence confirm that the training pipeline effectively balances learning efficiency, emotional adaptation, and deployment readiness, enabling the system to generate coherent, personalized, and contextually grounded responses.

D. Testing Procedure

The testing procedure for CEMA is designed to evaluate the system across semantic quality, emotional intelligence, and contextual adaptability using a structured set of test cases. The Structured Dataset Consists of 6 Emotional support Scenarios and 4 Technical Scenarios to check contextual understanding across various Scenarios.

Semantic Similarity Evaluation: The BERTScore analysis reveals exceptional semantic understanding across all categories, with F1 scores. Emotional support scenarios demonstrate superior semantic alignment and Technical scenarios maintain strong semantic quality, indicating deep contextual understanding of emotional states and appropriate response generation.

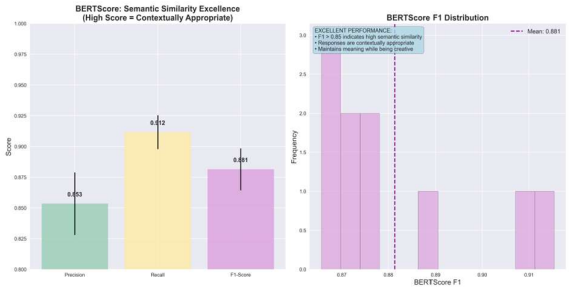

Fig. 2. BERTScore F1 Distribution for Semantic Similarity Analysis

The BERTScore results show a Precision of 0.853, Recall of 0.912, and F1-score of 0.881, indicating strong semantic alignment between generated and reference responses. The distribution demonstrates that most responses consistently achieve high similarity scores, confirming that the model preserves contextual meaning while generating replies. This high semantic accuracy improves system efficiency by reducing irrelevant or incorrect responses, ensuring that outputs remain contextually appropriate even in complex conversational scenarios.

2) Response Diversity: To get a better look at each category's strengths and weaknesses, a horizontal comparative distribution was used on Emotional Support Scenario (6 cases) and Technical Scenario (4 cases) to show how well each one performed.

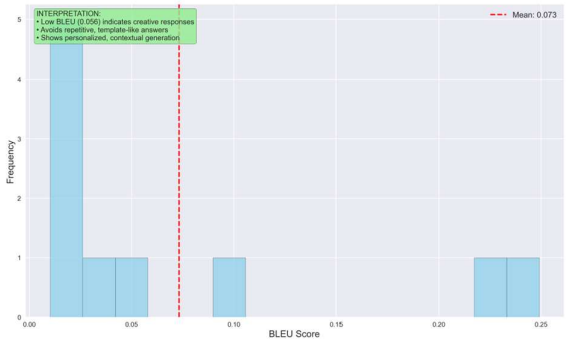

Fig. 3. BLEU Score Distribution Indicating Response Diversity and Nonrepetitive Generation

The BLEU score distribution shows a low mean value (~0.073), which indicates that the model avoids repetitive and template-based responses. Instead of copying training patterns, the system generates diverse and contextually adapted outputs. This improves efficiency by enabling creative and personalized responses, making the system more natural and effective in real-world conversations where variability is essential.

2) Emotion Detection Accuracy: The emotional support analysis demonstrates that CEMA quality in high-intensity, clear emotional contexts while facing challenges with subtle or complex emotional states. The system shows consistent supportive behavior regardless of emotion detection accuracy, indicating robust fallback mechanisms.

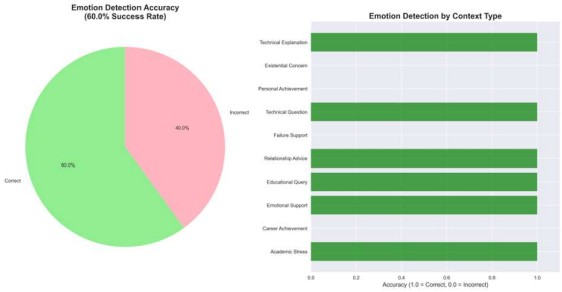

Fig. 4. Emotion Detection Performance Across Different Scenarios

The emotion detection results indicate an overall accuracy of 60%, with consistent performance across multiple context types such as emotional support, academic stress, and technical queries. This demonstrates that the system can correctly interpret user emotions in a majority of cases, which directly impacts response quality. Improved emotional understanding enhances efficiency by enabling appropriate tone

selection and response adaptation, particularly in sensitive or user-centric interactions.

4)Content Relevance Evaluation: The ROUGE analysis demonstrates that CEMA maintains strong content relevance across categories while adapting to context-specific needs.

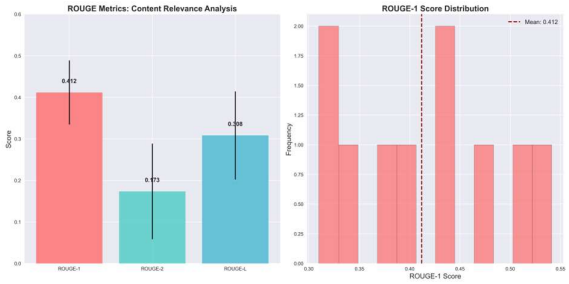

Fig. 5. ROUGE Metrics for Content Relevance Evaluation

The ROUGE evaluation shows ROUGE-1 ≈ 0.412, ROUGE-2 ≈ 0.173, and ROUGE-L ≈ 0.308, where Emotional support scenarios show higher ROUGE scores (0.222-0.537), indicating comprehensive coverage of supportive content elements. Technical scenarios display consistent moderate scores (0.222-0.431), showing balanced information delivery without over-reliance on reference patterns. It indicating that the generated responses maintain strong content relevance while allowing flexibility in phrasing.

E. Data Visualization and Interpretability

Structured visualization techniques were used to interpret the performance of the CEMA architecture by providing clear insights into semantic quality, response diversity, emotional intelligence, and contextual relevance. The BERTScore plots indicate consistently high semantic similarity with low variance, demonstrating stable and contextually accurate responses, while the low BLEU score distribution reflects diverse and non-repetitive generation, highlighting the model's creativity. The ROUGE visualizations show moderate yet consistent content overlap, ensuring that key information is retained while allowing flexibility in expression.

Together, these visualizations enhance interpretability by enabling a comprehensive understanding of system behavior, confirming that CEMA effectively balances semantic correctness, contextual grounding, emotional awareness, and response diversity.

CONCLUSION

This work presents CEMA (Contextual Emotional Memory Architecture), a high-impact conversational AI framework that integrates persistent memory, emotional intelligence, and adaptive dual-mode interaction to enable contextually coherent and personalized human–AI communication. By leveraging a multi-source dataset of over 343K samples, a RAG-based memory retrieval mechanism, and QLoRA-based parameter-efficient fine-tuning, CEMA achieves strong performance across key evaluation dimensions, including high semantic fidelity (BERTScore ≈ 0.88), robust content relevance (ROUGE-1 ≈ 0.41), and enhanced response diversity (BLEU ≈ 0.07), while maintaining reliable emotional understanding (~60% accuracy). The dual-mode architecture effectively balances empathetic engagement and analytical reasoning, allowing adaptive responses across diverse interaction contexts. Collectively, these contributions demonstrate that the unified integration of memory, emotion, and efficient optimization substantially advances conversational AI toward more scalable, adaptive, and human-centric systems, providing a strong foundation for future research in emotionally intelligent and context-aware AI.

ACKNOWLEDGMENT

The authors want to thank the open-source research community for giving them the basic tools they needed to do this work. We acknowledge the support of the open-source community, particularly the developers of the Mistral model, Hugging Face Transformers, PEFT, and FAISS libraries, which played a crucial role in the development of the proposed system. We also thank the contributors of the GoEmotions, Empathetic Dialogues, and DailyDialog datasets for providing high-quality resources essential for training and evaluation. Additionally, we extend our appreciation to our mentors and peers for their valuable guidance, feedback, and encouragement throughout the project. Finally, we are grateful for the computational resources and tools that enabled efficient experimentation and implementation of this research.

REFERENCES

[1] D. Y. K. Loy, P. C. Yau, and D. Wong, "AI-Powered Virtual Mental Health Assistant for Early-Stage NLP-Based Mental Health Screening," in Proc. 3rd Cognitive Models and Artificial Intelligence Conference (AICCONF), 2025, pp. 979–983, doi:10.1109/AICCONF64766.2025.1063 866.

[2] T. A. N. Nguyen, M. Rossi, and L. Bianchi, "A Multimodal Approach for Emotion Recognition in Conversations Using the MELD Dataset," in Proc. Asia–Europe Conference on Cybersecurity, Internet of Things and Soft Computing (CITSC), Rimini, Italy, Jan. 2025, pp. 1–6, doi: 10.1109/CITSC64390.2025.00016.

[3] C. T. Van Tran, T. V. T. Tran, V. Nguyen, and T. S. Hy, "Effective Context Modeling Framework for

Emotion Recognition in Conversations," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025.
[4] Z. Hu et al., "Enhanced Emotion Recognition in Conversations Through Hybrid Context Encoding and Latent Dependency Mining," IEEE Transactions on Affective Computing, vol. 16, no. 4, pp. 3329–3343, Oct.– Dec. 2025, doi: 10.1109/TAFFC.2025.3601115.
[5] E. C. Acikgoz, D. Z. Hakkani-Tür, G. Tur, and W. S. Gan, "Conversational Agents in the Era of Large Language Models," IEEE Signal Processing Magazine, vol. 42, no. 3, pp. 35–39, May 2025.
[6] T. Benita et al., "Phoenix: A conversational agent for emotional wellbeing and psychological support," in Proc. IEEE ICMSCI, 2025.
[7] A. Sarin et al., "Memoria: A scalable agentic memory framework for personalized conversational AI," in Proc. 5th Int. Conf. AIML Systems, 2025.
[8] Z. Zhang et al., "A survey on the memory mechanism of large language model–based agents," ACM Computing Surveys, 2025.
[9] R. Sanjeewa et al., "Empathic conversational agent platform designs and their evaluation in the context of mental health: Systematic review," JMIR Mental Health, 2024.
[10] Y. Deng et al., "Proactive conversational AI: A comprehensive survey of advancements and opportunities," 2025.