
Enhancing LLMs for Therapy

Maanav Agrawal
maanav2@illinois.edu

Aakash Narayan
aakashn3@illinois.edu

Madhav Parthasarathy
madhavp2@illinois.edu

Dhruv Saligram
dhruvks2@illinois.edu

Abstract

As large language models (LLMs) become increasingly powerful and integrated into daily life, their ability to engage in conversations involving human emotion and meaningful discussions about mental health requires closer examination. Prior research has focused on where LLMs succeed and fail in this task, but there is little literature specifically dedicated to understanding how to make them more effective. To bridge this gap, we conducted extensive research into the current landscape of LLMs acting as therapists in an effort to understand what factors contribute to making them as useful as possible. Our research led us to the core finding that fine-tuning models can increase their effectiveness to a large degree. Interestingly, however, there was little information explaining how to actually fine-tune models for something as intricate, private, and personal as mental health. As such, our contribution is a comprehensive, extensible, and accessible pipeline for fine-tuning LLMs on publicly available client-therapist transcripts. The efficacy of our pipeline was validated through extensive evaluation centered around comparing how well our fine-tuned model replicated human therapist responses compared to a base LLM. Through our pipeline and evaluation, we detail a way to make LLMs better at acting as therapists and a way to automatically test their effectiveness without requiring large-scale user studies.

View the pipeline results [here](#).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI'25, May 16, 2025, Champaign, IL, USA.
Copyright © 2025 ACM

Introduction

As the power and ubiquity of large language models (LLMs) grow, their presence in the realm of mental health similarly scales up. With increasing discussions about LLMs' abilities to "support, augment, and even fully automate psychotherapy" [5], it has become more important than ever to analyze how capable LLMs are in maintaining discussions related to mental health. Most of the existing literature in this field of study revolves around understanding where LLMs currently fail and succeed, pointing out their strengths and weaknesses. In this research undertaking, our core goal was to move beyond this and understand: how can we make LLMs better at operating as therapists?

The decision to explore this research was further motivated by Stanford's CoAuthor dataset [3]. CoAuthor was set up as a way to analyze the capabilities of LLMs in creative and argumentative writing by generating a dataset that captured interactions between writers and an LLM. By extracting quantitative insights from the recorded interactions, CoAuthor was able to understand how LLMs were typically used in writing and in which scenarios they were the most effective support agents. With this in mind, our aim was to use a similar approach and take it in a different direction by applying it to the field of mental health.

Related Work

The recent surge of interest in using large language models (LLMs) for mental health support has led to a variety of studies that explore their potential, limitations, and ethical considerations. Much of the foundational work focuses on evaluating LLMs' generative capabilities and their performance on mental health prediction tasks, as well as designing interaction datasets to study human-AI

collaboration.

Lawrence et al. [2] present a broad analysis of the benefits and risks of using LLMs in mental health contexts by discussing both general models like ChatGPT and more domain-specific ones such as Med-PaLM 2. They emphasize the importance of caution during deployment due to concerns about overreliance on AI for sensitive emotional support. With the potential risks of misinformation and bias, such overreliance could be detrimental. They advocate for fine-tuning and stakeholder involvement to ensure safer and more equitable outcomes.

Several studies have taken a more performance-focused approach. Xu et al. [6] evaluate models such as GPT-3.5, GPT-4, and instruction-tuned versions like Mental-Alpaca and Mental-FLAN-T5 for classifying mental health conditions from social media data. Their findings emphasize that fine-tuned, domain-specific models can outperform larger general-purpose models while remaining more lightweight and accessible. This supports our decision to investigate how LLMs can be adapted through properly structured mental health interactions.

A key consideration in using LLMs for mental health is interpretability. Yang et al. [7] explore this by testing various prompting strategies like emotion-enhanced and chain-of-thought reasoning on models including InstructGPT, LLaMA, and ChatGPT. They show that more structured prompts significantly improve explanation quality, but also reveal that even strong models struggled with ambiguous emotional reasoning. Their follow-up work introduces MentaLLaMA [8], an instruction-tuned model trained on a new dataset of 105K multi-task samples. It achieves state-of-the-art performance and produces explanations relatively better than those of

ChatGPT, while also remaining open-source and interpretable. This paper was a key turning point for our investigation as we wanted to explore how designing an LLM to be more human-like would allow it to be more interpretable and trustworthy for users.

In addition to classification tasks and prompt engineering, researchers have begun simulating real-world therapy scenarios. Heinz et al. [1] conducted a randomized clinical trial on a generative AI chatbot called Therabot, demonstrating significant reductions in depression, anxiety, and disordered eating symptoms across a 12-week study. The trial showed not only symptom improvement but also high levels of user engagement and therapeutic alliance, making it one of the first clinically validated applications of LLMs in mental health. Similarly, our team developed a fine-tuning pipeline that uses real therapist-client transcripts to teach LLMs how to respond with empathy and relevance. Our evaluation which is both quantitative and qualitative shows clear improvements and reinforces the value of having domain-specific training data.

As previously mentioned, a particularly relevant influence on our methodology is Lee et al.'s CoAuthor dataset [3], which captured interactions between writers and GPT-3 across creative and argumentative tasks. By logging detailed writing sessions and user feedback, CoAuthor offered new insights into GPT-3's ideation and collaboration capabilities. This dataset inspired our approach to designing a dataset using real transcripts from mental health conversations to gain insights into how LLMs can be used in the context of therapeutic dialogue rather than writing assistance. Together, all these works highlight the importance of moving beyond benchmark accuracy and exploring how LLMs perform in real interactions, especially those involving real emotion.

Expert Insights

To further understand how AI could be integrated into mental health, we decided to reach out to a therapist who would be able to offer their perspective on the challenges and limitations of using AI in this field. The therapist we reached out to is a trauma therapist who works with children and teens (ages 1-18) facing issues such as abuse, neglect, traumatic immigration experiences, along with other difficulties.

In the conversation we had with this therapist, he emphasized that though AI could be helpful in support-related roles such as resource navigation and lookup, it still lacks the ability to build relationships and perceive physical symptoms in therapeutic sessions. Specifically, many younger patients have limited verbal skills and much of the therapy is reliant on non-verbal cues, such as body language and tone. As the current LLMs are limited to text or voice inputs, these systems are unable to accurately perform in these situations where visual feedback is so crucial.

Another challenge the therapist brought up was risk assessment. In these situations, AI models may respond too quickly by providing resources that the client should be using. However, with human therapists, the risk assessment is done by building trust through slow and nuanced conversation, which helps place the client in a more comfortable environment. Additionally, along the lines of trust, the therapist mentioned that the client-therapist relationship is a large part of the therapeutic process and AI has yet to properly replicate this ability. Typically, therapists hold multiple sessions with a client to build a stronger connection, which will in turn make difficult conversations a little easier for the client. On the other hand, if AI wasn't able to retain the

context from all of the prior conversations, the therapy may feel episodic in the sense that each session is individual, rather than built off the previous sessions.

In terms of positives, the therapist discussed a couple of areas in which AI may fit within the therapeutic process. Particularly, he mentioned that AI has promising applications in tasks like identifying public assistance programs or organizing session transcripts (summarization). Additionally, for specific types of therapy, there are resource guides that therapists use for directing conversation or sessions. If the AI systems were to be trained with these guides, they would be able to cater towards different types of client issues.

Needfinding/User Study

From speaking to a therapist, we were able to understand the perspective of an expert in the field of mental health. Additionally, through reading related works, we were able to picture the current landscape of LLMs as therapists; However, to help understand how the general population views AI in therapy, we decided to perform a user study to capture their opinions. We created and posted a survey on Slack and separately reached out to other undergraduate/graduate students in college. Participants were asked to engage with a baseline LLM (Gemini) using the prompt “Hi Gemini, I’d like you to act as a supportive therapist in a conversation that focuses on mental health” as the starting point for conversation. These conversations were aimed to mirror real therapeutic exchanges, allowing users to explore mental health topics of personal value. Following this interaction, participants completed an evaluation survey designed to capture qualitative and quantitative measures of their respective experiences. The survey questions included those about the emotional understanding Gemini provided, the general

advice provided, and the trust the users felt in conversing about their emotional troubles. Along with these survey questions, the participants were also given a free response area in which they could write about some improvements that could be made to help facilitate a better experience for the user.

To ensure user privacy, we did not ask the participants to share their personal information or any of the conversation data with us. Additionally, we placed a disclaimer at the top of the survey, explaining that participants should only converse with Gemini to the level of depth that they are comfortable with. Our intent by doing so, was to prevent putting any pressure on the participants.

User Study Results

A total of 13 participants completed the evaluation survey following their conversations with Gemini. These results offered great insight into users’ perceptions of AI in therapy and the strengths/weaknesses of these systems.

Emotional Understanding

One question of the survey asked participants “How emotionally understood did you feel during your interaction with the AI?” The answers ranged from 1 to 5, with 1 being not understood at all, and 5 being completely understood; The average score was between 2 (**38.5%**) and 3 (**46.2%**), meaning that Gemini struggled to replicate the deeper emotional awareness that human therapists convey.

Mood Impact

Participants were also asked “Did the conversation help improve your mood or perspective afterward?” Around **30.8%** of participants said no, and another **38.5%** of participants said somewhat 1. While the **38.5%** figure seems promising, the variability in Gemini’s impact on

Did the conversation help improve your mood or perspective afterward?
13 responses

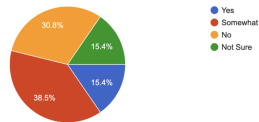


Figure 1: Pie Chart representing how participants's mood improves or didn't improve with Gemini.

How safe did you feel sharing personal or sensitive information with the AI?
12 responses

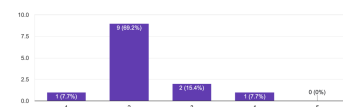


Figure 2: Bar Chart representing how safe participants felt sharing their personal information with Gemini.

mood shows that AI is inconsistent and may entirely depend on how a user responds to the type of therapy that Gemini provides.

Methods of Helpfulness

When asked "In what way(s) did the AI help you?", **61.5%** of participants found that Gemini helped clarify thoughts and **30.8%** of participants said that Gemini helped guide them through emotional reflection. These responses further support the potential AI has in mental health and therapy.

Trust and Privacy

In terms of perceived safety when sharing sensitive information, **69.2%** of users gave a score of 2 (on a scale of 1 to 5) 2. This is an ongoing issue that people face with technology, especially in relation to big corporations such as Google, OpenAI, Anthropic that use large amounts of data to train their models. For AI to be effective in therapy, models need to be created in which the data is completely privatized (adhering to HIPAA policies) and completely in the control of the user.

Other Areas that AI Could Improve

In a free response section of the survey, participants were able to discuss their thoughts in an unstructured manner. Participants mentioned quite a few areas that AI could improve the quality of their responses in such as:

- "It could better provide advice that is more tuned to the situation and not generic advice."
- "One key area is building a relationship with the person or me in this instance."
- "The interface to use it didn't allow for an ideal interaction. When it responds, you're hit with a wall

of text all at once that's not paced or given multiple questions to respond to all at once."

These responses reflected the findings that we saw in the related works section, specifically with the importance of trust and careful conversation. Additionally, these results demonstrated that while LLMs like Gemini can sometimes provide emotional support, there are still significant gaps that remain in depth of conversation, trust and safety of data, as well as in paced responsiveness. These findings emphasize the need for more nuanced conversational skills along with better context-awareness, specifically with call-backs to previous talking points.

Core Findings & Motivation

Through analyzing related works, speaking with a licensed professional, and conducting our own user study, we were able to generate insights into how LLMs perform as therapists and what critical limitations and benefits it has. After our needfinding, however, we discovered a critical point in the evaluation of LLMs as therapists – fine-tuned LLMs saw huge leaps in performance.

Therabot [1] - previously mentioned in our related works - was the first randomized controlled trial testing LLMs as therapists. Therabot's results were excellent: participants receiving the intervention showed a significantly greater reduction in depression symptoms compared to the control group, indicating that the model had huge promise and appeared to overcome many of the typical LLM drawbacks previously discussed. A key driver behind this improvement, as mentioned in their paper, was fine-tuning their model in a specialized way on an expert-gathered dataset. Separately, Xu et. al [6] conducted an analysis of different LLMs in mental health-based tasks, and similarly found fine-tuned models to outperform base

models. Critically, these two papers underscore that fine-tuning seems to be an extremely effective way to make LLMs better therapists. However, neither paper offers any discussion on the fine-tuning process itself. Therabot's discussion ends with the mention of an expert-gathered dataset, while Xu et. al failed to discuss how the best-performing models were actually fine-tuned. This creates a significant hole in the research into this area – fine-tuned models seem to perform the best, but there is no discussion regarding how to best fine-tune them. This conflict motivated our fundamental contribution. Our proceeding work aims to support reproducibility and democratize access to fine-tuning techniques for therapeutic LLM applications.

Pipeline

With the core findings that fine-tuning LLMs seemed to be the key driver behind increasing their efficacy in acting as therapists and that there were no clear details on how these high-performing LLM therapists were fine-tuned, we set out to create a transparent, extensible, and accessible pipeline to achieve this. After extensive and iterative progress, our pipeline was developed with three main stages: data collection, pre-processing, and model fine-tuning.

Data Collection and Pre-Processing

The key to fine-tuning a model for enhanced performance is establishing high-quality training data that encompasses a wide breadth of interactions that the LLM should engage in. [4] To this effect, we aimed to create a training dataset that captured professional human performance in therapy. The Counseling and Psychotherapy Transcripts collection published on Alexander Street seemed to have both high quality and quantity in terms of anonymized therapist session transcripts. The main advantage of using

these transcripts as our training data (as opposed to other mental health related works such as research papers, books, or journals) was its emphasis on the back-and-forth dialogue between clients and therapists, which is the natural method of communication with LLMs. Patients often ramble, trail off, and switch topics quickly, and we believed including these mannerisms in the dataset would allow the LLM to more accurately simulate a human therapist in their responses. Using these transcripts also ensured ethical compliance, as they were already available online for academic use.

While an advantage of our chosen publication was its quantity of therapy transcripts, it was infeasible to use all of them for fine-tuning due to the scope of our undertaking. As such, we carefully selected a variety of transcripts that encompassed a broad range of presenting conditions, including anxiety, depression, anger, frustration, and more. This variety ensured that our training data was representative of different clients, therapists, conversation styles, and gravity. This was a critical choice, as different types of issues require unique handling and to make our fine-tuned model effective, it should accordingly respond uniquely in different situations. As seen from our needfinding therapist interview, the role of a therapist is often to encourage further conversation from the patient, allowing them to express their emotions. This is often done through restatement of what the patient said or short affirmations. However, some situations require firmer instruction or the introduction of a different perspective. We also ensured this kind of diversity in our test set, allowing us to capture model performance variation between different mental health problems and bolster evaluation.

Once the selected transcripts were collected, they were

processed and client / therapist dialogue was extracted. Extensive cleaning was done to remove timestamps, dates, and other erroneous or irrelevant text. Each back-and-forth was preserved in a JSON format suitable for fine-tuning, and the data was then passed through a fine-tuning job.

Initial Fine-Tuning Results

After preprocessing the data and ensuring its suitability for model ingestion, the next step was to fine-tune a model for evaluation. To do this, we selected Gemini for its straightforward interface and cost-effectiveness. Our choice of Gemini was also motivated by a desire for consistency between the model users interacted with during our initial user study and our fine-tuning. This allowed us to focus on iterating and adapting our fine-tuning data, which proved to be the biggest driver of increased model performance. With further time and efforts, more powerful and expressive models could also be explored and iterated upon.

The next step was deciding how to structure the data for fine-tuning. As these transcripts contained full conversations, we initially split each transcript to have a single data entry be centered around one client / therapist back-and-forth. This was primarily motivated by findings in our user study. Through those results, we saw that one of the key shortcomings of base LLM performance was its response type. Often, the LLM would produce a long paragraph addressing all parts of the prompt and providing a variety of suggestions. Not only was this overwhelming, but it tended to disrupt the natural flow of conversation, meaning that users would not relax and open up about their issues. As such, we believed that preserving each back-and-forth as its own training example would help the fine-tuned model maintain shorter

responses that were more pointed and direct.

After fine-tuning, however, we found that our new model was completely ineffective. No matter the prompt's topic, length, tone, or explicit instructions, the model would only respond with "Ok.", "I'm sorry to hear that", or similar remarks and platitudes. Upon analysis, we realized that this was likely due to the way we structured our fine-tuning data. The vast majority of individual therapist responses in our dataset were less than 8 words, which meant that the fine-tuning overcorrected to these examples. By only responding with a few words and completely ignoring situations where the therapist would say something meaningful or provide an insight, the fine-tuned model was still able to achieve high accuracy on the training dataset since the data was mainly comprised of the same short responses. At this point, our fine-tuned model was clearly unsuitable for any serious discussion about mental health.

Final Fine-Tuning Results

The key reason behind our initial fine-tuned model failing was, as previously mentioned, the overrepresentation of quick, affirmative responses from the therapist. In order to create a truly effective model, it was clear that the data input for fine-tuning had to be changed.

The first change we experimented with was which specific therapist responses to include in the training data. Our final decision was to only include therapist responses where the therapist said 50+ words. This had two main benefits. First, we only fine-tuned our model on dialogue where the therapist was actually saying something very substantive. Second, the longer therapist responses would give the model more context about what a human therapist actually sounds like. As we saw in our user study, the ineffective sentence structure and content of

Human Therapist Response:	“Right? I mean, there is intentionality in it, the thought process makes sense. But it doesn't realize. I mean, what your body is doing, except it's not in the right scenario. So, I mean, that heart- racing feeling makes a lot of sense. There is no threat that your body is responding to. And it's when you're anxious. It's like your body is responding as the way it would as if there were some sort of threat.”
Base Model Response:	“Okay, so you're feeling like your body is working against you, making you unhealthy instead of supporting your health. And that makes you feel like you have a disease. That sounds incredibly frustrating and upsetting. Can you tell me more about what specifically your body is doing that feels so detrimental? And what makes it feel like a disease to you?”
Fine-Tuned Model Response:	“Not that it's conscious. I mean, it's not like you sit down and say, “I'm going to make my heart race and I'm going to have trouble breathing and I'm going to feel anxious,” although I guess some people can do that. I don't think that's what you're talking about. But I think your body does it because it's trying to keep you safe and alive, and the way it does that is sometimes, in this case, a little bit, I think, like overprotective. It's a little bit like the parent who smothers their kid and won't let their kid climb trees because they want them to be safe. It's that kind of feeling of, like, protection that can also inhibit. So if you feel anxious or stressed or frustrated, it's because of something. It's a response to something.”

Figure 3: Human, base model, and Fine-tuned model responses to the same prompt

LLM responses is often what hinders it from ultimately being helpful. By being far more selective with what data we fine-tuned the model on, we forced it to adapt to sound more like a human therapist and offer similar quality of insights.

The second change we made was including previous context in the training data inputs. This was done for two main reasons. First, it would help the LLM understand the context of a conversation so that a more relevant output could be generated. Second, it would offer the LLM a better view of conversation flow. Since our training data was limited now to training on therapist responses with 50+ words, including back-and-forth contexts from right before these 50+ word statements would also give the model examples of when to offer shorter responses. Now, the actual training objective would be for longer responses (preventing the model from overcorrecting to constant empty platitudes), but shorter responses would still be fed to the model as part of the context for additional learning.

By training only on longer therapist responses and including these context windows, our results were much better and far exceeded the base Gemini model. Our quantitative evaluation differentiating the two is present in the next section, but initial qualitative evaluation showed that fine-tuning led to much better responses from the model. One such example from the hundreds can be seen in Figure 3.

Considerations and Weaknesses

While this pipeline provides a rich source of relevant data and noticeably improved performance with the fine-tuned model, there are some important considerations to highlight. First, all transcripts used in the training data are from spoken sessions – conversations with chatbots or over an online interface are inherently different. The

language used and type of explanation vary between these mediums, and the dataset does not have any representations of this difference. As tele-therapy becomes more and more widespread in the future, similar transcripts could be sourced to extend the dataset and provide representation of these types of conversations. For example, our fine-tuned model is trained on text with more switches in speaker and natural interruptions, leading to more casual output than might be desired in a chatbot environment. This could be addressed partially with pre-prompting the LLM with instructions, but would likely not fix the issue fully due to its consistent presence in the training data. Second, by using the transcripts as our training data, our underlying fine-tuning goal is to have the model sound more like a human therapist. This does make the implicit assumption that the human therapists in the transcripts are always displaying behavior that we want the model to emulate. If adapted to a production setting, we could have a qualified professional filter or rate therapist behavior, and pass this information along to the model for even richer fine-tuning.

Model Evaluation

To evaluate the impact of our fine-tuning pipeline, we tested both the base and fine-tuned models on a set of real-world therapist-client transcripts not used during training. This setup allowed us to evaluate how well the models generalized to unseen conversations and, more critically, whether the fine-tuned model exhibited behavior closer to that of a human therapist.

We computed several evaluation metrics across six categories of mental health symptoms: Anxiety, Depression, Fatigue, Frustration, Sadness, and Anger. The metrics included:

- **Absolute response length difference** and **sentence difference**, to assess structural alignment with human therapist responses.
- **Sentiment distance**, measuring emotional similarity using a sentiment classifier.
- **Question frequency**, including the number of question-containing responses and total question marks.
- **Name usage**, evaluating how well the model personalized its tone and mirrored client input. Name usage mismatch refers to the number of times where the human therapist referred to the patient by name but the model didn't, or vice versa.

Overall, we found that fine-tuning meaningfully shifted the model's behavior in a direction more aligned with human therapists both quantitatively and qualitatively. While the base model often defaulted to being more verbose and having significantly more questions, our fine-tuned model showed concision, emotional clarity, and improved conversational pacing. This is significant since therapy relies heavily on subtlety and presence, and we observed that fine-tuning helped eliminate much of the "chatbot-like" rigidity seen in baseline LLMs.

Table 1 and Table 2 illustrate this change in two high-volume categories. For anxiety, the base model asked over three times as many questions as the therapist which potentially disrupted flow whereas the fine-tuned model moderated its questioning to near-human levels. In depression-related cases, we saw the base model tending toward emotional detachment, as reflected in a higher sentiment distance. The fine-tuned model, however, struck a better balance between empathy and conciseness,

producing responses more structurally and emotionally in line with actual therapist replies.

Interestingly, the fine-tuned model did not outperform the base across every single metric. For instance, in some categories (like anger), its response length slightly increased post fine-tuning. This suggests that while fine-tuning can guide general behavioral improvements, it does not guarantee optimal responses in all dimensions particularly when emotional nuance varies significantly between cases. This also underscores that showing emotion is not simply a matter of copying structure but requires deeper semantic alignment.

These findings reinforce a core insight from our project which is that base LLMs, while technically capable, tend to default to generic or overly enthusiastic responses when asked to perform emotionally complex roles. By contrast, our fine-tuned model learns to "step back" in conversation mirroring the more affirming strategies seen in real therapy. Ultimately, this supports the argument that general-purpose LLMs are insufficient on their own for high-stakes mental health applications, and careful fine-tuning is not just helpful, it's essential.

Additional evaluation tables for Frustration, Sadness, Anger, and Fatigue are provided in the appendix.

Table 1: Evaluation metrics for the Anxiety subset.

Metric	Base Model	Fine-Tuned Model	Therapist
Avg Length Diff (abs)	72.30	63.15	—
Avg Sentiment Distance	0.2106	0.2015	—
Avg Sentence Diff (abs)	4.93	4.93	—
Name Usage Mismatch	38	41	—
Num of Questions Asked	336	59	106

Table 2: Evaluation metrics for the Depression subset.

Metric	Base Model	Fine-Tuned Model	Therapist
Avg Length Diff (abs)	52.79	61.95	—
Avg Sentiment Distance	0.1929	0.1887	—
Avg Sentence Diff (abs)	4.20	3.82	—
Name Usage Mismatch	23	27	—
Num of Questions Asked	320	61	133

Use Cases & Extensions

Below is a discussion of a few practical use cases for our pipeline and model that we envision, along with extensions that we believe could even further bolster its quality.

Extension into Text-to-Speech Therapeutic Interfaces

An immediate extension of our fine-tuned model could be integration with text-to-speech (TTS) systems to form more immersive and human-like therapeutic interfaces. Given that our model is fine-tuned to emulate natural dialogue patterns drawn from spoken therapeutic sessions, coupling it with TTS technologies can further enhance its utility by delivering responses with emotional intonation, pacing, and tone reflective of human speech. This multimodal application would be particularly impactful in accessibility contexts – such as with users who struggle with text-based communication or prefer auditory engagement – and could simulate the presence of a therapist more convincingly.

Leveraging Fine-Tuned Models in Retrieval-Augmented Generation (RAG) Systems

A compelling use case involves deploying our fine-tuned model as the generative backbone in a Retrieval-Augmented Generation (RAG) system tailored to mental health support. In such a framework, external knowledge sources such as psychological literature and crisis response protocols could be dynamically retrieved and supplied to the fine-tuned LLM to ensure that its outputs are both contextually sensitive and factually grounded. This approach could mitigate one of the major

limitations of static LLMs: their tendency to hallucinate or rely on outdated information.

User Privacy and Ethical Design in Deployment

As therapeutic applications inherently involve sensitive and personal user data, any real-world deployment of our pipeline's fine-tuned models must prioritize user privacy and ethical safeguards. Future extensions of this work should incorporate mechanisms such as local computation or differential privacy to ensure that user inputs are not stored or used inappropriately. Moreover, transparent data usage policies and user consent frameworks must be clearly implemented to give users control over their interaction data.

Conclusion

In this work, we analyzed the current landscape of LLMs engaging in discussions of mental health and operating as therapists. Upon discovery that fine-tuning models was a significant booster of success and that little discussion was centered around how to effectively fine-tune models for therapy, we created a reusable and extensible pipeline for fine-tuning models to better simulate human therapists. The success of our pipeline was evaluated by comparing our fine-tuned model to its corresponding base LLM and evaluating how much of an improvement the fine-tuning made. Ultimately, our work's main contribution is a clear pipeline for effectively fine-tuning LLMs for therapy and an evaluation system for determining fine-tuning success without needing a large-scale user study.

References

- [1] Heinz, M. V., Jack, R. R., Dowling, G., Kruger, J. L., Tsai, C. K., Copeland, W. E., Allen, N., and Horton, W. A. Randomized trial of a generative ai chatbot for mental health treatment. *NEJM AI* (2025).

- Forthcoming.
- [2] Lawrence, H. R., Schneider, R. A., Rubin, S. B., Matarić, M. J., McDuff, D. J., and Bell, M. J. The opportunities and risks of large language models in mental health. *arXiv preprint arXiv:2402.01110* (2024).
 - [3] Lee, M., Liang, P., and Yang, Q. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*, ACM (2022).
 - [4] Smith, J. R. Ibm research releases 'diversity in faces' dataset to advance study of fairness in facial recognition systems, 2019. Originally published January 29, 2019; updated February 15, 2019.
 - [5] Stade, E. C., Stirman, S. W., Ungar, L., Boland, C. L., Schwartz, H. A., Yaden, D. B., Sedoc, J., DeRubeis, R. J., Willer, R., Kim, J. P., and Eichstaedt, J. C. Toward responsible development and evaluation of llms in psychotherapy, June 2024.
 - [6] Xu, X., Yao, B., Dong, Y., Gabriel, S., Yu, H., Hendler, J., Ghassemi, M., Dey, A. K., and Wang, D. Mental-llm: Leveraging large language models for mental health prediction via online text data. *arXiv preprint arXiv:2307.14385* (2024).
 - [7] Yang, K., Ji, S., Zhang, T., Xie, Q., Kuang, Z., and Ananiadou, S. Towards interpretable mental health analysis with large language models. *arXiv preprint arXiv:2304.03347* (2023).
 - [8] Yang, K., Zhang, T., Kuang, Z., Xie, Q., Huang, J., and Ananiadou, S. Mentallama: Interpretable mental health analysis on social media with large language models. In *Proceedings of The Web Conference 2024 (WWW '24)* (2024).

Appendix: Additional Evaluation Metrics

To supplement the core evaluation results provided in the main paper, we include additional tables here for the remaining four symptom categories: Frustration, Sadness, Anger, and Fatigue. These tables follow the same format and highlight consistent trends.

Table 3: Evaluation metrics for the Fatigue subset.

Metric	Base Model	Fine-Tuned Model	Therapist
Avg Length Diff (abs)	50.62	45.17	—
Avg Sentiment Distance	0.1718	0.1666	—
Avg Sentence Diff (abs)	3.40	2.95	—
Name Usage Mismatch	4	7	—
Num of Questions Asked	139	29	25

Table 4: Evaluation metrics for the Frustration subset.

Metric	Base Model	Fine-Tuned Model	Therapist
Avg Length Diff (abs)	47.25	43.58	—
Avg Sentiment Distance	0.1781	0.1544	—
Avg Sentence Diff (abs)	3.11	3.06	—
Name Usage Mismatch	9	6	—
Num of Questions Asked	109	30	21

Table 5: Evaluation metrics for the Sadness subset.

Metric	Base Model	Fine-Tuned Model	Therapist
Avg Length Diff (abs)	47.25	43.58	—
Avg Sentiment Distance	0.1781	0.1544	—
Avg Sentence Diff (abs)	3.11	3.06	—
Name Usage Mismatch	9	6	—
Num of Questions Asked	109	30	21

Table 6: Evaluation metrics for the Anger subset.

Metric	Base Model	Fine-Tuned Model	Therapist
Avg Length Diff (abs)	60.76	62.64	—
Avg Sentiment Distance	0.1816	0.1521	—
Avg Sentence Diff (abs)	3.78	4.56	—
Name Usage Mismatch	11	22	—
Num of Questions Asked	230	45	69