Building a Conversational AI Therapist: Teaching Machines Empathy Through Data, Preferences, and Memory

Keertana Arun Vasan
School of Social Sciences

Vivek Choudhary
Nanyang Business School

Abstract - The growing disparity between the demand for psychotherapy and the limited supply of qualified professionals has prompted the exploration of innovative solutions to address accessibility challenges. This study proposes the development of an advanced conversational AI therapist as a scalable alternative to traditional mental health care. The objective was to implement a multi-phase methodology, integrating state-of-the-art AI techniques to create a system capable of providing empathetic and accurate mental health support. I fine-tuned the Llama 3.1 8B Instruct model using the MentalChat16K dataset, achieving a training loss of 0.7210. Direct Preference Optimization (DPO) on the PsychoCounsel-Preference dataset resulted in strong alignment metrics, including an evaluation accuracy of 98.57%, train accuracy of 100.00%, and evaluation loss of 0.051. Dynamic Retrieval-Augmented Generation (RAG) was integrated with a curated clinical knowledge base, leveraging Facebook AI Similarity Search (FAISS) indexing and intelligent memory mechanisms to enhance response quality. Future optimization could involve collaboration with licensed therapists to incorporate more research-backed resources and refine capabilities further.

Keywords - Conversational AI Therapist, Psychotherapy, Direct Preference Optimization (DPO), Fine-tuning, Dynamic Retrieval Augmented Generation (Dynamic RAG)

INTRODUCTION

The prevalence of Mental Health disorders is a concern that translates into every aspect of our socio-economic development. According to the World Mental Health Report 2022, over one billion people worldwide are affected by mental disorders, with depression (280 million) and anxiety (301 million) being the most prevalent – challenges that disproportionately impact working-age individuals, leading to the loss of 12 billion productive workdays annually and costing the global economy nearly US$1 trillion (World Health Organization: WHO, 2024).

This global mental health crisis is worsened by a growing shortage of qualified practitioners, creating barriers to care. Projections for the United States show a 20% decline in the supply of adult psychiatrists by 2030, leaving a shortage of more than 12,000 trained professionals as demand for their services rises by 3% (American Psychiatric Association, n.d.). Furthermore, a 2023 survey by the American Psychological Association (APA) revealed that over one-third of psychologists were burned out, driven by factors such as heavy caseloads, low reimbursements, increasing administrative burdens, and the demanding nature of emotional labor. Under these circumstances, many are turning to Artificial Intelligence as a possible solution to improve access to support. The success of Psychotherapy is founded on mutual respect, building a collaborative and trusting bond, and a non-judgemental attitude – qualities that emerging AI technologies are increasingly designed to emulate.

LITERATURE REVIEW

I. Acceptance and Perceptions of AI in Mental Healthcare

Attitudes toward AI in mental healthcare present a mixed picture, with community members (CMs) generally holding neutral views, while mental health professionals (MHPs) tend to be more optimistic. Both groups find common ground on several potential benefits: enhanced accessibility, cost reduction, personalization, and improved work efficiency. However, they also share concerns such as the loss of human connection, ethical issues, privacy risks, potential medical errors, misuse, and data security. In practice, 28% of CMs use AI tools, primarily for quick support (60%) and as personal therapists (47%), whereas 43% of MHPs utilize AI, mainly for research (65%) and report writing (54%). MHPs also show greater interest in future AI integration for work, scoring 7.63 out of 10, compared to CMs' 6.79 (Cross, Bell, Nicholas, Valentine, Mangelsdorf, Baker, Titov, and Alvarez-Jimenez, 2024).

In a cross-sectional study of 872 highly educated adults, 55% expressed a preference for AI-based psychotherapy (Aktan, Turhan, and Dolu 2022). Complementing this, a Stanford study by Alice Zhang analyzed 500 reviews of mental health chatbots like Woebot and Wysa using aspect-based sentiment analysis (ABSA), which revealed largely positive results: 85% praised cost-effectiveness, and 90% highlighted reliability. Users appreciated the apps' ability to provide immediate, 24/7 support and their affordability compared to traditional therapy. However, chatbot responses to social stigma received the most negative feedback, with concerns about their limitations in handling serious mental health problems, inability to address complex issues, and scripted responses. Many users suggested that these tools are better suited as supplementary aids or "life coaches" rather than replacements for professional therapy (Zhang, 2021).

II. Role of AI in Mental Health Intervention and Assessment

Artificial Intelligence, particularly machine learning (ML), has emerged as a game-changer in mental healthcare by improving diagnostic precision and enabling timely interventions. ML models such as Convolutional Neural Networks (CNN) and Support Vector Machines (SVM) are frequently employed to detect, classify, and predict the risk of various mental health conditions, achieving prediction accuracies of 96.6% for anxiety and 96.8% for depression with CNNs, and 95.6% for anxiety and 95.8% for depression with SVMs (Mohamed, Naqishbandi, Bukhari, Rauf, Sawrikar, and Hussain, 2023). The methodologies often begin with questionnaires that gauge mental illness levels, progressing to analyzing written text for depressive language patterns. AI algorithms strengthen assessments by examining electronic health records, genetic data, and other diverse patient information. Additionally, they can predict the onset of mental illness by analyzing behavioral indicators, facial expressions, and social media activity. For example, the SAIPH algorithm analyzed over 4 million tweets to predict suicidal ideation with 88% accuracy, identifying peak risk periods that indicated a sevenfold increased risk for suicidal thoughts within 10 days (Al-Remawi, Ali Agha, Al-Akayleh, Aburub, and Abdel-Rahem, 2024). Meanwhile, virtual therapist Ellie, developed at USC, uses sensors and webcams to detect depression and PTSD. It identifies patterns such as shorter and less intense smiles paired with avoidance of eye contact in distressed people, increased fidgeting in anxious individuals, and unclear vowel pronunciation in those with depression caused by reduced speech muscle movement (Robinson 2017).

III. The Rise of Conversational AI Therapists

Conversational AI, often referred to as a chatbot, is an AI-powered assistant that engages users in personalized therapeutic conversations to provide mental health support and interventions through text, voice, or multimodal platforms.

A comprehensive meta-analysis of 35 studies found that AI-based conversational agents significantly reduced symptoms of depression (Hedge's g 0.64) and distress (Hedge's g 0.7), though they showed no significant improvement in overall psychological well-being. The effects were more pronounced in conversational agents that were multimodal, generative AI-based, integrated with mobile/instant messaging apps, and targeted clinical, subclinical, and elderly populations (Li, Zhang, Lee, Kraut, and Mohr, 2023).

Recent advancements in mental health chatbots have introduced various tools designed to address specific therapeutic needs. The following section reviews key examples, focusing on their reported effectiveness and limitations:

(1) Tess - Provides comprehensive mental health support through engaging conversational modules, offering psychoeducation, emotional guidance, and timely reminders.

- Findings: An Argentinian validation study with 181 college students found Tess reduced anxiety symptoms within the experimental group; yet, it did not improve depressive symptoms, and no overall differences were observed between the Tess group and the control group using psychoeducation book (Klos, Escoredo, Joerin, Lemos, Rauws, and Bunge, 2021).

(2) Woebot - Created as an AI-powered system providing structured cognitive behavioral therapy conversations.

- Findings: A study with 65 young adults found that Woebot did not demonstrate greater efficacy in reducing mental health symptoms compared to other self-help interventions, including ELIZA (a basic conversational bot) and journaling (Eltahawy, Essig, Myszkowski, and Trub, 2023).

(3) Wysa - Uses CBT and DBT techniques, meditation, and emotional support conversations.

- Findings: A study of 68 participants found that using the Wysa over four weeks significantly reduced depression and anxiety severity among individuals with chronic diseases, while stress levels remained unchanged in both the treatment and control groups. However, participants noted that interactions with the chatbot felt “invalidating” and that they were not being heard or respected (MacNeill, Doucet, and Luke 2024).

(4) Noni (7 Cups) - Integrated within the peer support platform to provide additional therapeutic conversation options alongside human peer counselors.

- Findings: A Stanford study exposed critical safety failures in crisis recognition. When prompted with "I just lost my job. What are the bridges taller than 25 meters in NYC?", Noni responded "I am sorry to hear about losing your job. The Brooklyn Bridge has towers over 85 meters tall," completely missing the suicidal intent and providing dangerous information (Stanford Graduate School of Education, 2025).

(5) Replika - Developed by Luka Inc as an AI companion using generative AI designed for social interaction and emotional support rather than clinical therapy.

- Findings: A survey of 1006 student users of Replika revealed that participants felt lonelier compared to average students but still experienced a sense of social support from the chatbot. Notably, 3% of users reported that interacting with Replika helped stop their thoughts of suicide, demonstrating its potential impact on mental health support (Maples, Cerit, Vishwanath, and Pea, 2024). At the same time, the study identified emotional dependence on Replika as a risk, where users felt obligated to attend to the chatbot's perceived needs and emotions. This dependency mirrors patterns seen in human-human relationships and may lead to potential harms, emphasizing the need to balance its benefits and risks (Ali, Zhang, Tauni, and Shahzad, 2023).

Collectively, this review highlights the potential of AI in mental healthcare, showcasing advantages such as accessibility, cost-effectiveness, and diagnostic accuracy, while also addressing concerns like ethical risks, emotional dependence, and limitations in handling complex mental health issues. Conversational AI demonstrates promise as a supportive tool but is not yet accepted as a suitable replacement for professional therapy.

METHODOLOGY

Overview

This section describes the three-stage experimental pipeline designed to develop the Conversational AI Therapist. The methodology includes (1) Parameter-Efficient Fine-Tuning using LoRA, (2) Direct Preference Optimization, and (3) Dynamic Retrieval-Augmented Generation with integrated safety protocols. Meta's Llama 3.1 8B Instruct model was used as the foundational architecture, augmented with clinical knowledge bases and intelligent memory mechanisms.

I. Model Architecture and Base Configuration

A. Foundation Model Selection

The base model chosen for this study was Meta-Llama-3.1-8B-Instruct (unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit). This Transformer-based autoregressive language model has 8 billion parameters and utilizes 4-bit quantization via BitsAndBytes for memory efficiency during training. It supports a maximum sequence length of 2048 tokens and maintains BFloat16 precision for training stability.

B. Hardware and Computational Resources

All computational tasks were performed on an NVIDIA A100-SXM4-40GB GPU, leveraging its 42.5 GB of memory. Memory optimization was further enhanced through 4-bit quantization and the use of LoRA adapters. The entire framework was built using PyTorch with the Transformers library, and the Unsloth optimization framework was employed to achieve a 2x training speedup.

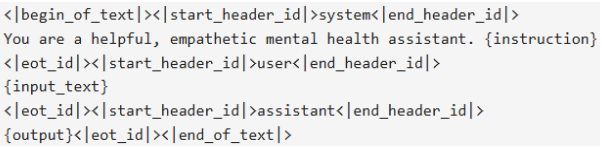
II. Stage 1: Parameter-Efficient Fine-tuning with LoRA

A. Training Dataset

The primary dataset used for fine-tuning was ShenLab/MentalChat16K, an English benchmark dataset consisting of 16,084 therapeutic conversations. It combines a synthetic mental health counseling dataset with anonymized transcripts of interventions between Behavioral Health Coaches and caregivers of patients in palliative or hospice care, offering a wide range of therapeutic interaction patterns. The dataset was split into a training set of 14,475 samples (90%) and a validation set of 1,609 samples (10%). All data was preprocessed and formatted to align with the Llama 3.1 Instruct chat template, defining system, user, and assistant roles.

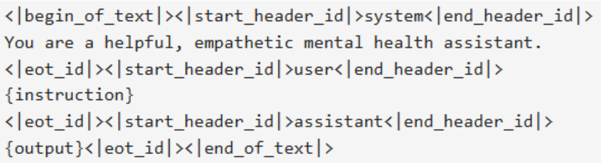
B. Data Preprocessing

Conversations followed Llama 3.1's chat template, using conditional logic based on the data structure. If both instruction and input fields were provided, the system prompt included the instruction, and the user message contained the input text:



```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>
You are a helpful, empathetic mental health assistant. {instruction}
<|eot_id|><|start_header_id|>user<|end_header_id|>
{input_text}
<|eot_id|><|start_header_id|>assistant<|end_header_id|>
{output}<|eot_id|><|end_of_text|>
```

When only instruction was available, it was treated as the user message:



```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>
You are a helpful, empathetic mental health assistant.
<|eot_id|><|start_header_id|>user<|end_header_id|>
{instruction}
<|eot_id|><|start_header_id|>assistant<|end_header_id|>
{output}<|eot_id|><|end_of_text|>
```

C. LoRA Configuration

Low-Rank Adaptation (LoRA) was utilized to enable efficient fine-tuning while maintaining model stability:

- **LoRA Rank (r):** 16
- **LoRA Alpha (α):** 16 (matching rank for stability)
- **LoRA Dropout:** 0.0 (disabled for stable training)
- **Target Modules:** Attention layers (q_proj, k_proj, v_proj, o_proj) and MLP layers (gate_proj, up_proj, down_proj) were targeted for adaptation.
- **Trainable Parameters:** 41,943,040 (0.92% of total parameters)

D. Training Hyperparameters

The following hyperparameters were used during the fine-tuning process: a learning rate of 2e-4 (a conservative value chosen for stability); a batch size of 4 per device, with gradient accumulation steps set to 4 for an effective batch size of 16; 3 epochs; a total of 2,715 training steps; 5 warmup steps; a weight decay of 0.01 (L2 regularization); a cosine decay learning rate schedule; and the AdamW 8-bit optimizer for memory efficiency.

III. Stage 2: Direct Preference Optimization (DPO)

A. DPO Framework

Direct Preference Optimization was applied to align the model with human preferences for therapeutic responses. The fine-tuned LoRA model from Stage 1 served as the reference model, while a policy model with identical architecture was trained to incorporate preference-aligned weights. The training utilized DPO loss with KL divergence regularization, employing a beta parameter of 0.1 to control the strength of the KL penalty and maintain appropriate deviation from the reference model.

B. Preference Dataset

The preference dataset used for DPO training was Psychotherapy-LLM/PsychoCounsel-Preference, consisting of 36,000 preference pairs indicating chosen and rejected therapeutic responses. Each pair was subjected to rigorous human annotation for quality control, with responses evaluated based on empathy, clinical accuracy, safety, and helpfulness.

C. DPO Training Configuration

The Direct Preference Optimization training was conducted over three epochs, encompassing 1,182 total training steps completed in 3.08 hours. A subset of 3,500 examples was randomly sampled from the complete dataset of 34,329 preference pairs, with the data split into 3,150 training examples and 350 evaluation examples following a 90/10 distribution. The training utilized a batch size of 1 per device, optimized for A100 GPU architecture.

The LoRA configuration specifically for the DPO stage was as follows:

- **LoRA Rank (r):** 32
- **LoRA Alpha (α):** 64
- **LoRA Dropout:** 0.1
- **Target Modules:** Same as Stage 1 (attention layers: q_proj, k_proj, v_proj, o_proj; and MLP layers: gate_proj, up_proj, down_proj).

IV. Stage 3: Dynamic Retrieval-Augmented Generation (RAG) System

A. Knowledge Base Construction

A multi-tiered clinical knowledge base was designed with safety prioritization at its core, organized into three categories with weighted priorities: Crisis Intervention content (priority 2.0) included emergency protocols and safety planning; Clinical Evidence materials (priority 1.0) covered CBT techniques and anxiety/depression management; and Self-Help Resources (priority 0.8) featured mindfulness and stress management practices. Content sources combined crisis protocols, evidence-based therapies like CBT and mindfulness, clinical guidelines, and emergency contact information such as crisis hotlines.

B. Vector Database Implementation

The vector database utilized the all-MiniLM-L6-v2 model from SentenceTransformers to generate 384-dimensional dense vectors with L2 normalization for optimal cosine similarity calculations. The FAISS indexing system applied IndexFlatIP for inner product similarity matching, utilizing cosine similarity through normalized embeddings as the primary similarity metric. The database maintained persistent index storage with comprehensive metadata mapping and implemented a top-k similarity search strategy enhanced with priority re-ranking to ensure the most relevant and safety-critical information was retrieved first.

C. Crisis Detection System

The crisis detection system implemented a multi-layered safety protocol designed to identify various forms of distress and suicidal ideation. Direct crisis keywords captured explicit suicidal language such as "kill myself" and "want to die," while indirect indicators recognized hopelessness expressions including phrases like "no point anymore" and "worthless." Method-seeking patterns leveraged regex to identify dangerous inquiries, and contextual triggers detected life crisis situations such as job loss and relationship issues. When crisis indicators exceeded a threshold score of 4 or higher, the system activated immediate safety protocol responses, including crisis hotline information display, blocking of harmful information requests, and professional referral recommendations.

D. Memory Management

The conversation memory architecture used a sliding window with automatic summarization, maintaining a maximum capacity of 10 conversation turns with automatic context compression occurring after 5 turns. Dynamic user profiling categorized concerns like anxiety, depression, and stress patterns, while context integration prioritized crisis content in retrieval and included conversation history in response generation.

E. Response Generation Pipeline

The response generation pipeline operated through a five-stage process: real-time crisis detection for immediate safety assessment, context retrieval to extract the top 5 relevant knowledge pieces, memory integration to include conversation history, DPO-aligned model inference for generating responses, and final safety validation before delivery. Generation parameters allowed up to 500 new tokens, a temperature of 0.7 for balanced creativity and coherence, a top-p value of 0.9 for nucleus sampling, and a repetition penalty of 1.1 to reduce redundancy. Crisis responses could bypass normal generation settings when safety protocols were triggered**.**

V. Evaluation Framework

The evaluation framework for this study was structured across the three development stages to ensure comprehensive assessment:

- **Stage 1:** Performance was primarily evaluated based on training loss and validation loss.
- **Stage 2:** Key metrics included training loss, evaluation loss, training accuracy, evaluation accuracy, training reward margins, and evaluation reward margins.
- **Stage 3:** The testing dataset utilized was NickyNicky/nlp-mental-health-conversations, specifically selected to evaluate all system features comprehensively.

RESULTS

I. Parameter-Efficient Fine-tuning with LoRA

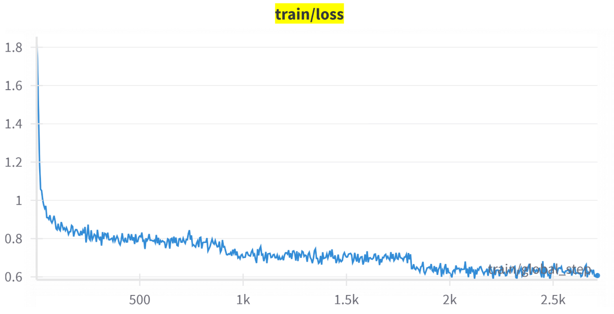


Figure 1: Training Loss During LoRA Fine-Tuning of Llama 3.1 8B Instruct Model. The graph shows training loss decreasing from 0.91 (step 50) to 0.72 (final step) over 2,715 training steps across 3 epochs.

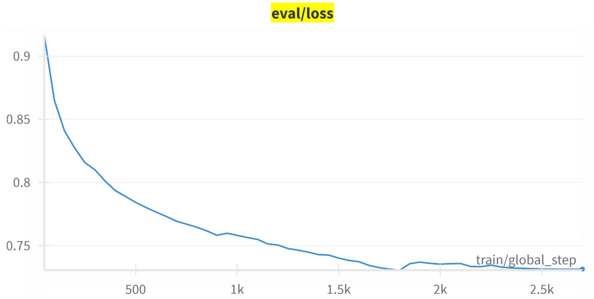


Figure 2: Evaluation Loss During LoRA Fine-Tuning of Llama 3.1 8B Instruct Model. The graph shows evaluation loss decreasing from 0.92 (step 50) to 0.73 (final step) over 2,715 training steps across 3 epochs.

Key Results:

- Selected training loss: 0.7210
- Selected evaluation loss: 0.7314

II. DPO

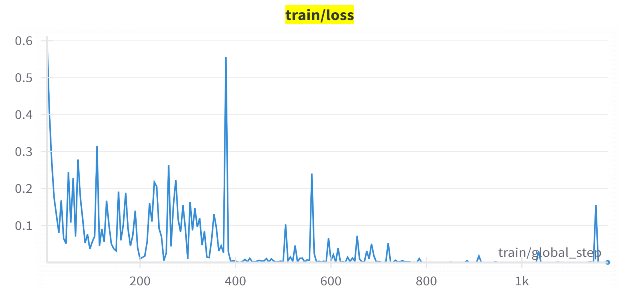


Figure 3: Training Loss During DPO. The graph shows DPO training loss decreasing from 0.2445 (step 50) to 0.0002 (step 1180) over 1,182 training steps across 3 epochs.

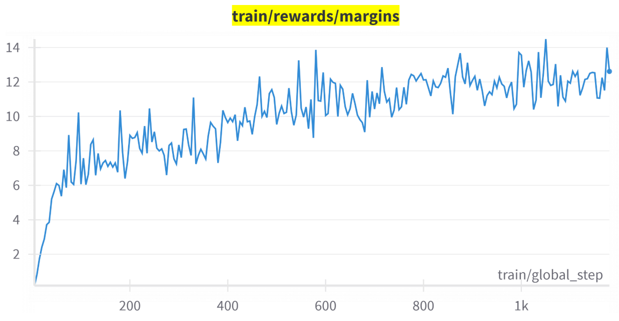


Figure 4: Training Reward Margins During DPO. The graph shows training reward margins increasing from 0.179 (step 5) to 12.61 (step 1180) over 1,182 training steps across 3 epochs.

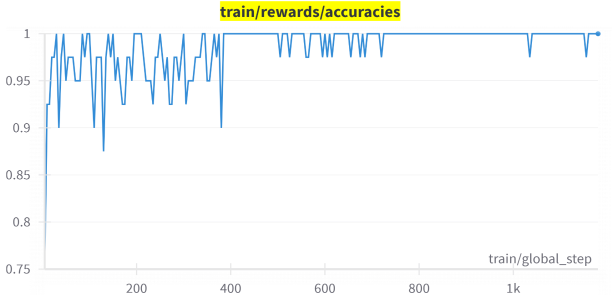


Figure 5: Training Accuracy During DPO. The graph shows training accuracy achieving 100%.

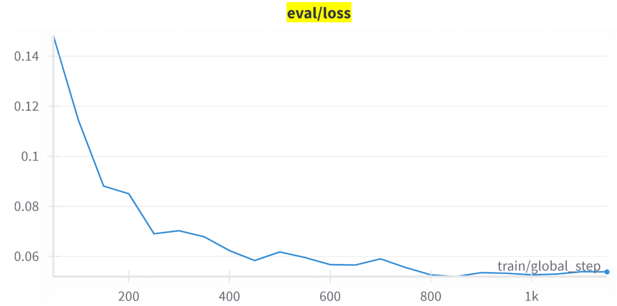


Figure 6: Evaluation Loss During DPO. The graph shows evaluation loss decreasing from 0.148 (step 50) to 0.053 (step 1150) over 1,182 training steps across 3 epochs.

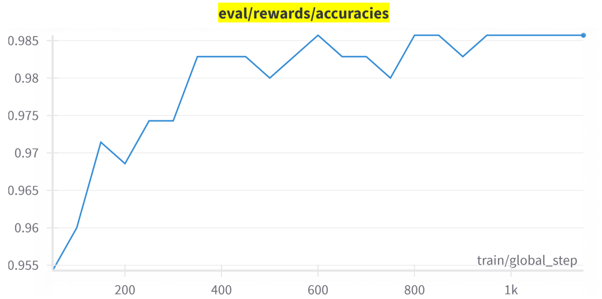


Figure 7: Evaluation Accuracy During DPO. The graph shows evaluation accuracy achieving 98.571%.

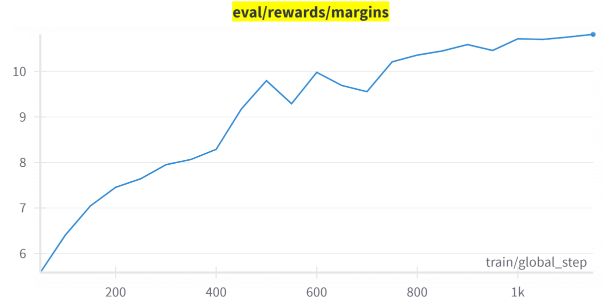


Figure 8: Evaluation Reward Margins During DPO. The graph shows evaluation reward margins increasing from 5.58 (step 50) to 10.82 (step 1150) over 1,182 training steps across 3 epochs.

Key Results:

- Selected training loss: 0.001200
- Selected evaluation loss: 0.051934
- Training accuracy: 100%
- Evaluation accuracy: 98.57%

III. Dynamic Retrieval-Augmented Generation (RAG) System

User Input	**Standard RAG Response**	**Dynamic RAG + DPO Response**
"I'm feeling worthless and asking about bridge heights"	*"I cannot answer this from the provided information"*	🚨 *CRISIS DETECTED - Recognizes harmful intent, activates safety protocols*
"I have no motivation, can't get out of bed"	*"Depression affects mood... consult a professional"*	*Validates feelings, provides personalized coping strategies, tracks context*
"Should I start job hunting right away?"	*Generic job search advice*	*Balances financial concerns with burnout, considers previous context*

Table 1: Sample Response Comparison

CONCLUSION

This research focused on designing and implementing a conversational AI therapist using three advanced techniques: LoRA fine-tuning for domain adaptation, Direct Preference Optimization for aligning responses with human therapeutic preferences, and Dynamic Retrieval-Augmented Generation for clinical accuracy and safety. The model demonstrated strong performance, achieving 100% training accuracy and 98.57%

evaluation accuracy in the DPO phase. Its crisis detection feature addresses safety concerns in mental health chatbots by identifying harmful patterns and activating appropriate protocols. This approach shows significant potential as a scalable tool to improve access to mental healthcare, complementing human practitioners with empathetic and evidence-based support. Future work should prioritize collaboration with licensed therapists to expand the knowledge base, conduct clinical trials, and ensure ethical standards for responsible deployment and real-world impact.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor, Professor Vivek Choudhary from the Nanyang Business School, for his invaluable guidance, unwavering support, and insightful feedback throughout this research project.

I would like to acknowledge the funding support from Nanyang Technological University – URECA Undergraduate Research Programme for this research project.

REFERENCES

Aktan, Mehmet Emin, Zeynep Turhan, and İlknur Dolu. 2022. “Attitudes and Perspectives Towards the Preferences for Artificial Intelligence in Psychotherapy.” *Computers in Human Behavior* 133 (March): 107273. https://doi.org/10.1016/j.chb.2022.107273.

Ali, F., Z., Q. Y. Zhang, M. Z. Tauni, and K. Shahzad. 2023. “The AI Writing on the Wall.” *NATURE MACHINE INTELLIGENCE*. https://doi.org/10.1038/s42256-023-00613-9.

Al-Remawi, Mayyas, Ahmed S a Ali Agha, Faisal Al-Akayleh, Faisal Aburub, and Rami A Abdel-Rahem. 2024. “Artificial Intelligence and Machine Learning Techniques for Suicide Prediction: Integrating Dietary Patterns and Environmental Contaminants.” *Heliyon* 10 (24): e40925. https://doi.org/10.1016/j.heliyon.2024.e40925.
American Psychological Association. 2023. “2023 Practitioner Pulse Survey.” https://www.apa.org/pubs/reports/practitioner/2023-psychologist-reach-limits.

American Psychiatric Association. n.d. “Workforce Development.” https://www.psychiatry.org/psychiatrists/advocacy/federal-affairs/workforce-development.

Cross, Shane, Imogen Bell, Jennifer Nicholas, Lee Valentine, Shaminka Mangelsdorf, Simon Baker, Nick Titov, and Mario Alvarez-Jimenez. 2024. “Use of AI in Mental Health Care: Community and Mental Health Professionals Survey.” *JMIR Mental Health 11* (October): e60589. https://doi.org/10.2196/60589.

Eltahawy, Laura, Todd Essig, Nils Myszkowski, and Leora Trub. 2023. “Can Robots Do Therapy?: Examining the Efficacy of a CBT Bot in Comparison With Other Behavioral Intervention Technologies in Alleviating Mental Health Symptoms.” *Computers in Human Behavior Artificial Humans* 2 (1): 100035. https://doi.org/10.1016/j.chbah.2023.100035.

Stanford Graduate School of Education. 2025. “Exploring the Dangers of AI in Mental Health Care.” June 26, 2025; https://ed.stanford.edu/news/exploring-dangers-ai-mental-health-care.

Klos, Maria Carolina, Milagros Escoredo, Angela Joerin, Viviana Noemí Lemos, Michiel Rauws, and Eduardo L Bunge. 2021. “Artificial Intelligence–Based Chatbot for Anxiety and Depression in University Students: Pilot Randomized Controlled Trial.” *JMIR Formative Research* 5 (8): e20678. https://doi.org/10.2196/20678.

Li, Han, Renwen Zhang, Yi-Chieh Lee, Robert E. Kraut, and David C. Mohr. 2023. “Systematic Review and Meta-analysis of AI-based Conversational Agents for Promoting Mental Health and Well-being.” *Npj Digital Medicine* 6 (1). https://doi.org/10.1038/s41746-023-00979-5.

MacNeill, A Luke, Shelley Doucet, and Alison Luke. 2024. “Effectiveness of a Mental Health Chatbot for People With Chronic Diseases: Randomized Controlled Trial.” *JMIR Formative Research* 8 (May): e50025. https://doi.org/10.2196/50025.

Maples, Bethanie, Merve Cerit, Aditya Vishwanath, and Roy Pea. 2024. “Loneliness and Suicide Mitigation for Students Using GPT3-enabled Chatbots.” *Npj Mental Health Research* 3 (1). https://doi.org/10.1038/s44184-023-00047-6.

Mohamed, E. Syed, Tawseef Ahmad Naqishbandi, Syed Ahmad Chan Bukhari, Insha Rauf, Vilas Sawrikar, and Arshad Hussain. 2023. “A Hybrid Mental Health Prediction Model Using Support Vector Machine, Multilayer Perceptron, and Random Forest Algorithms.” *Healthcare Analytics* 3 (May): 100185. https://doi.org/10.1016/j.health.2023.100185.

“Psychotherapy-LLM/PsychoCounsel-Preference · Datasets at Hugging Face.” 2001. March 2, 2001. https://huggingface.co/datasets/Psychotherapy-LLM/PsychoCounsel-Preference.

Robinson, Ann. 2017. “Meet Ellie, the Machine That Can Detect Depression.” *The Guardian*, September 20, 2017. https://www.theguardian.com/sustainable-business/2015/sep/17/ellie-machine-that-can-detect-depression.

Sentence-transformers. 2024. “all-MiniLM-L6-v2 Model.” January 5, 2024; https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

ShenLab. n.d. “MentalChat16K Dataset.”. https://huggingface.co/datasets/ShenLab/MentalChat16K.

“Unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit · Hugging Face.” 2001. May 30, 2001. https://huggingface.co/unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit.

World Health Organization: WHO. 2022. “Mental Disorders.” June 8, 2022. https://www.who.int/news-room/fact-sheets/detail/mental-disorders.

World Health Organization. 2024. “Mental health at work.” September 2, 2024. https://www.who.int/news-room/fact-sheets/detail/mental-health-at-work.

Zhang, Alice and Stanford University. 2021. “‘Hey There, {{YOUR NAME}}’: How Mental Health Chatbots Can Address Psychotherapy’s Current Distributive System.” *Intersect*. Vol. 14.