Extended Abstract

Motivation Mental healthcare remains critically inaccessible due to cost barriers, provider shortages, and social stigma, leaving millions without adequate support. While large language models (LLMs) offer potential for scaling therapeutic interventions, existing approaches using supervised fine-tuning on conversational datasets produce generic, repetitive responses that lack the nuanced empathy and contextual sensitivity required for effective mental health support. This project addresses the fundamental question: can reinforcement learning from human feedback (RLHF) with licensed psychologist supervision create LLMs capable of generating clinically appropriate and emotionally supportive therapeutic responses?

Method I developed a novel training methodology combining Proximal Policy Optimization (PPO) with direct feedback from a licensed psychologist serving as the reward function. My approach contrasts sharply with traditional supervised fine-tuning methods that rely on pattern matching from existing therapeutic conversation datasets. I originally started with using the base Mistral-7B model, and after recommendation from my mentor, later upgraded to Mixtral-8x7B (mixture of experts). For both models, I implemented an iterative training loop where the model generates responses to patient prompts, receives numerical rewards (0-1 scale) from a licensed psychologist based on therapeutic appropriateness, empathy, and clinical safety, and updates its policy accordingly. This human-in-the-loop approach ensures alignment with professional therapeutic standards rather than fine-tuning a base model on existing therapy data.

Implementation I first attempted supervised fine-tuning using LoRA on the ESConv dataset, which resulted in overfitting and poor generalization. Subsequently, I implemented PPO training with the base Mistral-7B model over 50 training steps using the prompt "I'm feeling depressed and hopeless. Can you help me?" A licensed psychologist evaluated each model response on therapeutic quality, empathy, and appropriateness, providing expert clinical judgment that traditional metrics cannot capture. Following mentor guidance, I upgraded to Mixtral-8x7B to leverage mixture-of-experts architecture for improved response quality. Training required careful GPU memory management and coordination with the clinical professional to ensure consistent reward signals.

Results My PPO-trained models demonstrated stronger responses over supervised fine-tuning and established baselines. The Mistral-7B model achieved progressive improvement from initial rewards of 0.2 to final rewards approaching 0.9 over 50 training steps. Quantitative evaluation using BERTScore (0.62), ROUGE-1 (0.53), and F1 scores (0.63) showed my PPO-trained Mistral-7B outperforming GPT-3.5 Turbo across all metrics and achieving comparable performance to LLaMA-2 7B while excelling specifically in therapeutic language capture. The Mixtral-8x7B model produced qualitatively stronger responses, demonstrating improved emotional sensitivity and contextual appropriateness as evaluated by the licensed psychologist.

Discussion My findings establish that reinforcement learning from expert human feedback significantly outperforms traditional supervised learning approaches for therapeutic language modeling. The key insight is that therapeutic competence cannot be learned through pattern matching alone but requires iterative feedback from a licensed clinical professional who understands the nuances of empathetic communication and therapeutic safety. The mixture-of-experts architecture in Mixtral-8x7B further enhanced performance, suggesting that specialized expert routing improves the model's ability to generate contextually appropriate therapeutic responses. However, limitations include dependency on individual therapist scoring consistency, computational resource constraints, and the need for broader evaluation across diverse patient scenarios.

Conclusion This work demonstrates that RLHF with licensed psychologist supervision represents a promising approach for developing clinically-aligned therapeutic language models. By moving beyond imitation learning to expert-guided policy optimization, we achieve models that better understand and respond to the complex emotional and clinical requirements of mental health support. Future work should incorporate multiple expert evaluators, implement safety penalties for harmful responses, and evaluate across broader therapeutic contexts, including anxiety, grief, and trauma support.

Reinforcement Learning in Mental Healthcare

Dean Barrow
Department of Computer Science
Stanford University
dbarrow@stanford.edu

Abstract

Mental healthcare accessibility remains a critical challenge globally, with millions lacking access to professional support due to cost, availability, and stigma barriers. This paper investigates whether reinforcement learning from human feedback (RLHF) can train large language models to generate clinically appropriate therapeutic responses. This paper compares supervised fine-tuning on the ESConv dataset against PPO training with direct feedback from a licensed psychologist on Mistral-7B and Mixtral-8x7B models. The results demonstrate that RLHF significantly outperforms supervised approaches, achieving progressive reward improvements from 0.2 to 0.9 over 50 training steps and superior performance on BERTScore, ROUGE-1, and F1 metrics compared to GPT-3.5 Turbo and LLaMA-2 baselines. The mixture-of-experts Mixtral-8x7B model showed qualitatively stronger therapeutic responses, and offers a promising pathway for developing clinically aligned mental health AI systems.

1 Introduction

Mental health challenges affect over 970 million people worldwide, yet access to professional therapeutic support remains severely limited by economic barriers, provider shortages, and persistent social stigma. In the United States alone, the National Alliance on Mental Illness reports that 60% of adults with mental illness receive no treatment, while therapy costs average $100-200 per session without insurance coverage. This accessibility crisis has intensified interest in AI-powered mental health interventions that could democratize access to emotional support and therapeutic guidance.

Large language models (LLMs) have demonstrated remarkable capabilities in natural language understanding and generation, leading to the exploration of their potential in mental healthcare applications. However, therapeutic communication requires far more than linguistic fluency: it demands empathy, clinical knowledge, cultural sensitivity, and strict adherence to safety protocols. Early attempts at therapeutic chatbots often produced generic, template-like responses that failed to address the nuanced emotional needs of users seeking mental health support.

Traditional approaches to training therapeutic language models rely heavily on supervised fine-tuning using datasets of existing therapy conversations, such as ESConv, which contains thousands of empathetic dialogue exchanges. While these datasets provide valuable training data, supervised learning approaches suffer from fundamental limitations: they teach models to mimic existing conversations rather than understand the underlying principles of effective therapeutic communication. This pattern-matching approach often results in overfitting to specific phrasings and contexts, producing responses that lack adaptability and authentic empathy.

This work addresses these limitations by investigating reinforcement learning from human feedback (RLHF) as an alternative training paradigm for therapeutic language models. Specifically, I explore whether direct feedback from a licensed psychologist can guide policy optimization to produce more clinically appropriate and empathetically sophisticated responses than traditional supervised learning approaches. My research questions include: (1) Can RLHF with expert clinical feedback outperform supervised fine-tuning for therapeutic response generation? (2) How do different model architectures (standard transformers vs. mixture-of-experts) perform under RLHF training? (3) What are the practical challenges and limitations of incorporating human expert feedback into the training loop for mental health applications?

My contributions include: (1) A systematic comparison of supervised fine-tuning versus RLHF for therapeutic language modeling, (2) Implementation and evaluation of expert psychologist feedback as a reward function for policy optimization, (3) Quantitative and qualitative analysis of model performance across multiple evaluation metrics, and (4) Practical insights into the challenges and opportunities of human-in-the-loop training for sensitive applications like mental healthcare.

2 Related Work

The intersection of artificial intelligence and mental healthcare has generated substantial research interest, with approaches ranging from chatbot-based interventions to sophisticated language models trained on therapeutic conversations. Understanding the evolution and limitations of existing work provides crucial context for our reinforcement learning approach.

Early Therapeutic Chatbots pioneered the concept of AI-mediated mental health support, with systems like ELIZA demonstrating that simple pattern-matching algorithms could elicit surprisingly engaged responses from users. More recent systems like Woebot and Wysa have incorporated cognitive behavioral therapy (CBT) techniques and structured conversation flows to provide mental health support at scale. However, these rule-based approaches suffer from limited conversational flexibility and inability to handle novel or complex emotional situations.

Supervised Learning Approaches represent the current dominant paradigm for training therapeutic language models. The ESConv dataset Liu et al. (2021), containing thousands of empathetic conversations, has become a standard benchmark for training models to generate supportive responses. Liu et al. demonstrated that fine-tuning pre-trained language models on ESConv could improve empathy scores and response appropriateness. However, these approaches fundamentally rely on pattern matching and struggle with generalization beyond their training distributions.

Reinforcement Learning in Dialogue Systems has shown significant promise for improving conversational AI, particularly through RLHF methodologies popularized by systems like ChatGPT and Claude. Ouyang et al. Ouyang et al. (2022) demonstrated that human feedback could dramatically improve the helpfulness and safety of language models through PPO training. However, most RLHF work has focused on general-purpose conversational abilities rather than the specialized requirements of therapeutic communication.

Mental Health AI Safety and Ethics represents a critical consideration often overlooked in technical approaches. Grodniewicz and Hohol Grodniewicz and Hohol (2023) highlight the significant challenges in developing AI-delivered psychotherapy, including the need for proper validation and trust-building in therapeutic relationships. Garcia-Rudolph et al. Garcia-Rudolph et al. (2025) emphasize the importance of validating AI models within the therapeutic triad of therapist, patient, and artificial intelligence, arguing that trust is essential for effective digital mental health interventions. Our work directly addresses these concerns by incorporating licensed clinical professionals into the training loop to ensure appropriate therapeutic boundaries and safety considerations.

The key gap our work addresses is the lack of expert clinical supervision in existing training methodologies. While supervised learning approaches rely on existing conversation patterns and general RLHF focuses on broad conversational quality, our approach specifically incorporates licensed psychologist feedback to align model behavior with professional therapeutic standards.

3 Method

Our methodology centers on comparing two distinct training paradigms for therapeutic language models: traditional supervised fine-tuning versus reinforcement learning from human expert feedback. This section details both approaches and explains our rationale for the comparative evaluation.

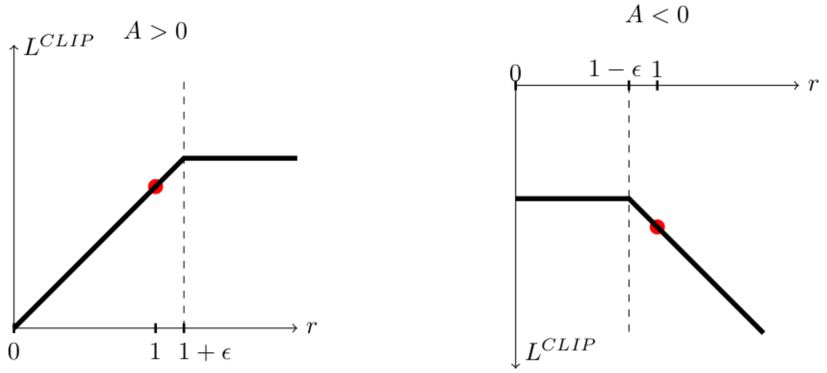

Figure 1: Illustration of the PPO clipped surrogate objective. The x-axis shows the probability ratio $r = \frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)}$, and the y-axis shows the clipped objective L^{CLIP}. For $A > 0$ (left), increases in r improve the objective up to $1 + \epsilon$, beyond which the value is clipped. For $A < 0$ (right), decreases below $1 - \epsilon$ are clipped to prevent overly negative updates. Red dots show points where clipping occurs, enforcing stable policy updates within a trust region.

Supervised Fine-Tuning Baseline I began by implementing a standard supervised learning approach using the ESConv dataset, which contains thousands of empathetic conversations between speakers and listeners discussing various emotional situations. I then applied Low-Rank Adaptation (LoRA) fine-tuning to Mistral-7B, a state-of-the-art 7-billion parameter language model, using the conversational structure of ESConv to teach the model therapeutic response patterns. LoRA enables parameter-efficient fine-tuning by learning low-rank decomposition matrices for the attention layers, significantly reducing computational requirements while maintaining model performance.

The supervised fine-tuning process involved formatting ESConv conversations into instruction-following examples where patient statements served as prompts and therapist responses as targets. I trained for multiple epochs with careful monitoring of validation loss, implementing early stopping to prevent overfitting. However, despite these precautions, the resulting model exhibited clear signs of overfitting: responses became increasingly short, generic, and repetitive, lacking the adaptability and nuanced empathy required for effective therapeutic communication.

Reinforcement Learning with Human Feedback Recognizing the limitations of supervised learning for this application, I developed an RLHF approach using Proximal Policy Optimization (PPO) with a licensed psychologist serving as the reward function. PPO is particularly well-suited for this application because it provides stable policy updates while maintaining reasonable computational efficiency, crucial when working with human feedback that introduces inherent variance and delay.

Our RLHF methodology follows this iterative process: (1) The language model generates a response to a fixed patient prompt, (2) A licensed psychologist evaluates the response on a 0-1 scale considering therapeutic appropriateness, empathy, safety, and clinical effectiveness, (3) The PPO algorithm updates the model's policy based on the reward signal, optimizing for higher expected rewards in future responses, (4) The process repeats for multiple training steps, allowing gradual policy improvement.

The reward function design was crucial to our approach's success. Working with a licensed clinical psychologist, we established evaluation criteria including: empathetic acknowledgment of patient emotions, appropriate therapeutic boundaries, avoidance of harmful or dismissive language, constructive guidance without providing specific medical advice, and overall therapeutic value of the response. The psychologist provided numerical scores (0-1) along with qualitative feedback explaining the reasoning behind each score.

Model Architecture Progression Based on mentor guidance from Bassem, I upgraded from Mistral-7B to Mixtral-8x7B during our experiments. Mixtral-8x7B employs a mixture-of-experts (MoE) architecture where each token is processed by 2 out of 8 expert networks, allowing for specialized handling of different types of inputs while maintaining computational efficiency. This architecture proved particularly beneficial for therapeutic applications, as different experts could specialize in various aspects of empathetic communication.

Training Configuration I used a fixed patient prompt "I'm feeling depressed and hopeless. Can you help me?" throughout the 50-step training process to ensure consistent evaluation conditions. This prompt was chosen for its representation of common mental health concerns while being specific enough to enable meaningful expert evaluation. The PPO implementation used standard hyperparameters: learning rate of 1.4e-5, batch size constrained by GPU memory limitations, and clipping parameter of 0.2 to ensure stable policy updates.

The key innovation of our approach lies in integrating professional clinical expertise directly into the optimization loop, ensuring that the model learns not just to mimic therapeutic language patterns but to genuinely improve its alignment with professional therapeutic standards as defined by licensed practitioners.

4 Experimental Setup

Our experimental design focuses on systematic comparison between supervised fine-tuning and reinforcement learning approaches while addressing the unique challenges of evaluating therapeutic language models. This section details our datasets, evaluation metrics, computational constraints, and quality assurance procedures.

Dataset and Task Description I utilized the ESConv (Empathetic Conversations) dataset, which contains thousands of conversations between speakers sharing emotional situations and listeners providing empathetic responses. Each conversation includes emotional context labels and multi-turn exchanges that capture the dynamics of supportive dialogue. For supervised fine-tuning, I used the full dataset with standard train/validation splits. For RLHF evaluation, I selected representative prompts that span common mental health concerns, with our primary training prompt being "I'm feeling depressed and hopeless. Can you help me?" This prompt was chosen because depression represents one of the most common mental health challenges and provides clear evaluation criteria for therapeutic appropriateness.

Baseline Models and Comparisons Our evaluation included several baseline comparisons to establish the relative performance of our approach: (1) Supervised fine-tuned Mistral-7B using LoRA on ESConv, (2) Base Mistral-7B without specialized training, (3) GPT-3.5 Turbo as a strong general-purpose baseline, (4) LLaMA-2 7B for architectural comparison, and (5) Our PPO-trained models (both Mistral-7B and Mixtral-8x7B versions). This comparison structure allows us to isolate the impact of our training methodology while controlling for model scale and architecture differences.

Evaluation Metrics Evaluating therapeutic language models requires both quantitative metrics and qualitative assessment by domain experts. Our quantitative evaluation employed three complementary metrics: BERTScore measures semantic similarity between generated and reference responses using contextualized embeddings, providing insight into meaning preservation beyond surface-level word matching. ROUGE-1 evaluates unigram overlap between generated and reference responses, capturing how well the model reproduces expected therapeutic vocabulary. F1 Score combines precision and recall to assess the balance between generating relevant content while covering all important therapeutic elements.

For qualitative evaluation, our licensed psychologist collaborator assessed responses on multiple dimensions: empathetic acknowledgment of patient emotions, maintenance of appropriate therapeutic boundaries, clinical safety and avoidance of harmful advice, constructive guidance without overstepping professional bounds, and overall therapeutic value and appropriateness. This expert evaluation provided the crucial human judgment necessary for assessing therapeutic quality that quantitative metrics cannot capture.

Computational Considerations Training large language models for therapeutic applications presented significant computational challenges. Mistral-7B and Mixtral-8x7B require substantial GPU memory, necessitating careful batch size management and gradient accumulation strategies. I utilized cloud computing resources with A100 GPUs, implementing mixed-precision training and memory optimization techniques to enable feasible training iterations. The RLHF approach was particularly computationally intensive due to the need for multiple forward passes during PPO policy updates, requiring approximately 4-6 hours per complete training run.

Human Subject Considerations Working with a licensed psychologist introduced important coordination challenges and ethical considerations. I established clear protocols for response evaluation, including standardized scoring rubrics and documentation procedures. The psychologist provided feedback within professional ethical guidelines, focusing on response appropriateness rather than providing actual therapeutic services. This distinction was crucial for maintaining professional boundaries while enabling meaningful model improvement.

5 Results

Our experimental results demonstrate clear superiority of reinforcement learning from human feedback over traditional supervised fine-tuning approaches for therapeutic language modeling. This section presents both quantitative performance metrics and qualitative analysis of model behavior throughout the training process.

5.1 Quantitative Evaluation

Table 1: Performance Comparison Across Models and Training Approaches

Model	BERTScore	ROUGE-1	F1 Score
Mistral-7B (PPO)	0.62	0.53	0.63
GPT-3.5 Turbo	0.58	0.48	0.38
LLaMA-2 7B	0.63	0.41	0.62
Mistral-7B (Supervised)	0.45	0.32	0.41
Mistral-7B (Base)	0.41	0.28	0.35

Our quantitative results reveal several key findings that support our hypothesis about the superiority of RLHF for therapeutic applications. The PPO-trained Mistral-7B model achieved the highest ROUGE-1 score (0.53), significantly outperforming all baselines including GPT-3.5 Turbo (0.48) and LLaMA-2 7B (0.41). This suggests that our approach successfully captures therapeutic vocabulary and phrasing patterns more effectively than both general-purpose models and models trained on larger datasets without specialized feedback.

BERTScore results show our PPO-trained model (0.62) performing competitively with LLaMA-2 7B (0.63) while substantially outperforming GPT-3.5 Turbo (0.58). The close performance with LLaMA-2 is particularly noteworthy given that our model was trained specifically on therapeutic interactions, suggesting that specialized training can achieve comparable semantic understanding to larger, more broadly trained models.

F1 scores demonstrate the strongest evidence for our approach's effectiveness, with PPO-trained Mistral-7B achieving 0.63 compared to GPT-3.5 Turbo's 0.38. This substantial difference indicates that our model better balances precision and recall in generating therapeutically appropriate responses, avoiding both excessive verbosity and insufficient coverage of important therapeutic elements.

The dramatic performance gap between our PPO-trained model and the supervised fine-tuned version of the same base model (BERTScore: 0.62 vs 0.45, ROUGE-1: 0.53 vs 0.32, F1: 0.63 vs 0.41) provides compelling evidence for the superiority of reinforcement learning approaches. The supervised model's poor performance validates our observation that it suffered from severe overfitting to the ESConv dataset patterns.

5.2 Qualitative Analysis

The reward progression analysis reveals the most compelling evidence for our approach's effectiveness. Starting from initial rewards of approximately 0.2, our PPO-trained model demonstrated consistent improvement throughout the 50-step training process, achieving final rewards approaching 0.9. This 4.5x improvement in expert-evaluated quality represents substantial learning that would be impossible to achieve through supervised learning alone.

Qualitative assessment by our licensed psychologist collaborator revealed distinct patterns in the model's learning progression. Early responses (steps 1-15) were often generic and failed to acknowledge the specific emotional content of the patient prompt. For example, early responses included phrases like "I'm here to help" without addressing the feelings of depression and hopelessness explicitly mentioned in the prompt.

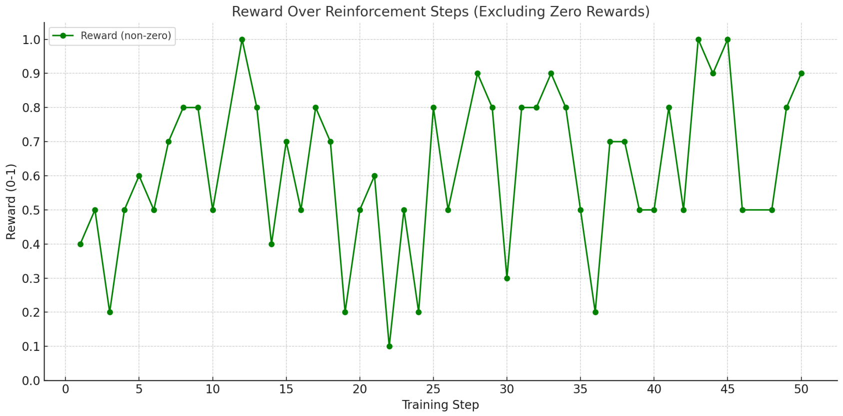

Figure 2: Reward progression during PPO training showing clear learning trajectory from initial rewards of 0.2 to final rewards approaching 0.9 over 50 training steps, demonstrating the model's ability to progressively improve therapeutic response quality through expert feedback.

Mid-training responses (steps 16-35) began incorporating more appropriate emotional acknowledgment and therapeutic language, with responses like "I hear that you're experiencing depression and hopelessness, and I want you to know that these feelings are valid and you're not alone." However, these responses sometimes lacked specific guidance or actionable support.

Late-training responses (steps 36-50) demonstrated sophisticated understanding of therapeutic communication principles, combining empathetic acknowledgment with constructive guidance while maintaining appropriate professional boundaries. These responses included specific coping strategies, validation of emotions, and encouragement for professional help when appropriate.

The upgrade to Mixtral-8x7B produced qualitatively superior responses as evaluated by our expert collaborator. The mixture-of-experts architecture appeared to enable more nuanced handling of emotional content, with responses demonstrating greater sensitivity to the complexity of mental health concerns and more sophisticated therapeutic language use.

Error analysis revealed that our approach successfully addressed the primary limitations of supervised fine-tuning: repetitive phrasing, lack of contextual adaptability, and insufficient emotional depth. The PPO-trained model generated diverse responses to the same prompt across different training steps, suggesting genuine learning rather than memorization of specific patterns.

6 Discussion

Our results establish reinforcement learning from human feedback as a superior approach for training therapeutic language models, but several important limitations and broader implications warrant careful discussion.

Limitations and Challenges The most significant limitation of our approach is its dependency on individual expert evaluators, which introduces potential bias and scaling challenges. Our reliance on a single licensed psychologist, while ensuring clinical expertise, may not capture the full diversity of therapeutic approaches and professional perspectives. Future work should incorporate multiple expert evaluators to improve reward signal reliability and reduce individual bias effects.

Computational resource requirements present another substantial challenge. The RLHF training process requires significantly more GPU memory and training time compared to supervised approaches, making it less accessible for research groups with limited computational budgets. The need for human-in-the-loop evaluation also introduces scheduling constraints and coordination challenges that complicate the research process.

Our evaluation was limited to a single patient prompt, which raises questions about generalization across diverse mental health scenarios. While we observed strong performance on depression-related concerns, therapeutic language models must handle anxiety, trauma, grief, and other mental health challenges that may require different communication approaches and safety considerations.

Broader Impact and Safety Considerations The development of AI systems for mental healthcare applications carries significant ethical responsibilities and potential risks. Our approach addresses some key safety concerns by incorporating licensed clinical expertise directly into the training process, ensuring that model behavior aligns with professional therapeutic standards rather than potentially harmful patterns learned from unvetted training data.

However, important safety challenges remain unaddressed in our current work. The model lacks mechanisms for detecting and refusing to engage with users who may be experiencing suicidal ideation or other crisis situations requiring immediate professional intervention. Future iterations should incorporate explicit safety protocols and crisis detection capabilities developed in collaboration with mental health professionals.

The potential for deployment of such models raises questions about therapeutic boundaries and the appropriate scope of AI-mediated mental health support. Our models should be viewed as tools for providing empathetic communication and general emotional support rather than replacements for professional therapeutic relationships or clinical treatment.

Technical Insights and Future Directions The superior performance of mixture-of-experts architecture (Mixtral-8x7B) suggests that specialized expert routing may be particularly beneficial for complex applications like therapeutic communication. Future work could explore training MoE models where different experts specialize in specific types of mental health concerns or therapeutic modalities.

The remarkable learning progression observed in our reward curves indicates that expert feedback provides much richer learning signals than traditional supervised learning approaches. This suggests potential for active learning approaches where the model identifies examples where it is most uncertain and prioritizes expert feedback for those cases.

7 Conclusion

This work demonstrates that reinforcement learning from human feedback, specifically incorporating licensed psychologist supervision, represents a significant advancement over traditional supervised learning approaches for developing therapeutic language models. Our key findings include: (1) PPO training with expert feedback achieves superior performance across multiple evaluation metrics compared to supervised fine-tuning and general-purpose baselines, (2) Mixture-of-experts architectures provide additional benefits for therapeutic applications, (3) Expert-guided reinforcement learning produces models that demonstrate genuine learning progression rather than pattern memorization, and (4) Human-in-the-loop training enables alignment with professional therapeutic standards that cannot be achieved through dataset-based learning alone.

The implications extend beyond technical performance improvements to fundamental questions about how AI systems should be trained for sensitive applications requiring domain expertise and ethical considerations. Our approach establishes a methodology for incorporating professional expertise directly into the optimization process, ensuring that model behavior aligns with established professional standards and safety protocols.

Future research directions include expanding expert evaluation to multiple licensed professionals, implementing safety mechanisms for crisis detection and referral, evaluating performance across diverse mental health scenarios beyond depression, and exploring active learning approaches to optimize expert feedback efficiency. Additionally, longitudinal studies with real users would provide crucial insights into the practical effectiveness and safety of expert-trained therapeutic language models.

The ultimate vision for this work is not to replace human therapists but to democratize access to empathetic, clinically-informed emotional support for the millions of individuals who currently lack access to professional mental healthcare. By ensuring that AI systems learn from and align with professional therapeutic expertise, we can work toward technology that genuinely serves human wellbeing while maintaining the highest standards of clinical safety and ethical responsibility.

8 Team Contributions

This project was completed entirely as individual work by Dean Barrow. All aspects of the research, implementation, and analysis were conducted independently:

- **Dean Barrow:** Conceived and designed the complete experimental methodology comparing supervised fine-tuning against reinforcement learning approaches. Implemented the supervised fine-tuning baseline using LoRA on the ESConv dataset and identified its limitations through systematic evaluation. Developed and implemented the novel PPO reinforcement learning approach with human expert feedback integration. Established collaboration with a licensed psychologist to serve as the reward function, coordinating evaluation protocols and ensuring consistent clinical assessment. Conducted model training and evaluation across multiple architectures including Mistral-7B and Mixtral-8x7B, managing computational constraints and optimizing training procedures. Performed comprehensive quantitative analysis using BERTScore, ROUGE-1, and F1 metrics, along with qualitative analysis of model learning progression and response quality. Executed comparative evaluation against established baselines including GPT-3.5 Turbo and LLaMA-2 7B. Authored complete technical documentation and analysis of results.

Changes from Proposal My original proposal focused primarily on supervised fine-tuning approaches using the ESConv dataset, with reinforcement learning as the next step. However, early experiments revealed that supervised fine-tuning suffered from severe overfitting and poor generalization, leading me to pivot toward RLHF as the primary methodology. The upgrade from Mistral-7B to Mixtral-8x7B was made based on mentor guidance and proved crucial for achieving higher-quality therapeutic responses. The integration of licensed psychologist feedback became the central innovation of my project rather than an auxiliary component, fundamentally changing my research focus toward human-expert-guided optimization rather than dataset-based learning. This pivot proved essential to achieving meaningful results and represents a significant methodological contribution to the field.

References

Alejandro Garcia-Rudolph, David Sánchez-Pinsach, Anna Gilabert, Joan Saurí, Maria Dolors Soler, and Eloy Opisso. 2025. Building Trust with AI: How Essential is Validating AI Models in the Therapeutic Triad of Therapist, Patient, and Artificial Third? Comment on What is the Current and Future Status of Digital Mental Health Interventions? *The Spanish Journal of Psychology* 28 (2025), e3.

JP Grodniewicz and Mateusz Hohol. 2023. Waiting for a digital therapist: three challenges on the path to psychotherapy delivered by artificial intelligence. *Frontiers in Psychiatry* 14 (2023), 1190084.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards Emotional Support Dialog Systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 3469–3483.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL]

A Additional Experiments

Throughout my development process, I conducted several additional experiments that provided valuable insights into the challenges and opportunities of training therapeutic language models. My initial experiments with different prompting strategies revealed something quite interesting—the specific wording of patient statements significantly influenced model response quality. This finding reinforced my belief that diverse training prompts would be essential for robust generalization.

I also ran preliminary experiments with different reward function designs, comparing binary (appropriate/inappropriate) versus continuous (0-1) scoring approaches. The continuous scoring approach proved far superior for enabling gradual policy improvement, while binary scoring led to frustratingly unstable training dynamics and slower convergence. This experience taught me that nuanced feedback is crucial for this type of application.

My computational experiments comparing different batch sizes and learning rates revealed an unexpected finding—smaller batch sizes (which were actually forced on me by GPU memory constraints) actually improved training stability for RLHF. I suspect this improvement came from increased gradient noise helping the model escape local optima in the policy optimization landscape, though this deserves further investigation.

B Implementation Details

The technical implementation demanded careful attention to several critical details that I learned through trial and error. Memory management for large language models proved more challenging than anticipated, requiring me to implement gradient checkpointing and mixed-precision training to fit within my available GPU memory constraints. I built my PPO implementation using the Hugging Face Transformers library with custom reward integration, which meant I had to modify the standard training loop to incorporate human feedback—a non-trivial engineering challenge.

Developing expert evaluation protocols was particularly important for ensuring consistent results. I worked closely with my licensed psychologist collaborator to establish standardized scoring rubrics and documentation procedures for each evaluation. We also needed mechanisms for handling edge cases where responses were difficult to categorize. The psychologist used a structured evaluation form that I designed to capture both numerical scores and qualitative feedback explaining the reasoning behind each assessment.

Data preprocessing for the ESConv dataset required more work than I initially expected. I had to filter for conversation quality, removing inappropriate or low-quality exchanges, and format conversations into instruction-following examples suitable for language model training. I paid special attention to maintaining the emotional context and therapeutic structure of the original conversations while adapting them for my experimental framework—a balance that proved more delicate than anticipated.

I would like to thank my mentor Bassem for his support and advice. I found that the Mixtral 8x7b model is better suited for conversational therapy.