Evaluating the Therapeutic Quality of LLMs in Mental Health Contexts

Roz Huang
ronghuang0604@berkeley.edu

Sohail Khan
sohailk2@berkeley.edu

Abstract

As conversational AI systems become increasingly integrated into daily life, their role as empathetic and trusted advisors—particularly in sensitive contexts like mental health therapy—has grown significantly. This study evaluates the empathetic capabilities of large language models (LLMs), with a focus on Meta's LLaMA. Building on prior research into empathy evaluation frameworks such as EmotionQueen, we fine-tune LLaMA 3.1 8B Instruct using curated datasets comprising mental health counseling scenarios. We establish baseline performance using models like FLAN-T5 and T5, assessed through a combination of lexical (BLEU, ROUGE) and semantic (BERTScore, Sentence Transformer similarity) metrics. Fine-tuning experiments reveal challenges such as input-label misalignment and demonstrate the effectiveness of qLoRA for resource-efficient adaptation. The results highlight LLaMA's consistent performance across prompts, the impact of prompt specificity on FLAN-T5, and the limitations of T5 for this domain. While LLaMA shows promise as a foundation model for empathetic applications, substantial gaps remain in achieving human-like responsiveness, underscoring the need for further research in empathetic AI.

1 Introduction

Artificial Intelligence and large language models (LLMs) are becoming increasingly intertwined in the interactions and lives of humans today. What began as relatively simple models for paraphrasing or text summarization have evolved into conversational tools capable of explaining foreign concepts or offering advice in specific situations.

Instead of turning to a friend, family, or even a Google search for help, it's increasingly more convenient to go straight to a product like ChatGPT to seek answers in a conversational format. With this shift in behavior and the growing role of AI as a trusted advisor, especially in sensitive contexts like personlized therapy (Iftikhar et al, 2024) [2]., it becomes crucial we not only evaluate the helpfulness of these models but also the emotional quality of their responses.

In this paper, we will evaluate the therapeutic standards that LLMs achieve. First, we will break down the methods for analyzing the responses of generative models and propose a strategy for evaluating the empathetic quality of language models. Then we will evaluate how empathetic LLaMA – Meta's open source LLM – is.

2 Related Work

A variety of strategies have been explored to analyze the deeper "emotional" quality of generative language models. The heart of this issue lies in building a vetted, diverse, and robust methodology to assess not only the accuracy of model outputs but also their ability to convey empathy and emotional nuance.

One promising direction has been the development of metrics and evaluation frameworks tailored to empathetic response generation. For instance, (Luo et al., 2024) [3] introduced a framework that evaluates models based on their ability to align responses with user emotions and contextual needs, emphasizing the importance of dynamic adaptability in conversational agents. Similarly, (Liu et al., 2024) [4] proposed a reinforcement learning-based approach to optimize generative models for empathetic conversation by integrating feedback loops from human evaluators and automated sentiment classifiers.

A significant advancement in this area is the EmotionQueen framework introduced by (Chen et al., 2024) [1] as part of their work on evaluating Meta's LLaMA. The EmotionQueen framework addresses the challenge of evaluating empathy in LLMs by building a carefully curated and contextually rich dataset. This dataset includes prompts derived from real-world conversational scenarios, particularly those requiring nuanced emotional responses. To develop this evaluation dataset, the researchers employed a multi-step strategy:

1. **Diverse Prompt Selection:** Prompts were sourced from a wide array of emotional contexts, including interpersonal conflicts, mental health scenarios, and celebratory moments, ensuring coverage across a spectrum of affective states.

2. **Human Annotation:** Each response was annotated by trained evaluators who assessed both the emotional alignment of the response with the prompt and its linguistic quality. These annotations also included detailed feedback on the depth of empathy expressed.

3. **Iterative Refinement:** Using the annotated responses, the dataset was refined iteratively, prioritizing prompts that revealed critical gaps in the model's empathetic capabilities.

The EmotionQueen framework was not only instrumental in evaluating LLaMA but also highlighted broader limitations in existing LLMs, such as a tendency to produce generic or overly formal responses in emotionally sensitive contexts. The study's findings emphasize the value of such domain-specific datasets for driving improvements in generative models' empathetic capacities.

From a healthcare perspective, the application of AI-driven conversational models for therapeutic contexts has garnered increasing attention. (Iftikhar et al, 2024) [2] explored the integration of AI systems in digital therapy, finding that while these systems can provide personalized and scalable support, they must be rigorously evaluated against clinical standards to ensure safety and effectiveness in emotionally sensitive interactions.

Together, these works underscore the need for holistic evaluation frameworks that combine quantitative metrics with qualitative insights, enabling a nuanced understanding of how well generative models meet both technical and emotional expectations in their interactions.

3 Dataset

To evaluate our baselines and fine-tune the model, we initially used the Amod/Mental Health Counseling Conversations dataset, which consists of real-world mental health questions and answers sourced from two online counseling and therapy platforms. While this dataset provided a solid foundation, the initial fine-tuning performance was suboptimal, likely due to its limited size.

To address this, we incorporated two additional datasets: mpingale/Mental Health Chat Dataset and heliosbrahma/Mental Health Chatbot Dataset. The integration of these datasets expanded the dataset size to 6,292 examples, offering improved diversity and a more comprehensive representation of mental health conversations.

This combined dataset was split into 75% for training, 12.5% for validation, and 12.5% for testing. After preprocessing steps—including removing unnecessary columns, reformatting entries, and filtering out empty responses—each data instance has the following standardized structure:

- **Context:** A question from a user seeking mental health advice.

- **Response:** A corresponding answer crafted by a qualified psychologist.

4 Baseline Model Evaluations

4.1 Baseline Models

To establish a baseline for our fine-tuned model, we compared three pre-trained models: **LLaMA 3.1 8B Instruct**, **FLAN-T5**, and **T5**. These models were carefully chosen to explore two key aspects:

1. **Instruction Fine Tuned**
   We selected T5 as a representative of models that are not instruction-fine-tuned.

2. **Non-Instruction Fine Tuned**
   In contrast, LLaMA 3.1 8B Instruct and FLAN-T5 are instruction-fine-tuned models designed to better understand and follow human instruction. This comparison allows us to assess the impact of instruction fine-tuning on model performance.

Within instruction-fine-tuned models, we compared LLaMA 3.1 8B Instruct and FLAN-T5 to evaluate their performance on our specific task. LLaMA 3.1 8B Instruct is a decoder-only model, while FLAN-T5 follows an encoder-decoder architecture. This comparison sheds light on how differences in architecture influence the models' ability to generate empathetic and supportive responses.

4.2 Evaluation Metrics

To evaluate the models, we employed a combination of word-based and embedding-based metrics, ensuring a comprehensive assessment of the generated responses.

For word-based metrics, we used BLEU and ROUGE (ROUGE-1, ROUGE-2, and ROUGE-L) to measure exact lexical overlap between generated responses and references. These metrics provide an efficient evaluation of surface-level alignment. However, they can undervalue responses that use different but equivalent wording.

Embedding-based metrics, BERTScore and Sentence Transformer (all-mpnet-base-v2), address this limitation by assessing the quality of responses based on their contextual and semantic similarity to references. By combining these two approaches, we capture complementary insights into both the lexical precision and semantic understanding of the models' outputs.

4.3 Prompt Engineering

Large Language Models (LLMs) are highly sensitive to the input context they receive, particularly in task-specific applications. In real-world scenarios, it is often unclear what context the model is expected to handle, as user inputs can vary greatly. To evaluate the performance of our baselines, we tested them across three different contexts:

1. **No Prompt**
   In this scenario, the model receives no specific guidance or instructions regarding the task at hand. This setup was chosen as a baseline to assess the model's performance in a more realistic, open-ended context, where the input might not include detailed task-specific directions.

2. **Prompt 1 - Minimal Guidance**
   You are a licensed therapist. Your role is to respond to the user's question with empathy, understanding, and support.
   User: Context
   Therapist:
   This prompt provides minimal direction, encouraging the model to assume the role of a therapist, but without offering deeper clarification on what constitutes empathy, understanding, or support. This was chosen to evaluate how the model responds with limited guidance in an emotionally complex context.

3. **Prompt 2 - Empathetic Guidance**
   "You are a licensed therapist. Your role is to respond to the user's question with empathy, understanding, and support. Empathy means recognizing and validating the user's feelings while offering thoughtful, non-judgmental responses.
   User: Context
   Therapist: "
   This prompt extends the guidance provided in Prompt 1 by defining empathy in more detail. It aims to measure how well the model incorporates empathy and thoughtfulness when generating responses.

The "No Prompt" condition was selected as a baseline to simulate real-world interactions where the model may not receive structured instructions. Prompt 1 was included to assess the model's response to basic role-playing instructions, while Prompt 2 was designed to evaluate how the model performs when given more specific, empathetic guidelines.

T5 was only evaluated using the "No Prompt" condition, as it is not specifically designed to handle prompt-based tasks and may not perform optimally when exposed to more complex, task-specific instructions. This allows for a clearer understanding of its performance in its native setup without additional task-specific fine-tuning.

4.4 Results and Analysis

The performance of the models was evaluated across multiple metrics including BLEU, ROUGE, BERTScore, and Sentence Transformer similarity (all-mpnet-base-v2). We focus on three models: LLaMA 3.1 8B Instruct, FLAN-T5, and T5, each evaluated with different prompts where applicable.

LLaMA 3.1 8B Instruct: LLaMA shows relatively stable performance across different prompts, with minimal variation in BLEU, ROUGE, and BERTScore. The model performs best under "Prompt 2," with improvements in BLEU (0.0139) and ROUGE-1 (0.2675) compared to the "No Prompt" and "Prompt 1" configurations. The Sentence Transformer similarity also increases slightly from 0.6331 (No Prompt) to 0.6204 (Prompt 2), indicating that the choice of prompt does not significantly affect the performance of LLaMA. The consistent performance across different prompts suggests that LLaMA may be relatively prompt-agnostic in generating responses.

Model	Prompts	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	Sentence Transformer (all-mpnet-base-v2)
Therapist Baseline		0.4362	0.5376	0.3680	0.4343	Precision: 0.6871, Recall: 0.6951, F1: 0.6881	N/A
LLaMA 3.1 8B Instruct	No Prompt	0.0113	0.2167	0.0366	0.1065	Precision: 0.4984, Recall: 0.5549, F1: 0.5234	0.6331
	Prompt 1	0.0134	0.2581	0.0380	0.1205	Precision: 0.4955, Recall: 0.5263, F1: 0.5079	0.6088
	Prompt 2	0.0139	0.2675	0.0388	0.1251	Precision: 0.5113, Recall: 0.5355, F1: 0.5208	0.6204
FLAN-T5	No Prompt	0.0000	0.0591	0.0079	0.0430	Precision: 0.4749, Recall: 0.2997, F1: 0.3641	0.2779
	Prompt 1	0.0023	0.1273	0.0179	0.0923	Precision: 0.5367, Recall: 0.3926, F1: 0.4510	0.4202
	Prompt 2	0.0021	0.1300	0.0208	0.0955	Precision: 0.5414, Recall: 0.3902, F1: 0.4512	0.4254
T5	No Prompt	0.0015	0.0914	0.0090	0.0673	Precision: 0.4556 Recall: 0.3703 F1: 0.4064	0.3574

Table 1: Comparison of models across various evaluation metrics.

FLAN-T5: FLAN-T5, however, demonstrates notable improvements when different prompts are used. The model performs poorly with "No Prompt," with BLEU, ROUGE, and BERTScore all being low (e.g., BLEU = 0.0000, ROUGE-1 = 0.0591). However, when prompted, particularly with "Prompt 1" and "Prompt 2," there is a noticeable increase in performance. For example, the BLEU score increases to 0.0023 for "Prompt 1" and 0.0021 for "Prompt 2," with corresponding improvements in ROUGE-1 and ROUGE-2 as well. Additionally, the Sentence Transformer similarity rises from 0.2779 (No Prompt) to 0.4254 (Prompt 2), suggesting that FLAN-T5 benefits significantly from structured prompts. This indicates that FLAN-T5 is more sensitive to prompt design than LLaMA.

T5: T5 performs the worst among the models, with a BLEU score of 0.0015 under the "No Prompt" condition. Other metrics, including ROUGE, BERTScore, and Sentence Transformer similarity, also reflect low performance. The model's inability to perform well across these metrics indicates that T5 may not be the best choice for tasks requiring higher-quality text generation and more nuanced semantic understanding. This suggests that T5 might struggle with generalization or context-awareness when compared to more advanced models like LLaMA and FLAN-T5.

Conclusion: Based on these results, LLaMA maintains consistent performance regardless of prompt, making it a reliable model for general use. In contrast, FLAN-T5 benefits significantly from prompted configurations, highlighting its sensitivity to input structure. T5, however, lags behind both LLaMA and FLAN-T5 in all metrics, underscoring its limitations for more complex tasks.

However, it's important to note that all these models—including LLaMA—still have significant room for improvement to meet the baseline standards of human-like responses.

5 Fine-Tuning

While pre-trained models are highly capable, they are designed for general-purpose language understanding and generation. This broad focus limits their effectiveness in addressing the unique sensitivities and nuances of mental health-related conversations. Fine-tuning addresses these limitations by training the model on a curated mental health dataset, allowing it to specialize in producing empathetic and contextually appropriate responses.

Our fine-tuning process was not without challenges. Early iterations revealed issues such as misalignment between inputs and labels, leading to inconsistent and incoherent outputs. Through iterative debugging and experimentation, we refined our approach to better align with the architecture of decoder-only models. These adjustments ultimately improved the model's ability to generate contextually appropriate and empathetic responses, highlighting both the potential and the complexities of fine-tuning large language models for specialized applications.

5.1 Fine-Tuning Setup

5.1.1 Model Selection

We selected LLaMA 3.1 8B Instruct as the base model for fine-tuning due to its scalability and strong performance across various natural language understanding and generation tasks. Our baseline evaluations confirmed its capability as a robust foundation for domain-specific fine-tuning. Its instruction-tuned foundation enhances its ability to follow prompts effectively, while its decoder-only architecture supports the generation of fluent and coherent text—both critical for producing high-quality responses. Furthermore, the 8B parameter size offers a balance between expressive power and computational efficiency, making it feasible to fine-tune on our limited resources while preserving the capacity to generate nuanced responses.

5.1.2 Fine-Tuning Strategy

To fine-tune LLaMA 3.1 8B Instruct, we adopted the **qLoRA (Quantized Low-Rank Adaptation)** technique. This method significantly reduces computational overhead and memory footprint, enabling us to fine-tune the 8B parameter model on consumer-grade Google Colab GPUs while retaining strong performance.

The model was quantized to 4 bits using the `nf4` quantization type. We focused fine-tuning on the most critical components of the model's architecture, specifically targeting the query and value projection layers in the attention mechanism. We set the rank (`r`) to 8 and the scaling factor (`lora_alpha`) to 16. This configuration ensured a manageable number of additional trainable parameters while providing sufficient influence on the pre-trained weights. Additionally, a 10% dropout rate was applied within the LoRA layers to prevent overfitting and enhance generalization. This setup allowed us to fine-tune the model efficiently under resource constraints while achieving high-quality outputs.

5.2 Fine-Tuning Experiments

5.2.1 Iteration 1: Initial Fine-Tuning with Prompt 2

We selected *"Prompt 2 - Empathetic Guidance"* for fine-tuning based on its better overall performance during baseline evaluations. However, the initial results were suboptimal.

- **Small Training Dataset (525 examples):** To assess the model's performance with minimal training, we fine-tuned it on a small subset of 525 examples. The model successfully generated complete and meaningful sentences, though some outputs contained noise. For instance:

  > *"==================It's completely normal to feel nervous or shaky when attending therapy sessions, especially if you're new to the process. Many people experience anxiety . . . "*

- **Full Training Dataset (4719 examples):** Separately, we fine-tuned the model on the entire dataset of 4719 examples to evaluate its performance with comprehensive training data. However,the model frequently failed to generate any response, and occasionally, it produced incoherent outputs consisting of random words or phrases such as:

  > *"with, that a the for is you and a to to what a you you, a you feel or to the to with. in in the you. to you,"*

These results suggest that while the model could handle minimal training data reasonably well, training with the full dataset introduced issues.

Root Cause Analysis Through extensive debugging, we identified the root cause of the inconsistent outputs: **misalignment between tokenized inputs and labels in the fine-tuning process**. As a decoder-only model, LLaMA 3.1 requires a unified sequence of input and output tokens to to compute loss effectively. Initially, we tokenized the user's context as *input_ids* and the therapist's response as *labels*. This approach disrupted loss computation, causing the model to misinterpret the training data.

Resolution To resolve this issue, we restructured the training data by concatenating the user's context and the therapist's response into a single sequence. This combined sequence was ued for both *input_ids* and *labels* during fine-tuning. After implementing this fix, the model stopped generating random words and began producing more coherent outputs, marking a significant improvement in its behavior. This iteration underscores the importance of aligning the input-output formats with the architectural requirements of decoder-only models to ensure effective fine-tuning.

5.2.2 Iteration 2: Using Special Tokens

Despite resolving tokenization misalignment in Iteration 1, unexpected behavior persisted during evaluation:

- **Small Training Dataset (525 examples):** The model failed to generate responses for $\sim$3% of examples, while $\sim$97% were contextually appropriate and task-aligned.

- **Full Training Dataset (4719 examples):** The model failed to generate responses for $\sim$70% of examples, with only $\sim$30% producing contextually appropriate outputs.

Given these inconsistencies, we hypothesized that our prompt design might not be well-suited for LLaMA's architecture. To test this, we restructured the prompt to include special tokens:

```
<system>You are a licensed
therapist. Respond to the user's
question with empathy.</system>
<user>Context</user>
<assistant>Response</assistant>
```

This structured prompt clarified roles and improved guidance. Although the quality of the outputs did not significantly improve compared to the previous iteration, adopting this approach aligns with best practices for fine-tuning decoder-only models like LLaMA. It also provides a solid foundation for future experimentation and refinement.

5.2.3 Iteration 3: Investigating Empty Responses

To address the inconsistency where some examples yielded no output, we analyzed the model's behavior. In a decoder-only model, a completely empty response suggests a flat or undefined probability distribution for token prediction. To verify this, we analyzed an example where the model generated no response. We passed the input context from this example through the fine-tuned model and extracted the logits from the model's output. These logits represent the model's raw confidence scores for predicting the next token. By applying a softmax function, we converted the logits into probability distributions, allowing us to examine the likelihood assigned to each token.

Contrary to our hypothesis, the model did generate tokens with valid probabilities. The issue was traced to a bug in the response generation function, where the decoding process mishandled the model's outputs, resulting in empty responses. After fixing this bug, the model consistently produced responses for all examples.

5.3 Evaluation Results and Discussion

From a qualitative perspective, the fine-tuned model consistently produced high-quality, empathetic, and contextually appropriate responses. However, its scores on evaluation metrics remained lower than the baseline LLaMA model.

Metric	LLaMA Baseline	LLaMA Fine-Tuned
BLEU	0.0139	0.0111
ROUGE-1	0.2675	0.2436
ROUGE-2	0.0388	0.0333
ROUGE-L	0.1251	0.1267
BERTScore (F1)	0.5208	0.5076

Table 2: Evaluation metrics for baseline and fine-tuned LLaMA models.

5.4 Time Constraints and Future Work

Given the time constraints of the project, we were unable to fully investigate the discrepancy between qualitative performance and quantitative scores. It is possible that another bug exists in the evaluation pipeline or that the current evaluation metrics do not adequately capture the nuanced empathy and contextual appropriateness of the responses. Although the project timeline limited our ability to dive deeper into these issues, the model's strong qualitative performance highlights its potential for real-world applications. Addressing the evaluation challenges and refining the fine-tuning process would be valuable directions for future work.

6 Conclusion

This study highlights LLaMA's strengths in delivering empathetic responses in mental health conversations, showcasing performance comparable to human-like therapists. However, the naturally empathetic tone of mental health dialogues raises questions about the model's adaptability in domains where empathy must be more subtly inferred, such as education or workplace environments. Can LLaMA and similar models detect subtle emotional shifts, adapt their tone dynamically, and balance informative and reassuring responses?

Language models are unequivocally integrating into our lives as trusted companions in information exchange, support, and decision-making. Ensuring that these systems maintain empathy is crucial to prevent undermining the humanity they are designed to reflect. While challenges remain, the progress thus far provides a solid foundation for future exploration, underscoring the promise of LLMs in transforming communication and support. And for now, it seems we are off to a pretty decent start.

References

[1] Yuyan Chen et al. *EmotionQueen: A Benchmark for Evaluating Empathy of Large Language Models*. Bangkok, Thailand, 2024. ACL Anthology: 10.18653/v1/2024.findings-acl.128 (cs.CL). URL: https://aclanthology.org/2024.findings-acl.128.

[2] Zainab Iftikhar et al. *Therapy as an NLP Task: Psychologists' Comparison of LLMs and Human Peers in CBT*. 2024. arXiv: 2409.02244

[cs.HC]. URL: https://arxiv.org/abs/2409.02244.

[3] Man Luo et al. *Assessing Empathy in Large Language Models with Real-World Physician-Patient Interactions*. 2024. arXiv: 2405.16402 [cs.CL]. URL: https://arxiv.org/abs/2405.16402.

[4] Y. H. P. P. Priyadarshana et al. *Prompt engineering for digital mental health: A short review*. 2024. DOI: 10.3389/fdgth.2024.1410947. Frontiers: 1410947. URL: https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2024.1410947.