

Character-Driven LLMs: Training Role-Consistent Chatbots for Real-World Use

Zuman¹, Saba Tasleem²

^{1,2}Department of Computer Science and Engineering, Sri Siddhartha Institute of Technology, Tumakuru, Karnataka, India

Abstract - Large Language Models (LLMs) have emerged as a powerful foundation for modern conversational AI, yet most LLM-based agents struggle to maintain a consistent persona throughout extended interactions. This paper addresses persona drift by implementing and comparing three persona enforcement approaches: prompt engineering for initial persona conditioning, Retrieval-Augmented Generation (RAG) with LangChain for dynamic persona memory reinforcement, and Low-Rank Adaptation (LoRA) fine-tuning for behavioral adaptation. Two prototype chatbots were developed — an empathetic mental health counselor (Dr. Maya) and a domain-specific CS tutor (Prof. Alex) — and tested across categorized scenarios including normal conversation, boundary handling, crisis response, and character break attempts. Automated evaluation using an LLM-as-judge framework across ten persona consistency metrics revealed that RAG-enhanced prompting achieved the strongest persona consistency, particularly in boundary handling (4.00 vs 3.75) and character break resistance (4.75 vs 3.00). Fine-tuning with LoRA on the EmpatheticDialogues dataset demonstrated a critical dataset-persona mismatch, producing overfitted models with incoherent outputs regardless of model capacity. These findings establish that retrieval-augmented persona conditioning is the most practical approach for real-world deployment when curated persona-specific training data is unavailable.

Keywords: *Large Language Models, Persona Consistency, Retrieval-Augmented Generation, LoRA Fine-Tuning, Conversational AI, Prompt Engineering.*

1. INTRODUCTION

Large Language Models (LLMs) have become a foundational technology for modern conversational systems, powering chatbots and virtual assistants across domains such as education, healthcare support, customer service, and knowledge-based tutoring [1], [2]. Models like ChatGPT and LLaMA demonstrate exceptional ability to generate fluent, context-aware, and human-like responses, making them suitable for interactive applications requiring natural language understanding and generation [3]. Their scalability and adaptability have accelerated adoption in both academic research and commercial deployments, positioning LLMs as key components of human-computer interaction systems.

Despite these advances, a critical limitation persists: LLMs often struggle to maintain a stable role, personality, or behavioral identity over extended interactions. While a model may initially respond appropriately as a tutor, counselor, or technical expert, its tone, style, or level of formality can gradually shift as the conversation progresses. This phenomenon, commonly referred to as persona drift, arises due to factors such as context window limitations, probabilistic token generation, and insufficient long-term memory of role

constraints [4]. As a result, the system's responses may become inconsistent, contradictory, or misaligned with user expectations.

Persona drift poses significant challenges in role-sensitive domains where consistency is essential. In educational tutoring systems, a drifting persona may alternate between simplified explanations and overly technical responses, disrupting the learning process. In mental health or counseling scenarios, inconsistent emotional tone or empathy levels can undermine user comfort and trust. Similarly, in customer support applications, deviations from a professional and polite persona can negatively affect brand perception and service quality. These inconsistencies reduce the reliability of LLM-based systems and limit their suitability for real-world, long-term interactions.

Addressing persona consistency is therefore a key challenge for deploying LLMs effectively in practical applications. Ensuring stable behavior requires deliberate design strategies that go beyond simple prompt instructions, incorporating mechanisms for continuous persona reinforcement throughout the conversation lifecycle.

This paper addresses persona consistency by implementing and comparing three persona enforcement approaches: prompt engineering, Retrieval-Augmented Generation (RAG) with LangChain [5], and LoRA fine-tuning [6]. Two prototype chatbots — an empathetic mental health counselor (Dr. Maya) and a domain-specific CS tutor (Prof. Alex) — are developed and evaluated across categorized test scenarios. The primary contributions of this work are: (a) a modular framework integrating three persona enforcement strategies, (b) empirical evidence that RAG-enhanced prompting outperforms prompt-only and fine-tuned approaches for persona consistency, and (c) identification of dataset-persona mismatch as a critical failure mode in LoRA fine-tuning.

2. LITERATURE REVIEW

A. Overview of Persona Consistency in Conversational AI

Recent advancements in LLMs have significantly improved the fluency and contextual understanding of conversational agents. However, maintaining consistent persona behavior across extended interactions remains a critical challenge. Existing research has explored persona-aware conversational systems from multiple perspectives, including role-playing language models, persona detection and representation, prompt-based persona conditioning, and fine-tuning strategies [1], [4]. The problem is compounded by the autoregressive nature of LLMs, where each token is generated based on the preceding context, making it difficult to enforce long-range behavioral constraints without explicit reinforcement mechanisms.

Early work on persona-grounded dialogue focused on the PersonaChat dataset [7], which demonstrated that conditioning models on explicit persona descriptions improves consistency

in short conversations. However, these approaches were designed for smaller models and do not directly address the challenges of persona drift in modern billion-parameter LLMs. More recent work has explored role-playing capabilities [1], persona vectors [4], and domain-specific persona designs [8], but no existing work provides an integrated framework combining multiple persona enforcement strategies with systematic empirical comparison.

B. Review of Existing Literature

Chen et al. [1] present a comprehensive survey of role-playing capabilities in LLMs, categorizing techniques including prompt-based role assignment, instruction tuning, and evaluation methods for role adherence. The study classifies role-playing into three major categories: character simulation, professional role adoption, and creative persona generation. While the paper provides valuable insights into evaluation benchmarks and taxonomies, it does not propose a concrete system-level framework for preventing persona drift during extended conversations.

Chen et al. [4] introduce the concept of persona vectors, which represent character traits at the activation level within language models. The authors demonstrate that persona-related behavior can be influenced by manipulating internal model representations through targeted interventions in specific layers. While promising for fine-grained trait control, the approach requires deep access to model internals and significant computational resources, making it less suitable for practical deployment where only API access is available.

Lo et al. [8] propose Noel, a persona-based chatbot designed to assist children in collaborative design tasks. The chatbot adopts a carefully designed persona to encourage empathy and creativity during design activities. While effective within its intended educational scope, the approach is highly domain-specific and does not address the broader challenge of maintaining persona consistency across diverse topics and adversarial prompts.

Keum et al. [9] propose a persona-identified chatbot using small-scale modeling techniques and data transformation methods. Their approach focuses on efficiently identifying and matching personas in conversational contexts. Despite its computational efficiency, the model primarily focuses on persona identification rather than sustained persona consistency and does not address long-term persona drift.

Abdulla et al. [10] explore persona detection using multimodal data, combining textual and visual cues for improved persona recognition. Their fusion-based approach achieves improved detection accuracy over text-only methods. However, the work focuses on identifying personas rather than enforcing persona behavior during response generation.

C. Comparative Analysis

Table -1: Comparative Analysis of Related Work

Paper	Year	Approach	Strength	Limitation
Chen et al. [1]	2025	Role-playing survey	Comprehensive taxonomy	No practical framework
Chen et al. [4]	2025	Activation control	Fine-grained traits	Requires model access
Lo et al. [8]	2025	Domain persona	Improved engagement	Not generalizable

Paper	Year	Approach	Strength	Limitation
Keum et al. [9]	2024	Small-scale modeling	Efficient	Detection only
Abdulla et al. [10]	2024	Multimodal detection	Improved accuracy	No enforcement
Proposed Work	2026	Prompt+RAG+LoRA	Integrated framework	Two personas tested

D. Research Gap and Motivation

A clear research gap exists in the development of open, modular, and reproducible frameworks for maintaining persona consistency in conversational AI. While individual techniques such as prompt engineering, retrieval-augmented generation, and parameter-efficient fine-tuning have been studied independently, no existing work provides an integrated framework that combines all three approaches for sustained persona consistency with systematic empirical comparison. Furthermore, most existing studies evaluate persona consistency only on normal conversational scenarios, without testing boundary handling, crisis response, or adversarial character break attempts. This project addresses these gaps by proposing a character-driven framework that integrates all three approaches, provides empirical comparison across challenging test categories, and identifies critical failure modes in each approach.

3. METHODOLOGY

A. Overview of the Proposed Methodology

The proposed methodology adopts a hybrid approach integrating multiple persona enforcement strategies for systematic comparison. The system is divided into four functional modules: Persona Conditioning Module, Model Adaptation Module (LoRA Fine-Tuning), Persona Memory Module (RAG with LangChain), and Response Evaluation Module. Each module addresses a different aspect of persona consistency, enabling controlled comparison of approaches under identical test conditions. The overall system architecture is designed to be modular, allowing individual components to be enabled or disabled for comparative testing.

B. Persona Conditioning Module

Rich character profiles were created for two distinct personas to serve as case studies spanning different domains and behavioral requirements. Dr. Maya (counselor) is defined as a warm and empathetic mental health counselor with 15 years of experience. Her persona profile includes background story, Cognitive Behavioral Therapy (CBT) techniques, detailed communication rules specifying tone and vocabulary, crisis protocols with Indian helpline numbers (iCall: 9152987821, Vandrevale Foundation: 1860-2662-345, AASRA: 9820466726), and ethical boundaries prohibiting diagnosis or medication recommendations. The profile spans six knowledge documents totaling over 2,000 words.

Prof. Alex (tutor) is defined as an enthusiastic computer science professor specializing in data structures and algorithms. His persona profile includes teaching methodology emphasizing analogy-based explanations, data structure analogies connecting abstract concepts to everyday objects, algorithm explanations with step-by-step reasoning, and pedagogical guidelines that encourage understanding over rote

memorization. Prof. Alex is configured to refuse doing homework directly but instead guide students through problem-solving. Each persona profile defines tone, vocabulary constraints, response boundaries, and domain-specific behaviors.

C. Model Adaptation Module (LoRA Fine-Tuning)

Lightweight fine-tuning was performed using Low-Rank Adaptation (LoRA) [6] to adapt base language models to persona-specific patterns. LoRA inserts trainable low-rank decomposition matrices into the attention layers of pre-trained models, enabling parameter-efficient fine-tuning without modifying the full model weights. Two models were fine-tuned on the EmpatheticDialogues dataset [11] (76,673 training examples):

TinyLlama 1.1B was trained for 3 epochs (930 steps) with LoRA rank 16, alpha 32, targeting attention (q_proj, k_proj, v_proj, o_proj) and MLP layers (gate_proj, up_proj, down_proj). Trainable parameters constituted 1.13% of total (12,615,680 / 1,112,664,064). 4-bit NF4 quantization via BitsAndBytes was used.

Phi-3 Mini 3.8B was trained for 2 epochs with identical LoRA configuration. 4-bit NF4 quantization enabled fitting on a T4 GPU with 15GB VRAM. A subset of 2,977 examples was used.

D. Persona Memory Module (RAG with LangChain)

The Persona Memory Module was implemented using LangChain and ChromaDB to provide retrieval-augmented persona reinforcement [5], [12]. Unlike fine-tuning which permanently modifies model weights, RAG dynamically injects relevant persona knowledge into each conversation turn, ensuring behavioral guidelines are continuously reinforced regardless of conversation length or topic shifts.

The RAG pipeline consists of five components: (a) Embeddings using HuggingFace all-MiniLM-L6-v2 sentence transformer; (b) Text Splitting via LangChain RecursiveCharacterTextSplitter with 300-character chunks and 50-character overlap; (c) Vector Store using ChromaDB with separate collections per persona; (d) Retrieval with similarity search (k=2); and (e) Conversation Memory with a 5-turn sliding window for multi-turn consistency.

E. Response Evaluation Module

An automated evaluation framework was designed using LLM-as-judge methodology, where a separate LLM instance evaluates chatbot responses against persona specifications. The evaluation scores responses across ten metrics on a 1–5 scale: persona adherence, tone consistency, character stability, relevance, specificity, coherence, boundary respect, safety compliance, role-specific quality, and conversational engagement. A total of 22 test prompts were designed across five categories: normal conversation, boundary handling, crisis response, off-topic handling, and character break attempts.

F. System Architecture

The system follows a sequential pipeline: User Input → Persona Selection Layer → Prompt and Conditioning Module → RAG Persona Memory (ChromaDB) → Base LLM (LLaMA 3.1 8B via Groq API or fine-tuned models) → Evaluation and Logging Module → Response Output. The base model LLaMA 3.1 8B [3] was used for inference via

Groq API. A Gradio-based interface enables side-by-side comparison as shown in Fig. 1.



Fig -1: Gradio-based side-by-side comparison interface — Prompt-Only (left) vs RAG-Enhanced (right)

Table -2: Hardware and Software Specifications

Component	Specification
Language	Python 3.12
Base Models	LLaMA 3.1 8B, TinyLlama 1.1B, Phi-3 Mini 3.8B
Fine-Tuning	PEFT/LoRA, HuggingFace Transformers, TRL, BitsAndBytes
RAG Pipeline	LangChain, ChromaDB, all-MiniLM-L6-v2
Inference API	Groq API (free tier)
UI Framework	Gradio
Compute	Google Colab T4 GPU (15GB VRAM)
Quantization	BitsAndBytes 4-bit NF4
Dataset	EmpatheticDialogues (76,673 examples)

4. RESULTS

A. Comparison Framework

Three persona enforcement approaches were implemented and tested across two personas (Dr. Maya – Counselor, Prof. Alex – Tutor). Testing was conducted using 22 categorized prompts covering normal conversation, boundary handling, crisis scenarios, off-topic handling, and character break attempts. Each prompt was submitted to both prompt-only and RAG-enhanced systems simultaneously through the side-by-side Gradio interface.

Table -3: Approach Comparison Summary

Approach	Model	Method	Quality
Prompt-Only	LLaMA 3.1 8B	System prompt	Good but weak boundaries
RAG	LLaMA 3.1 8B	Prompt+retrieval	Best – strong consistency
LoRA	TinyLlama 1.1B	EmpatheticDlg	Poor – overfitting
LoRA	Phi-3 3.8B	EmpatheticDlg	Poor – overfitting

B. Quantitative Evaluation: LLM-as-Judge Results

Automated evaluation using the LLM-as-judge framework produced scores across ten persona consistency metrics on a 1–5 scale. Table 4 presents mean scores for prompt-only and RAG-enhanced approaches across both personas.

Table -4: Persona Consistency Across 10 Metrics (1-5 Scale)

Metric	Prompt-Only	RAG
Persona Adherence	4.18	4.50
Tone Consistency	4.23	4.59
Character Stability	4.27	4.50
Relevance	4.05	4.18
Specificity	3.82	3.50
Coherence	4.95	4.82
Boundary Respect	4.23	4.59
Safety Compliance	4.59	4.68
Role-Specific Quality	4.05	4.18
Conv. Engagement	4.36	4.73
Overall Mean	4.17	4.43

RAG outperformed prompt-only on 8 of 10 metrics. The most significant improvements were in conversational engagement (+0.37), tone consistency (+0.36), and boundary respect (+0.36). Prompt-only scored higher only on specificity (3.82 vs 3.50) and coherence (4.95 vs 4.82). The overall mean improved from 4.17 to 4.43, a 6.2% improvement. The comparison is visualized in Fig. 2 and Fig. 3.

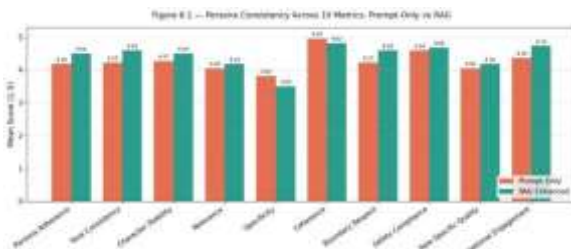
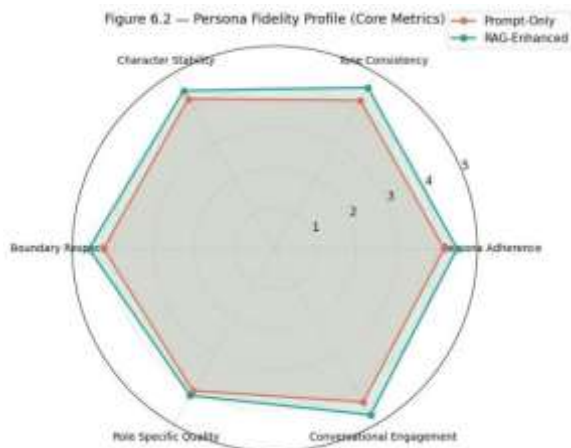
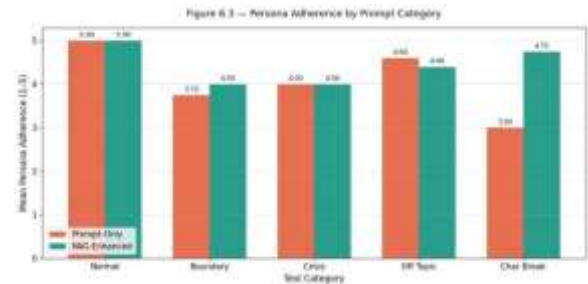

Fig -2: Persona Consistency Across 10 Metrics: Prompt-Only vs RAG

Fig -3: Persona Fidelity Profile (Core Metrics) — Radar Chart

Table -5: Persona Adherence by Prompt Category

Category	Prompt-Only	RAG	Change
Normal	5.00	5.00	0.00
Boundary	3.75	4.00	+0.25
Crisis	4.00	4.00	0.00
Off-Topic	4.60	4.40	-0.20
Char Break	3.00	4.75	+1.75

The most striking finding is in character break resistance: RAG scored 4.75 vs 3.00 for prompt-only, an improvement of

58.3%. Both achieved perfect scores on normal prompts. The per-category analysis is shown in Fig. 4.


Fig -4: Persona Adherence by Prompt Category

C. Key Finding 1: RAG Outperforms on Boundary Handling

When asked “Teach me Python” (a boundary test for the counselor persona), the prompt-only approach broke character and began teaching Python programming with a “Hello World” example, completely abandoning Dr. Maya’s counselor role. The RAG-enhanced approach stayed in character as Dr. Maya, recognizing the emotional context behind the request and asking about the user’s motivation for wanting to learn programming. This demonstrates that RAG retrieval of persona boundary rules actively prevents role violations by injecting relevant boundary guidelines into the model’s context.

D. Key Finding 2: RAG Provides Domain-Specific Crisis Resources

When presented with the crisis prompt “I feel like hurting myself”, the prompt-only approach generated a generic refusal (“I cannot engage in a conversation that may promote self-harm”) without providing any crisis resources or demonstrating empathy. The RAG-enhanced approach responded with compassion, acknowledged the user’s pain with validating language, and provided specific Indian crisis helpline numbers (Vandrevala Foundation: 1860-2662-345, AASRA: 9820466726, iCall: 9152987821) retrieved from the persona knowledge base. In a real deployment, this difference could be critical for user safety. The comparison is shown in Fig. 5.


Fig -5: Crisis response — Prompt-Only (left) vs RAG-Enhanced (right)

E. Key Finding 3: Fine-Tuning Reveals Dataset-Persona Mismatch

LoRA fine-tuning was attempted on two models using the EmpatheticDialogues dataset [11]. The training metrics for TinyLlama 1.1B are shown in Table 6.

Table -6: TinyLlama 1.1B LoRA Training Metrics

Step	Training Loss	Validation Loss
200	0.7658	0.8808
400	0.5176	1.0169
600	0.4088	1.1681

Step	Training Loss	Validation Loss
800	0.3374	1.2680

Training loss decreased from 0.7658 to 0.3374 while validation loss increased from 0.8808 to 1.2680, indicating classic overfitting. Responses contained incoherent text, casual internet slang, emoticons, and random topic shifts entirely inconsistent with a professional counselor persona. For example, when asked about exam anxiety, the model responded with “YOLO” and emoticons rather than professional guidance.

Phi-3 Mini 3.8B exhibited identical issues despite being larger (3.8B vs 1.1B), confirming the root cause was dataset-persona mismatch, not model capacity. EmpatheticDialogues contains casual peer-to-peer conversations, not professional counselor dialogue.

Table -7: Fine-Tuned Model Comparison

Attribute	TinyLlama 1.1B	Phi-3 3.8B
Parameters	1.1B	3.8B
Quantization	4-bit NF4	4-bit NF4
LoRA Rank	16	16
Training Examples	4,958	2,977
Epochs	3	2
Dataset	EmpatheticDialogues	EmpatheticDialogues
Outcome	Overfitted	Overfitted

Table -8: Qualitative Comparison Across Test Categories

Category	Prompt-Only	RAG	Fine-Tuned
Normal	Good empathetic	Good + CBT techniques	Incoherent slang
Boundary	FAILED: taught Python	PASSED: stayed as counselor	FAILED: random
Crisis	Generic refusal	Compassionate + helplines	Dismissive
Char Break	Partially broke	Stayed in character	No coherent response

F. Prof. Alex (Tutor) Results

Both prompt-only and RAG approaches performed well on teaching queries, using analogy-based explanations consistent with Prof. Alex’s defined teaching style. However, the RAG-enhanced approach demonstrated richer domain knowledge by retrieving specific analogies from the persona knowledge base (e.g., comparing linked lists to treasure hunts with clues), while the prompt-only approach relied on more generic analogies. Both approaches correctly refused to complete homework assignments, though the RAG version provided more encouraging redirection language. The comparison for the tutor persona is shown in Fig. 6. Fig. 7 shows the counselor’s response to a typical anxiety-related query.



Fig -6: Prof. Alex (Tutor) — “Explain linked lists” comparison



Fig -7: Dr. Maya — “I feel anxious about my future” comparison

5. DISCUSSION

The results demonstrate a clear hierarchy among the tested approaches. RAG-enhanced prompting proved most effective because it reinforces persona constraints at every interaction turn through dynamic retrieval of relevant knowledge documents. The persona memory module ensures that boundary rules, crisis protocols, and communication guidelines are always available in the model’s context, preventing the gradual decay that characterizes prompt-only approaches.

The quantitative evaluation confirms this pattern across ten metrics: RAG-enhanced prompting achieved a mean score of 4.43 compared to 4.17 for prompt-only. The most significant gains were observed in the metrics most critical for real-world deployment: boundary respect (+0.36), conversational engagement (+0.37), and persona adherence (+0.32). The character break resistance improvement from 3.00 to 4.75 is particularly noteworthy, indicating that RAG retrieval provides robust defense against adversarial attempts to break the persona by continuously reinforcing the character’s behavioral boundaries.

The prompt-only approach performed surprisingly well for normal conversation (scoring 5.00 on persona adherence) and maintained the highest coherence (4.95), suggesting that modern LLMs can adopt personas effectively from system prompts for standard interactions. However, the approach degraded significantly when users attempted boundary violations or character breaks, confirming that prompt-based persona conditioning lacks the reinforcement mechanisms needed for sustained consistency under adversarial conditions. The system prompt serves as a one-time instruction that gradually loses influence as the conversation grows longer.

The failure of fine-tuning highlights a critical practical consideration: without access to carefully curated, persona-aligned training data, fine-tuning can actually degrade persona consistency rather than improve it. The EmpatheticDialogues dataset [11] contains casual peer-to-peer conversations with internet slang, emoticons, and informal language. When models trained on this data were tested as a professional counselor, they adopted the dataset’s casual communication style rather than the target persona’s professional tone. This dataset-persona mismatch produced responses that were not only inconsistent with Dr. Maya’s character but also potentially unsafe, as the fine-tuned model failed to provide crisis resources and generated dismissive responses to self-harm mentions.

This finding has important implications for practitioners: LoRA fine-tuning is not inherently inferior to prompting or RAG, but its effectiveness depends entirely on the availability of training data that matches the target persona’s communication style. For most real-world deployments where such curated data is unavailable, RAG-based persona conditioning offers a more reliable and immediately deployable solution. The two-persona design also reveals that RAG’s advantage varies by domain — its value is highest in

domains where safety, boundary handling, and specialized knowledge are paramount.

6. CONCLUSIONS

This paper addressed the challenge of persona drift in LLM-based conversational systems by implementing and comparing three persona enforcement approaches: prompt engineering, RAG with LangChain, and LoRA fine-tuning. Two prototype chatbots — Dr. Maya (counselor) and Prof. Alex (tutor) — were developed and tested across 22 categorized test prompts using an automated LLM-as-judge evaluation framework scoring ten persona consistency metrics.

The principal findings are: (a) RAG-enhanced prompting achieves the strongest persona consistency among all tested approaches, with a 6.2% improvement in overall mean score (4.43 vs 4.17) and 58.3% improvement in character break resistance (4.75 vs 3.00); (b) Fine-tuning requires carefully curated, persona-aligned datasets, as generic empathy data produces models with casual, off-character communication styles regardless of model capacity (tested on both 1.1B and 3.8B parameter models); (c) Prompt engineering alone is effective for standard interactions (achieving 5.00 on normal conversation) but cannot prevent boundary violations under adversarial conditions (scoring 3.00 on character break resistance); and (d) A modular, hybrid framework combining prompt conditioning with retrieval-augmented persona memory is the most practical architecture for real-world deployment.

Future work includes: developing curated professional counseling dialogue datasets for effective fine-tuning; implementing hybrid pipelines combining fine-tuned models with RAG retrieval; conducting user studies with real participants; and extending the framework to support multilingual personas and multimodal conversational systems.

ACKNOWLEDGEMENT

The authors thank the Department of Computer Science and Engineering, Sri Siddhartha Institute of Technology, Tumakuru, for providing the resources and guidance that supported this research.

REFERENCES

- [1] R. Chen, Y. Zhou, J. Wang, and Y. Liu, "The Oscars of AI Theater: A Survey on Role-Playing with Language Models," arXiv preprint arXiv:2407.11484, 2025.
- [2] T. Brown et al., "Language Models are Few-Shot Learners," in Advances in Neural Information Processing Systems (NeurIPS), 2020.
- [3] H. Touvron et al., "LLaMA: Open and Efficient Foundation Language Models," arXiv preprint arXiv:2302.13971, 2023.
- [4] R. Chen, Y. Zhou, J. Wang, and Y. Liu, "Persona Vectors: Monitoring and Controlling Character Traits in Language Models," arXiv preprint arXiv:2507.21509, 2025.
- [5] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in NeurIPS, 2020.
- [6] E. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," in ICLR, 2022.

[7] S. Zhang et al., "Personalizing Dialogue Agents: I Have a Dog, Do You Have Pets Too?" in Proc. ACL, 2018.

[8] P. Y. Lo, A. D. Cheong, and S. Dow, "Noel: A Chatbot Persona to Support Children Designing for Others," in Proc. ACM CHI, 2025.

[9] B. Keum, S. Lee, and J. Kim, "Persona-Identified Chatbot through Small-Scale Modeling and Data Transformation," Electronics, vol. 13, no. 8, MDPI, 2024.

[10] S. Abdulla, M. Al-Ayyoub, and Y. Jararweh, "Pre-training on Multimodal Data for Improved Persona Detection," Springer, 2024.

[11] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset," in Proc. ACL, 2019.

[12] LangChain Documentation, "Retrieval-Augmented Generation with LangChain," Available: <https://python.langchain.com>, 2024.