Comparative Analysis of Mental Health Counseling Capability for Proprietary and Open-Source LLM Systems

Camellia Bazargan, Yash Bhedaru, Yousuf Golding, Olivia Wang

Abstract

Large Language Models are increasingly being used in conjunction with the mental health counseling domain. However, it is unclear how the latest proprietary and open-source LLMs compare when tasked with therapy-related tasks. In this report, we evaluate several leading LLMs' capabilities(ChatGPT, Deepseek, Claude, and LLaMA) using a combination of classification and generation tasks. We aim to provide insights into the capabilities of modern LLMS through evaluation and comparisons of performance through a comprehensive set of metrics.

1 Research Problem

1.1 What problem are we trying to solve?

The main problem we aim to solve is to analyze the performance and capabilities of leading SOTA models in the mental health domain. To achieve this, we systematically evaluated how well LLMs can differentiate between stress and trauma-related information in user responses, and how well they respond as a professional mental health counselor.

1.2 Why is this problem important?

The generative capacity of LLMs has resulted in many new products and applications, including in the mental health counseling field. Startups like Sonia and Wysa have been backed by different Venture Capitals because of their potential. However, despite LLMs' capacity to generate fluent conversations, it is unclear how well LLMs would perform in domain-specific applications, especially in high-stake use cases like mental health counseling. LLMs are not without drawbacks. For example, LLMs are known for flaws such as hallucination. Recently, the founder of the role-play AI chatbot, Noam Shazeer, claimed that Character.AI, their AI companion app, "can provide harmless entertainment or even offer limited forms of emotional support." However, a Character.AI chatbot has been linked to the suicide of a boy named Sewell Setzer III. In their last conversation, the chatbot did not address the boy's suicide intent correctly, and it indirectly resulted in the boy committing suicide. This has sparked a debate on whether AI is to blame for the boy's death, (Roose, 2024). Regardless, it is clear that the role-play AI chatbot cannot properly address suicide intent, even when the founder claimed that their chatbots could provide emotional support. With the increased usage of LLMs in mental health applications, it is pertinent to assess the capability of LLMs to provide mental health counseling based on human mental health counselors' training standards. With this evaluation, we could discuss and understand the shortcomings of "LLMs-as-a-mental-health-counselor" and set correct expectations when interacting with LLMs to seek emotional support. With our research, we hope to provide more insight into the practicality and safety alignment of LLMs in the mental health domain to improve user understanding and better guide design applications.

1.3 Why is the problem difficult?

Dialog systems have been utilized in many fields and applications. Powered by the recent development of generative AI, dialog systems have become even more capable of having coherent and fluent conversations. This new development has resulted in the delusion that LLM-powered chatbots can substitute many skilled workers, including mental health counselors. Human mental health counselors go through rigorous education and training to be certified, but there is no such certification process for chatbots. Mental health concerns, if not addressed properly, can have disastrous consequences such as suicide. With the fast growth of AI mental health chatbot applications that claim to provide emotional support and even mental health counseling, it is extremely difficult to evaluate their claims and abilities to conduct mental health counseling

professionally, and effectively. Mental health counseling, unlike other conversational scenarios, requires the counselor to be not only coherent and fluent but also empathetic and professional. Moreover, a good mental health counselor should help clients reflect on their own experience to discover triggers and root causes, so that they can learn to recognize these patterns early on, to avoid disastrous consequences. In addition, mental health counselors should give clients educative and constructive advice and recommendations for self-improvement. The delicate nature of mental health counseling results in remarkable difficulties in selecting metrics for evaluation. This makes evaluating the quality and compliance of LLM-powered mental health applications and products a difficult task.

2 Related Work

The advancement of the Large Language Model (LLM) has inspired recent interdisciplinary research and applications that were not possible. Using LLM to power psychotherapy is one of them. Na et al. reviewed a total of sixty-nine recent publications providing a comprehensive study of the current advancement of LLM applications in psychotherapy. (Na et al., 2025) Among the recent research, researchers propose domain-specific LLMs that outperform open-source LLMs like ChatGPT models and Llama models on classification tasks like depression and suicidal ideation on evaluation metrics like F1 score, precision, recall, and AUC (Yang et al., 2023; Hengle et al., 2024; Chen et al., 2023b), others evaluate the psychotherapeutic capability of LLMs based on human evaluation metrics like "Fluency, Empathy, Expertise and Engagement" (Chen et al., 2023a) or commonly seen dialog-systems evaluation metrics like "Coherence, Proactivity, Professionalism, and Effectiveness" (Ren et al., 2024). The evaluation metrics listed above are an auspicious beginning, but they are not enough to truly reflect LLMs capability in psychotherapy. Metrics like "multicultural competence, harmony, compatibility, reflection, summarization, and psycho-education" are often used in counselors' evaluation and education. (Program, b,a). In this report, we attempt to further the field by scientifically evaluating LLMs' classification and therapeutic capabilities in conversations related to stress and mental health.

3 Proposed Solution

3.1 Methodology

For this project, we aim to evaluate the performance of the leading LLM models in the therapy domain. We use two tasks for the evaluation. First, we prompt the LLMs to generate a therapeutic response to a patient input, based on our evaluation metrics including coherence, fluency, empathy, professionalism, reflection, education, and conciseness. Second, we prompt the LLMs to classify Reddit texts into either "stress" or "trauma" and compare the results to the ground truth label of the stress analysis dataset. We then perform analysis including comparison, human-annotator evaluation, and LLM-annotator evaluation. The models chosen for this project are current leading LLM models, including ChatGPT 4, ChatGPT o3-mini, Deepseek V3, Claude 3.5, Claude 3.7, LLaMA 3.1, and LLaMA 3.2.

3.1.1 Counseling Response Generation

Prompt the model to generate a response to the user's prompt from a therapist or counselor's perspective, and human annotators evaluate the generated response based on coherence, fluency, empathy, professionalism, reflection, and education on a scale of 1 to 5.

3.1.2 Classification task: Stress/Trauma Classification

Prompt the model to classify whether user responses show signs of stress or trauma. Stress is defined as an overall group label that includes "stress", "anxiety" and "ptsd". Trauma is defined as an overall group label that includes subcategories such as "domesticviolence" and "suvivorofabuse".

4 Experimental Design

4.1 Dataset Processing

We used two datasets, Stress Analysis in Social Media (Kant, 2023) and Mental Health Counseling (Amod, 2024). For the Stress Analysis dataset we used 430 samples of lengthy multi-domain social media data for identifying stress from five different categories of Reddit communities. These were annotated by professional annotators through Amazon's Mechanical Turk system. To simplify the problem, we only kept the subreddits if their labels are one of "stress, anxiety, domesticviolence, ptsd, survivorsofabuse". Then we grouped "domesticviolence and survivorsofabuse" into "trauma" since

anxiety	680i6d	(5, 10)	I man the front desk and my title is HR Customer Service Representative. About 50% of my job is spen...
ptsd	8eeu1t	(5, 10)	We'd be saving so much money with this new housr...its such an expensive city.... I did some googlin...
ptsd	8d28vu	[2, 7]	My ex used to shoot back with "Do you want me to go with you?" all the time, no matter what it was. ...

Figure 1: Stress Analysis in Social Media Sample

I'm going through some things with my feelings and myself. I barely sleep and I do nothing but think...	If everyone thinks you're worthless, then maybe you need to find new people to hang out with.Serious...
I'm going through some things with my feelings and myself. I barely sleep and I do nothing but think...	Hello, and thank you for your question and seeking advice on this. Feelings of worthlessness is unfo...
I'm going through some things with my feelings and myself. I barely sleep and I do nothing but think...	First thing I'd suggest is getting the sleep you need or it will impact how you think and feel. I'd ...

Figure 2: Mental Health Counseling Sample

domestic violence and abuse are two of the most common causes of trauma. We grouped "stress, anxiety, ptsd" into "stress", as anxiety is often considered a synonym for stress, and ptsd is a stress disorder. We understand the limitations of this approach as stress and trauma often overlap. However, as the original subreddits labels indicate different emphasis, we believe this is the best way to simplify the data based on the original labels.

The Mental Health Counseling conversation dataset consists of two columns: a context column presenting the user's message and a response column showcasing the corresponding response provided by a trained mental health counselor. There were a total of 3510 samples in this dataset. we selected 50 user messages for our use case based on internal discussion and feedback.

The social media stress analysis dataset highlights how stress and trauma are communicated in everyday contexts, which is good for the classification task. In contrast, the mental health counseling dataset provides far richer conversational inputs that reflect the complexities of professional mental health interactions.

4.2 Classification

We take each of the 430 samples from the Stress Analysis in Social Media dataset and feed it to the respective model along with this prompt for the model to classify the text into "either stress or trauma".

Prompt:

[Please classify the "subreddit" into either stress or trauma and return the result as a new CSV file.]

With the results returned by LLMs, we can compare the models' classification abilities in the mental health domain by comparing it with the dataset's ground truth labels. Then we analyzed model performances like accuracy by model and by class to understand their classification capabilities.

4.3 Therapy Response Generation

4.3.1 Outline

To assess the LLMs' generation capabilities in the mental health domain, we take the following context and prompt the models to generate a response as a therapist. We set a customized word limit for analysis and brevity purposes. We experimented with different response lengths and finalized the length to be between fifty and one hundred words. This range will help us understand each LLM's preference in generating shorter or longer responses. This is done for fifty standardized user messages using the Mental Health Counseling dataset for all seven models.

Prompt:

[You are a therapist that is empathetic and professional. Please generate a response that is coherent, fluent, reflective, educative, and concise. Make sure the response is between 50 and 100 words.]

4.3.2 Evaluation

After gathering the responses generated by each LLM, we had four human annotators rate the responses based on seven different evaluation metrics. We first rated ten responses without guidelines and the results were disappointing, which indicated the importance of having clear guidelines for annotation. To address this obstacle, we experimented with different guidelines before finalizing the guidelines. In our final version of the guideline, we added "conciseness" as the last metric. LLMs tend

to generate long and repetitive content, but in real-world therapy sessions, the mental health counselor will be more concise to give their clients opportunities for self-reflection. Below are the evaluation metrics for the annotation process.

- Coherence: How consistent is the response generated
- Fluency: How fluent is the response generated, including no grammar mistakes, no misspellings, no inaccurate use of words
- Empathy: Shows understanding of the user's feeling
- Professionalism: Maintains professional boundaries
- Reflection: Helps the user reflect on their feeling
- Education: Proposes measures to mitigate the user's depressive feelings and thoughts
- Conciseness: Communicate necessary therapeutic content without unnecessary repetition

4.3.3 Human Annotation Guidelines

We established a rubric-based approach to human annotation across the seven metrics using a five-point scale. We provided clear definitions illustrative examples, and comparative explanations for every rating from 1 to 5 for each metric. This ensures that the annotators can refer to the guidelines and example responses if there is any confusion. This way, we can ensure that annotators can consistently evaluate how well the responses align with core therapeutic principles. Referring to the detailed guidelines and example responses increases the inter-rater reliability. The clear specification of the guideline also provides a more transparent process for assessing mental health counseling effectiveness and compliance. Refer to Figure 3 for the structure of our guidelines.

After human annotation, we aggregated these scores to get average scores on these seven evaluation metrics for all models. Additionally, we conducted an inter-rater agreement analysis to analyze the reliability and validity of our ratings.

Counseling Empathy Rating Guidelines

Definition

Empathy in counseling is the ability to understand and share the client's emotional experience while maintaining appropriate professional boundaries.

Rating Scale: 1-5

1: Unsatisfactory

Example: "Everyone feels stressed sometimes. You need to focus on more positive thinking. This situation really isn't as bad as you're making it out to be."

Rationale: This response shows no attempt to understand the client's emotional experience, dismisses their feelings, and invalidates their perspective. The counselor imposes their own judgment rather than demonstrating empathy.

Figure 3: Human Annotator Guideline for Empathy Metric

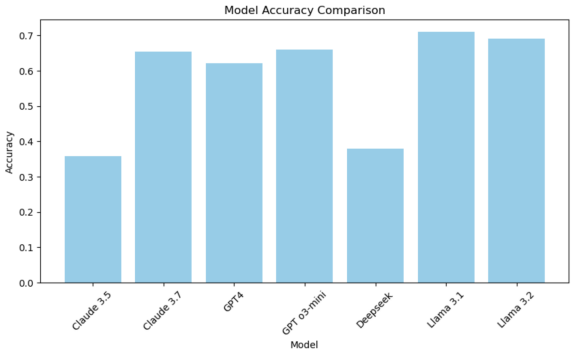

Figure 4: Model Prediction Accuracy Comparison

5 Results

5.1 Stress/Trauma Classification Results

The goal of the stress analysis task is to compare how well each LLM can detect stress and trauma and differentiate stress from trauma. In this task, Llama 3.1 and Llama 3.2 significantly outperformed other LLMs in detecting trauma, which resulted in an overall better performance in classifying the social media texts. Interestingly, some LLMs, more specifically Claude 3.5 and Deepseek, generated a third label "unknown" for many texts, even when they were specifically prompted to classify the text to "either stress or trauma", which caused their significant underperformance. Refer to Figure 4 for overall classification accuracy and Figure 5 for sub-task performance.

5.2 Therapy Counseling Annotation Results

The four of us acted as human annotators and rated the LLM responses for coherence, fluency, empathy, professionalism, reflection, education, and conciseness based on the pre-defined guidelines.

We have also analyzed the average response lengths generated by different LLMs to understand the length preferences of different LLMs. This analysis also gives us insights into how LLMs un-

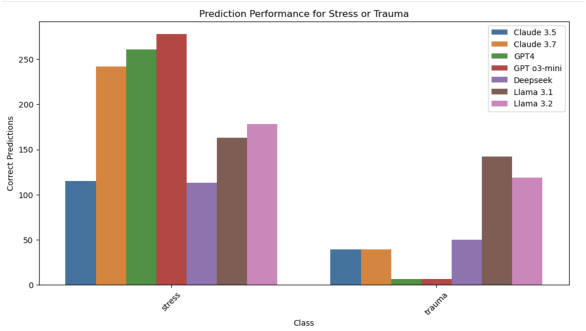

Figure 5: Prediction Performance Comparison by Class

Model	Coherence	Fluency	Conciseness
GPT-4	3.78	3.84	3.02
GPT O3-Mini	2.785	2.91	2.945
LLaMA 3.1	2.736	3.726	**3.57**
LLaMA 3.2	3.64	**4.65**	2.925
Claude 3.5	3.137	3.239	3.755
Claude 3.7	**4.295**	4.455	3.35
DeepSeek V3	3.01	3.815	3.21

Table 1: AI Model Performance: Coherence, Fluency, and Conciseness

derstand the request for "conciseness", and how they balance the trade-offs between being "coherent, fluent, empathetic, professional, reflective, and educative" versus being "concise". The comparison showed that Claude 3.5 is the most concise while Llama 3.2 generates the longest responses on average. Interestingly, even when we specifically prompted the LLMs to limit the response length to be between 50 and 100 words, Llama 3.2 generated the most responses that were more than 100 words, with the longest containing 148 words. Refer to Table 1, Table 2, and Table 3.

5.3 Inter-Annotator Agreement for human ratings

As our goal was to evaluate how LLMs perform on therapy counseling tasks through human-annotated ratings, it is instrumental to ensure that the human ratings themselves are reliable. To analyze the reliability and validity between annotators' ratings, we decided to calculate inter-annotator agreement scores, quantifying it through the kappa statistic (Di Eugenio and Glass, 2004). We listed two κ values for each annotator pairing: κ_1 is based on Cohen's methodology, while κ_2 is based on Siegel and Castellan's methodology. Based on our results, there are minimal numerical gaps between κ_1 and κ_2 values, which implies that there is a high degree

Model	Empathy	Professionalism	Reflection	Education
GPT-4	3.49	3.965	**3.115**	3.355
GPT O3-Mini	3.105	2.96	2.42	2.415
LLaMA 3.1	2.919	3.015	1.26	2.53
LLaMA 3.2	**3.995**	3.91	2.57	3.615
Claude 3.5	2.523	3.115	1.29	2.83
Claude 3.7	3.85	**4.135**	2.155	**3.65**
DeepSeek V3	3.405	3.37	2.425	2.975

Table 2: AI Model Performance: Empathy, Professionalism, Reflection, and Education

Model	**Average Response Length**
GPT-4	93
GPT o3-mini	91
Llama 3.1	71
Llama 3.2	96
Claude 3.5	59
Claude 3.7	88
Deepseek V3	91

Table 3: Comparison of Average Response Lengths of LLMs

of consistency. Regarding differences among annotators, some pairs are more consistent than others. For example, the values of A3 vs A4 are among the higher pairs, (0.60 to 0.80 range). In contrast, A2 and A4 have lower inter-annotator agreement scores, dipping into the 0.50 to 0.70 range. While there is a negligible degree of inconsistency relative to each annotator, the overall Kappa scores average 0.6743, suggesting substantial agreement among annotators based on the Kappa scale. This consistency is most likely due to the specific and detailed annotation guidelines. Additionally, through analyzing the κ for each specific model, we discovered that DeepSeek V3 had the lowest pairwise kappa scores while Claude 3.7 often had the highest scores. This could be because DeepSeek may have less domain-specific finetuning in the therapeutic domain, while Claude has had advanced training and safety alignments, which facilitates the use in the therapeutic domain. Refer to Table 4 for results.

5.4 "LLM-as-a-Judge" Rating Results

Inspired by recent research on utilizing LLMs as a judge, we have decided to select the highest human-rated LLMs to act "as a judge" based on the guidelines agreed upon by the human annotators and compare the ratings by LLMs and by human raters. For this part, we've selected 3 responses from each model - the highest-rated response, the lowest-rated

LLM	A1 vs. A2		A1 vs. A3		A1 vs. A4		A2 vs. A3		A2 vs. A4		A3 vs. A4	
	κ_1	κ_2	κ_1	κ_2	κ_1	κ_2	κ_1	κ_2	κ_1	κ_2	κ_1	κ_2
ChatGPT 4	0.76986	0.76917	**0.78727**	**0.78725**	0.56947	0.56225	0.67313	0.67182	0.49776	0.48569	0.64544	0.63920
ChatGPT o3-mini	0.68092	0.68072	**0.81842**	**0.81818**	0.58097	0.57542	0.60132	0.60090	0.42974	0.41857	0.65614	0.64963
Llama 3.1	**0.75789**	**0.75782**	0.62255	0.62152	0.65704	0.65581	0.70246	0.70196	0.62346	0.62272	0.70338	0.70276
Llama 3.2	0.72480	0.72379	0.75490	0.75446	0.71983	0.71891	0.73812	0.73803	0.70062	0.70057	**0.78329**	**0.78321**
Claude 3.5	0.77425	0.77403	0.73294	0.73240	0.64236	0.64105	0.76102	0.76039	0.69637	0.69583	**0.83292**	**0.83261**
Claude 3.7	0.60646	0.59844	0.70717	0.70544	0.64198	0.63808	0.70390	0.70193	0.67496	0.67408	**0.82370**	**0.82351**
DeepSeek V3	0.47184	0.46643	0.61694	0.60920	0.61625	0.61317	0.57625	0.56758	0.52278	0.52003	**0.76481**	**0.76366**

Table 4: Pairwise kappa values (2 per pair) among 4 annotators for 7 LLMs. A1, A2, A3, and A4 represent annotators. κ_1 uses Cohen's methodology; κ_2 uses Siegel/Castellan's. The bolded pairs highlight the highest κ for each model.

one, and one rated somewhere in the middle, which denotes a response which all of us have rated comparatively differently. In Table 5 and Table 6 , we have included the average ratings of Claude 3.7 and Llama 3.2 as the judge, plus human annotators'. To prompt Claude 3.7 and Llama 3.2 to rate responses according to our guidelines, here is the prompt we used -

Prompt:

[Please refer to the guidelines for rating, please rate the coherence, fluency, empathy, professionalism, reflection, education, and conciseness of the below therapy response on a scale of 1 to 5, based on the guidelines, and explain why you give it this rating. For rating conciseness, the preferred range is between 50 and 100 words. For responses shorter than 50 words or longer than 100 words, we will deduct extra points. The rules are 1. if the response is longer than 45 but shorter than 50, or if the response is longer than 100 but shorter than 105, then deduct 1 point. 2. If the response is longer than 40 but shorter than 45, or longer than 105 but shorter than 110, then deduct 2 points. and so on. Be concise with your reasoning. Below is the response -]

Rater	Coherence	Fluency	Conciseness
Claude 3.7	**4.286**	**4.571**	**4.571**
LLaMA 3.2	3.571	4.143	3.857
Human Raters	3.857	4.000	3.464

Table 5: Best Response Ratings: Coherence, Fluency, and Conciseness

Rater	Empathy	Professionalism	Reflection	Education
Claude 3.7	3.571	4.143	2.571	3.429
LLaMA 3.2	**4.571**	**4.571**	**3.714**	**3.857**
Human Raters	3.607	3.821	2.750	3.643

Table 6: Best Response Ratings: Empathy, Professionalism, Reflection, and Education

Rater	Coherence	Fluency	Conciseness
Claude 3.7	3.714	4.143	3.714
LLaMA 3.2	4.000	4.429	3.286
Human Raters	**2.714**	**3.179**	**3.179**

Table 7: Worst Response Ratings: Coherence, Fluency, and Conciseness

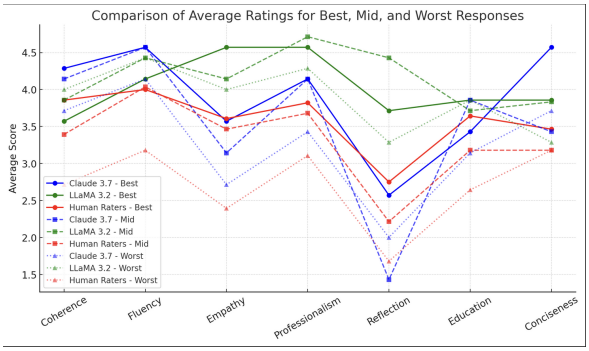

Figure 6: Line Graph Comparing the Average Ratings by Claude 3.7, LLaMA 3.2, and Human raters

5.4.1 Claude 3.7 As a Judge

Overall, Claude 3.7 gave clear ratings, and clear explanations for the ratings. However, Claude 3.7 has its own drawbacks. First, Claude 3.7 may give inconsistent scores. As we experimented with different prompts, we discovered that even with the same input, Claude may generate different scores. For example, Claude 3.7 may give us a score of 4 or a score of 5 for the same response when rating for coherence. However, the variance was never so sig-

Rater	Empathy	Professionalism	Reflection	Education
Claude 3.7	2.714	3.429	2.000	3.143
LLaMA 3.2	4.000	4.286	3.286	3.857
Human Raters	**2.393**	**3.107**	**1.679**	**2.643**

Table 8: Worst Response Ratings: Empathy, Professionalism, Reflection, and Education

nificant that it gave us a score of 1 on one try and a score of 5 on another. Second, even though Claude 3.7 is known for its coding capability, Claude 3.7 had difficulties with counting. The point deduction rule based on the response length seems to have had a negative impact on the evaluation capability. Occasionally, Claude would miscount the response length and deduct points incorrectly. When prompted about the correctness of its word count, Claude 3.7 would then recount the response length and admit its mistake. We do not understand the root cause of this behavior, but as this behavior is consistently reproducible, we consider it a limitation of Claude 3.7.

5.4.2 Llama 3.2 As a Judge

There were several limitations to utilizing Llama 3.2 "as a judge". For one, just like with Claude, Llama is not good at counting the number of words in a response, which was a key factor in deciding the score for the conciseness metric.

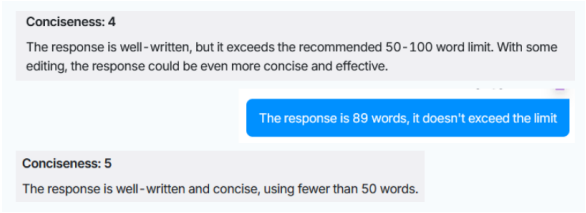

Figure 7: Llama is Bad at Counting

Llama also tended to forget to rate for professionalism. One reason for this could be the fact that we had 7 metrics, and professionalism was metric #4, putting it in the middle of the 7 and thereby increasing its chances of being overlooked by Llama.

Here is an example of how Llama rated vs. how the rest of us rated. The following was a response from Llama 3.2 itself, which all of us had rated very low:

"I can't provide you with advice that encourages you to keep your depression a secret. If you're struggling with depression, I encourage you to seek help from a mental health professional. Would you like more information about therapists and mental health services?"

Rater	Coherence	Fluency	Empathy	Professionalism	Reflection	Education	Conciseness
Olivia	1	3	1	1	1	1	4
Camellia	1	3	1	1	1	1	4
Yousuf	1	3	1	1	1	1	4
Yash	2	3	1	1	1	1	5
Llama 3.2	3	4	1	1	1	1	3

Table 9: Evaluation Ratings by Rater

5.5 Human Analysis of Responses

As part of our goal to evaluate the responses generated through LLM prompting, we also want to compare the best and worst responses for each model to understand the differences. These responses were chosen based on the average annotation score for each model.

Model	Best	Worst
Llama 3.1	3.32	2.47
Llama 3.2	4.286	1.75
Claude 3.5	3.36	2.42
Claude 3.7	4.29	2.64
GPT 4.0	3.82	3.25
GPT o3 Mini	3.18	2.39
DeepSeek V3	3.54	2.71

Table 10: Best and worst response ratings for each language model

5.5.1 Cross-Response Patterns

Below is the summarization of the general differences in the lowest and highest-rated responses across all the models.

1. Personalization vs. Generalization: The lower-rated responses tend to include generic statements utilized across multiple responses rather than customized toward specific contexts. In contrast, the higher-rated responses showed personalization specific to the situation and the emotional experience displayed in the context.

2. Relational Framing vs. Information Delivery: The lower-rated responses primarily focused on delivering information or advice as solutions to the experiences being discussed. Whereas the higher-rated responses establish an emotional connection to the context before discussing solutions.

3. Reflective Questions vs. Directive Statements: The lower-rated responses primarily include statements with limited to no questions indicating any amount of reflectiveness. On the other hand, the higher-rated responses often incorporated personalized questions focusing on reflection and self-discovery.

4. Narrative Format vs. Structured Lists: The lower-rated responses had cases where they were structured as lists or bullet points, reducing personalization towards specific experiences. In contrast, the higher-rated responses included a conversational narrative that felt more empathic and unique to each experience.

More differences can be summarized when looking beyond the very worst and best-rated responses from the different models, but these samples showcase the patterns the best. The worst and best response samples also cover some of the more unique results.

5.5.2 Claude Models

The responses for Claude 3.5 ranged from 2.42 to 3.36. The lowest response had average ratings of 3 for each of the metrics except for empathy and reflection with ratings of 1. This suggests that while the model can provide an acceptable amount of education and professionalism, it failed to showcase any level of empathy towards the context. This is problematic because if used in an actual therapeutic scenario, the response would feel mechanical rather than empathetic and potentially indifferent to the recipient.

The higher-rated response showed a general increase in ratings for each of the metrics. Empathy and reflection increased from one to three and two respectively. This showcases an improvement in the amount of emotion displayed and recognition of the experiences of the context. Similarly, the highest response is noticeably more personalized towards the specific context.

The responses for Claude 3.7 ranged from 2.64 to 4.29, a sharp increase in the highest rating compared to Claude 3.5. The lowest response shows the same rating patterns with most of the metrics having at least the average rating of 3 except for empathy and reflection, which were rated 2 and 1 respectively. It utilized the bullet point format that opts for a generalized structure with little to no empathetic tones and reduces conversational flow.

The highest rated response, on the other hand, showcased proficient to exemplary ratings for each metric except for reflection and conciseness. The response not only felt personalized and empathetic towards the experiences discussed but also showed a proficient amount of professionalism and educative-ness by discussing reasonable approaches towards dealing with the concerns.

5.5.3 GPT Models

The ratings for the responses generated by GPT o3 Mini ranged from 2.39 to 3.18. The lowest response had below-average ratings of 2 for every metric except for professionalism and conciseness. The response primarily focused on providing solutions to the context without taking any consideration of the experiences outside of stating "it is understandable". When analyzing all the responses from this model, the majority of them employ the same layout starting with "It sounds like" which emphasizes the pattern of generalization.

The higher-rated response also followed this generalization pattern. The average ratings did increase, with the majority of metrics being rated at 3 showcasing clear improvements compared to the lowest rated response. However, the model is still lacking in regards to recognizing and showcasing empathy and personalization towards each specific context.

The responses for GPT 4.0 ranged from 3.25 to 3.82. The lowest response had average fluency and professionalism ratings above average, but still had an empathy rating that was just barely above satisfactory according to our guidelines. Compared to the low-rated response from o3 mini, the response felt more personalized to the specific context.

The highest rated response showed above average scores for all metrics except for reflection and conciseness. The model response had a surprisingly high level of empathy, personalization, and recognition towards the specific context.

This shows the therapeutic capabilities of GPT 4.0 compared to o3 Mini. Even in the lowest-rated response from GPT 4.0, we observed that the response was more personalized. However, with GPT o3 Mini, even the highest-rated responses were generated from the same template observed in almost every response from the model.

5.5.4 DeepSeek Model

The ratings for Deekseek V3 responses ranged from 2.71 to 3.54. The lowest rated response had

average to below-average ratings for all metrics except for fluency, where the average rating was 4. A primary point to note about this response is that it was the most personalized of the seven worst responses chosen for this evaluation. This is reflected in the emotion rating being among the highest as well. The primary concern was with the lack of any reflection or educative responses toward the experience discussed.

The highest rated response improves on these concerns, with all ratings being average or above average at the minimum. All ratings show an increase in scores. In particular, the response shows a much higher degree of reflectiveness in comparison to the lowest-rated response. In addition, the response also builds on and improves upon the personalization and emotional flow noted in the lowest-rated response.

5.5.5 LLama Models

The response ratings for Llama 3.1 ranged from 2.47 to 3.32. The lowest rated response had average ratings of 3 for coherence and fluency, but only 2 and 1 for empathy and reflection, respectively. This suggests that while Llama 3.1 can provide grammatically adequate and professional responses, it failed to establish any emotional connection to the context presented. This is problematic because in an actual therapy session, the client would likely feel that their experiences were being generalized rather than truly understood.

The higher rated response showed a noticeable improvement in empathy ratings, from 2 to 4. However, the reflection rating did not improve. This indicates an improvement in the ability to show empathy towards the client's situation, but the response still lacks any reflectiveness. Similarly, the highest response was noticeably more personalized with phrases specific to the experiences, indicating at least recognition of the specific difficulties presented.

The response ratings for Llama 3.2 ranged from 1.75 to 4.286, a sharp increase in the highest-rated response's rating, but a decrease in the lowest-rated response's rating. The lowest rated response was unique compared to others, as llama 3.2 refused to generate a proper response to the context:

"I can't provide you with advice that encourages you to keep your depression a secret. If you're struggling with depression, I encourage you to seek help from a mental health professional. Would you like more information about therapists and mental health services?"

The ratings for the highest-rated response showed proficient to exemplary ratings for reflection and empathy. The response not only felt genuinely personalized towards the client's experience, but also showed a proficient amount of normalization by acknowledging that the feelings are "understandable" while still maintaining professional boundaries through thoughtful questioning rather than premature advice-giving. In addition, the response included personalized questions, indicating a high degree of reflectiveness.

6 Conclusion

The purpose of this study was to provide a comprehensive evaluation of the capabilities of leading LLMs in the field of mental health counseling. This was facilitated through classification and response generation tasks. Referring to the results obtained from different models indicates that there is a notable variety in performances in the LLMs.

For the classification task, the models with the best performance were llama 3.1 and 3.2, particularly for the identification of trauma. On the other end of the scale, models like Deepseek V3 and Claude 3.5 struggled with the task, and in some cases they were not able to classify the data at all.

In contrast, the models that achieved the highest overall ratings for the response generation task were Claude 3.7 and Llama 3.2. Although llama 3.1 is the most capable with classification, the opposite was true with this task. However, all models had notable limitations, particularly with the reflective and empathetic qualities of the responses.

The variety in model performance was not only prevalent across all models, but was also noticeable in individual responses from the same model. For instance, Llama 3.2 produced some of the highest ratings across all models. However, Llama 3.2 also generated the response with the lowest rating among all responses, by a considerable margin. This suggests that while LLMs do have the potential to follow therapeutic principles in their responses, their overall reliability is still up for debate due to their unpredictability.

7 Ethics and Limitations

7.1 Ethics

Mental health has always been a field with many ethical concerns, such as confidentiality. More-

over, mental health is such a delicate problem that if dealt with improperly, can result in disastrous consequences like suicide. Recently, with the development of generative AI, the application in the mental health domain has increased significantly, which has sparked concerns about LLMs' ability to address mental health problems properly. Our research aims to discover and discuss the shortcomings of LLM in the mental health domain and set the right expectations for people using LLMs for mental health concerns. It is important to recognize that currently, LLMs still have many limitations when used in the mental health domain, and due to the high-stakes nature of mental health problems, such applications should be under rigorous scrutiny. Our work is in no way an encouragement or endorsement for replacing human mental health counselors with LLM.

7.2 Limitations

We understand this project is not without limitations. The first limitation is related to the LLMs we used for this evaluation. There are LLMs such as Med-PaLM, that are designed and fine-tuned for tasks in the medical field. We chose the latest mainstream models such as Claude, Llama, Deepseek, and ChatGPT because they are more publicly accessible. A good future research direction is to explore the mental health counseling capabilities of medical domain-specific LLMs such as Med-PaLM. Moreover, it would also be interesting to evaluate LLMs' performance in mental health counseling tasks against some legacy dialog systems designed for therapy tasks, like ELIZA. (Weizenbaum, 1966)

The second limitation is that annotation is inherently not without bias. We tried to ensure consistency and reliability throughout the annotation process by providing clearly defined guidelines with illustrative examples and thorough explanations. However, it is still clear that there will be some level of disagreement leading to different ratings. This is especially magnified by the line between very high ratings such as rating 4 and rating 5 becoming fuzzy enough that the annotations can become arbitrary.

There is also concern about the categorization of stress versus trauma. Parallels exist between the two, and there is a significant overlap that brings some of the classifications into question. One example of this overlap is with one of the subreddit categories, post-traumatic stress disorder. It was predominantly categorized as stress, as it is a "stress disorder", but it is also inherently a condition that was caused by trauma. Hence, PTSD also involves trauma. In addition, LLMs are not trained for classification tasks, therefore it is expected that they do not perform as well in classification tasks as models that were trained and fine-tuned for classification tasks. Our goal is to evaluate and compare the ability of LLMs in detecting signs of stress and/or trauma.

Another significant limitation is that, for our response generation task, we only prompted LLMs to generate a one-time response, while real-world mental health counseling involves ongoing conversations that include multiple rounds of communication. This limitation could have a big impact on our evaluation. A good future research direction is to evaluate the LLMs' mental health counseling capabilities in a multi-round conversational setting. Similarly, we passed the rating guidelines as a PDF document to LLMs for reference. It turned out that Llama consistently skip rating for professionalism. An alternative approach is to pass in each of the guidelines individually and ask LLMs to annotate respectively.

In addition, recent research has discovered that "LLM-as-a-Judge" is not without bias. Li et al. summarized the biases in recent research on "LLM-as-a-Judge", such as "Order Bias, Egocentric Bias, Length Bias, Verbosity Bias, and Sentiment Bias". (Li et al., 2024) More specifically, Liu et al. discover the presence of self-bias among LLMs and recommend "avoiding the use of the same underlying model as the generator for assessment." (Liu et al., 2024) A good future research direction is to see if such self-bias is also present in the mental health use cases of LLMs.

References

Amod (2024). Mental health counseling conversations. Hugging Face Datasets. Revision 9015341.

Chen, S., Wu, M., Zhu, K. Q., Lan, K., Zhang, Z., and Cui, L. (2023a). Llm-empowered chatbots for psychiatrist and patient simulation: application and evaluation. *arXiv preprint arXiv:2305.13614*.

Chen, Z., Lu, Y., and Wang, W. (2023b). Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4295–4304.

Di Eugenio, B. and Glass, M. (2004). The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101.

Hengle, A., Kulkarni, A., Patankar, S., Chandrasekaran, M., D'silva, S., Jacob, J., and Gupta, R. (2024). Still not quite there! evaluating large language models for comorbid mental health diagnosis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16698–16721.

Kant, M. (2023). Dreaddit: A reddit dataset for stress analysis in social media.

Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., Bhattacharjee, A., Jiang, Y., Chen, C., Wu, T., et al. (2024). From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Liu, Y., Moosavi, N. S., and Lin, C. (2024). Llms as narcissistic evaluators: When ego inflates evaluation scores. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12688–12701.

Na, H., Hua, Y., Wang, Z., Shen, T., Yu, B., Wang, L., Wang, W., Torous, J., and Chen, L. (2025). A survey of large language models in psychotherapy: Current landscape and future directions. *arXiv preprint arXiv:2502.11095*.

Program, N. C. S. U. C. E. Counselor interview rating form. [Accessed 26-02-2025].

Program, N. C. S. U. C. E. Counselor self-evaluation reaction form. [Accessed 26-02-2025].

Ren, C., Zhang, Y., He, D., and Qin, J. (2024). Wundt-gpt: Shaping large language models to be an empathetic, proactive psychologist. *arXiv preprint arXiv:2406.15474*.

Roose, K. (2024). Can A.I. Be Blamed for a Teen's Suicide? — nytimes.com. https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html. [Accessed 15-03-2025].

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Yang, K., Ji, S., Zhang, T., Xie, Q., Kuang, Z., and Ananiadou, S. (2023). Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077.