MIRROR AI

LLM BASED MENTAL HEALTH ASSITANCE

Student Name: Rikzy Jezuli (IIT ID: 20210667 | RGU ID: 2122107)
Supervisor: Ms. Dileeka Alwis

INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

ROBERT GORDON UNIVERSITY ABERDEEN

Presentation Date:

Why

1. To find out how LLMs can be trained to provide therapeutic responses .
2. And how similar and effective they are compared to a human therapist.

PROBLEM STATEMENT

LLMs can act like therapists with prompt engineering—but we still don’t fully understand how to fine-tune them for mental health or evaluate their responses.

RESEARCH QUESTIONS

- Fine Tuned LLM Vs Human Therapist
- How Training Parameters Influences Model’s responses
- Prompt Engineering Vs Fine Tuning

PROJECT SCOPE

Researching existing AI mental health tools and identifying gaps

Fine-tuning open-source LLMs using mental health datasets

Analyzing performance, empathy, accuracy, and usability

Implementing a user-friendly interface for real time chat with the LLM

Significance of the Study

Utilizes LLMs to generate human-like therapeutic responses making mental health care accessible to many.

Advance research in mental health-specific LLM fine-tuning

Improve evaluation methods for AI-driven therapeutic tools

Contribute to safer, more effective AI companions for emotional wellbeing

RELATED WORK

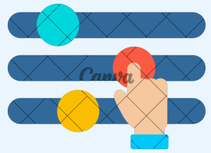
- **ChatCounselor (Liu et al., 2023)** – Fine-tuned LLM on counseling data; showed improved empathy but limited scalability.
- **CBT-LLM (Na, 2024)** – Domain-specific LLM using CBT principles, but limited to Chinese language context.
- **SERMO (Denecke et al., 2020)** – Focused on emotion regulation but lacked long-term engagement evidence.

EXISTING SOLUTIONS AND THEIR LIMITATIONS

- **Rely heavily on scripted or prompt-based interactions**
- **Limited personalization due to not leveraging LLMs**
- **Limited Interaction and Understanding**
- **Lack of Empathy and Emotional Intelligence**

Research Gap

Lack of Evaluation Standards for LLM-Based Mental Health Tools



Limited Research on Fine-Tuning vs. Prompt Engineering in Mental Health Contexts

Insufficient Diversity in Training Datasets

Lack of Real-World User Testing in AI Mental Health Tools

Data Sources

Primary Dataset

- Psych8k, published by EmoCareAI, contains 8000 client-therapist responses.
- The dataset is organized into three levels, each focusing on different key aspects of CBT, including basic knowledge recitation, cognitive model understanding, and therapeutic response generation.

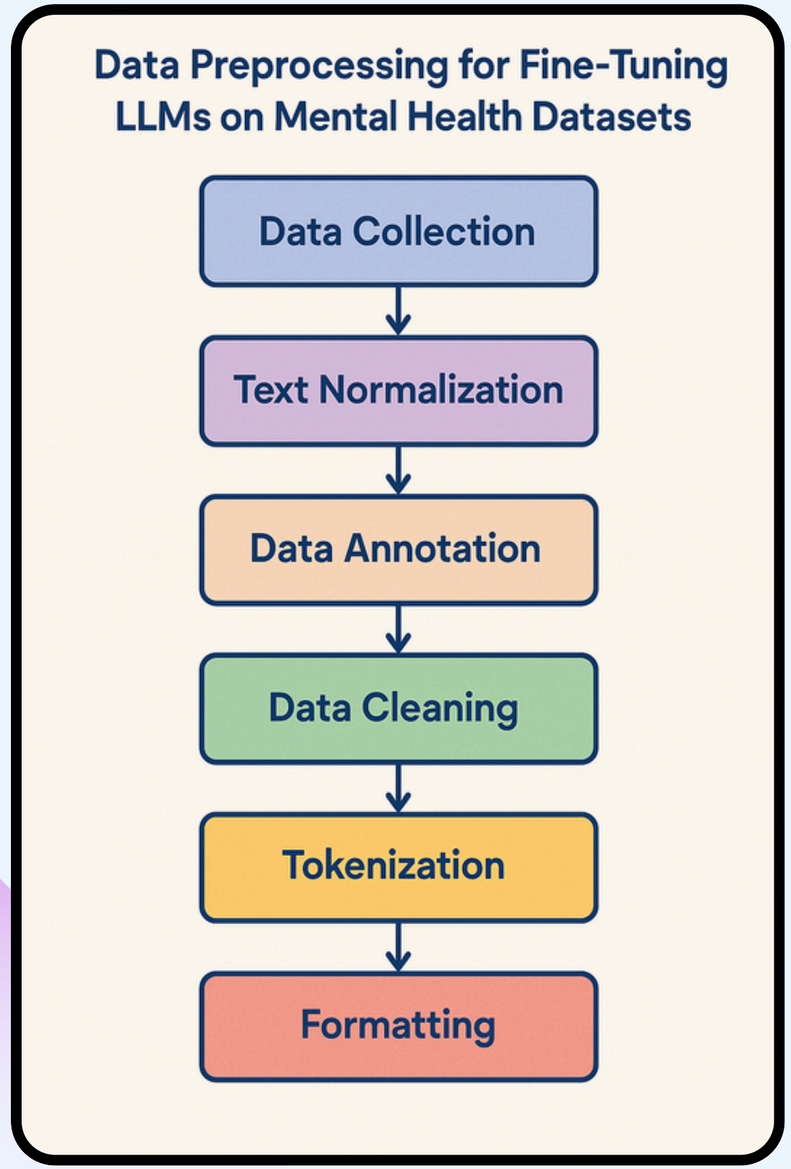
	A	B	C	D	E	F
1	instruction	input	output			
2	If you are a	Lately, I've	Thank you for sharing your experience wit			
3	If you are a	Lately, I've	Thank you for sharing your experience wit			
4	If you are a	Lately, I've	I appreciate you sharing your concerns wi			
5	If you are a	Lately, I've	Thank you for sharing your feelings and ex			
6	If you are a	Lately, I've	I appreciate you sharing your experiences			
7	If you are a	Lately, life	Thank you for sharing your concerns and			
8	If you are a	Lately, I've	It's important to pay attention to these fee			

Secondary Dataset

- •Synthetic dataset, generated using OpenAI API based on real patient input.
- Dataset contains 3000+ conversations which goes back and forth between client and therapist.
- Used to train the model highly engaging and empathetic but shorter responses that fits the CBT framework.


```
[
    {
        "conversations": [
            {
                "role": "user",
                "content": "I have applied for something and I'm not worried about it
            },
            {
                "role": "therapist",
                "content": "That's an interesting perspective. It's great that you're
'everything going well' would look like for you? How would that make you feel?"
            },
            {
                "role": "user",
                "content": "If everything went well, I would get the job, and I think
            },
```

Data Preprocessing



🔁 Role Normalization
Replaced "therapist" with "assistant" to align with LLM-compatible formats for chat-based training.

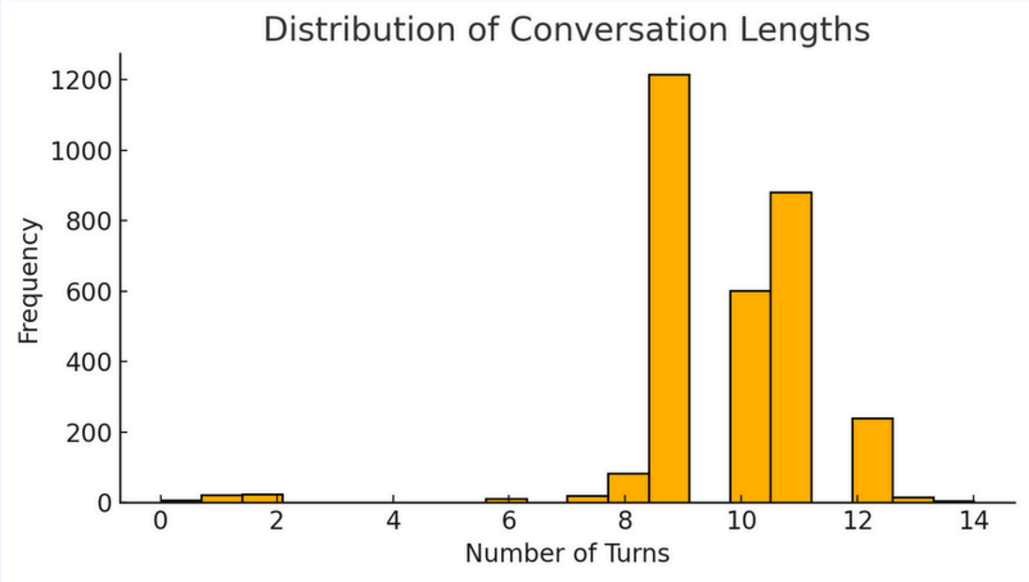
🧽 Data Cleaning
Skipped empty conversations.
Ensured only lists of structured turns (with role and content) are included.

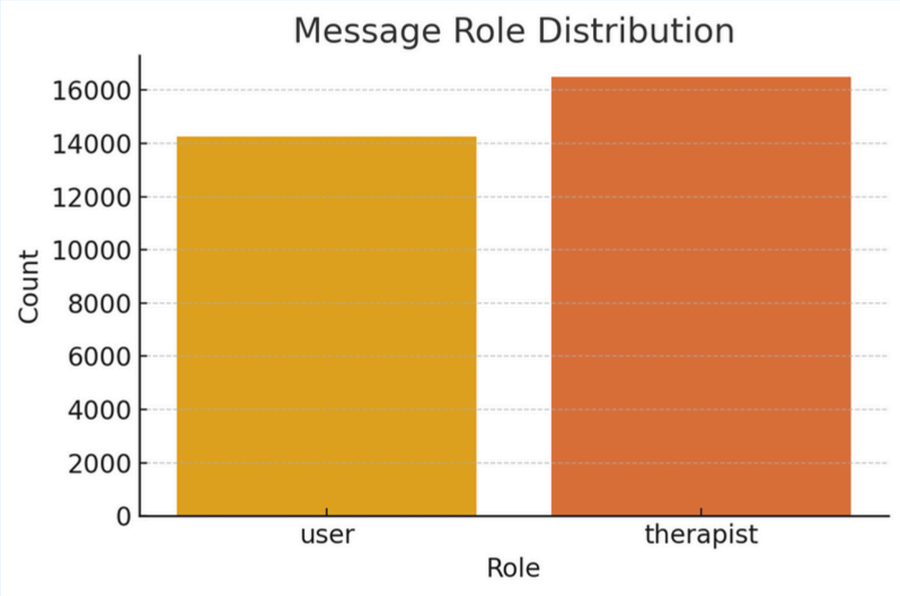
📐 Dataset Conversion
Converted the cleaned list of dictionaries to a HuggingFace Dataset format.

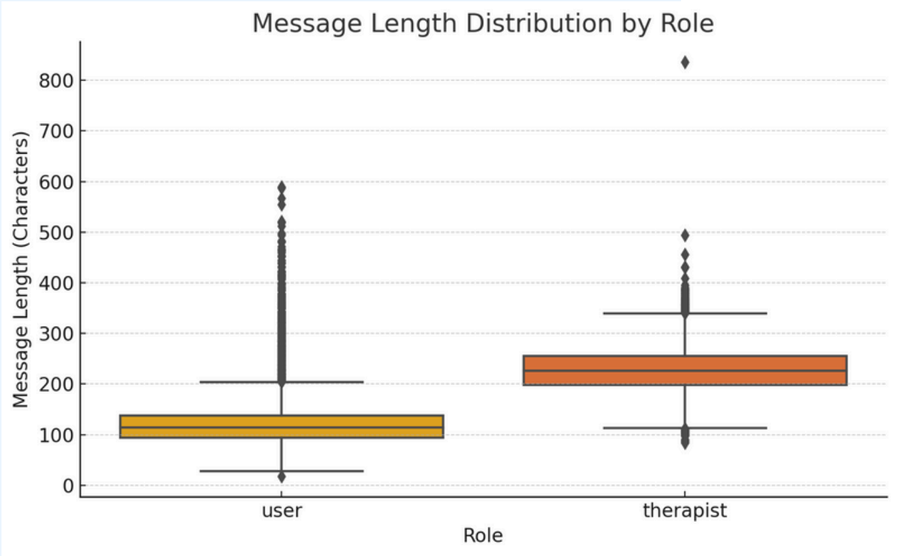
🧠 Chat Format Standardization
Used standardize_sharegpt() to align the conversation format with instruction-tuned models, simulating a chat-like context.

🗨️ Prompt Formatting
Applied a chat template using apply_chat_template()

Initial Data Insights







Model Development & Deployment

Base Model Selection
LLAMA 3.2 1B Vs LLAMA 3.2 3B

Finetuning Method
PEFT > LORA

Experimenting With LORA Rank
r=8, r=16, r=32, r=64

Experimenting With Learning Rates
1e-5, 2e-4, 5e-4

- Successfully implemented Parameter-Efficient Fine-Tuning (PEFT) with LoRA
- Deployed model on an online server and created an inference API
- Created a user-friendly chat interface for real time chat via API

Model Evaluation & Validation

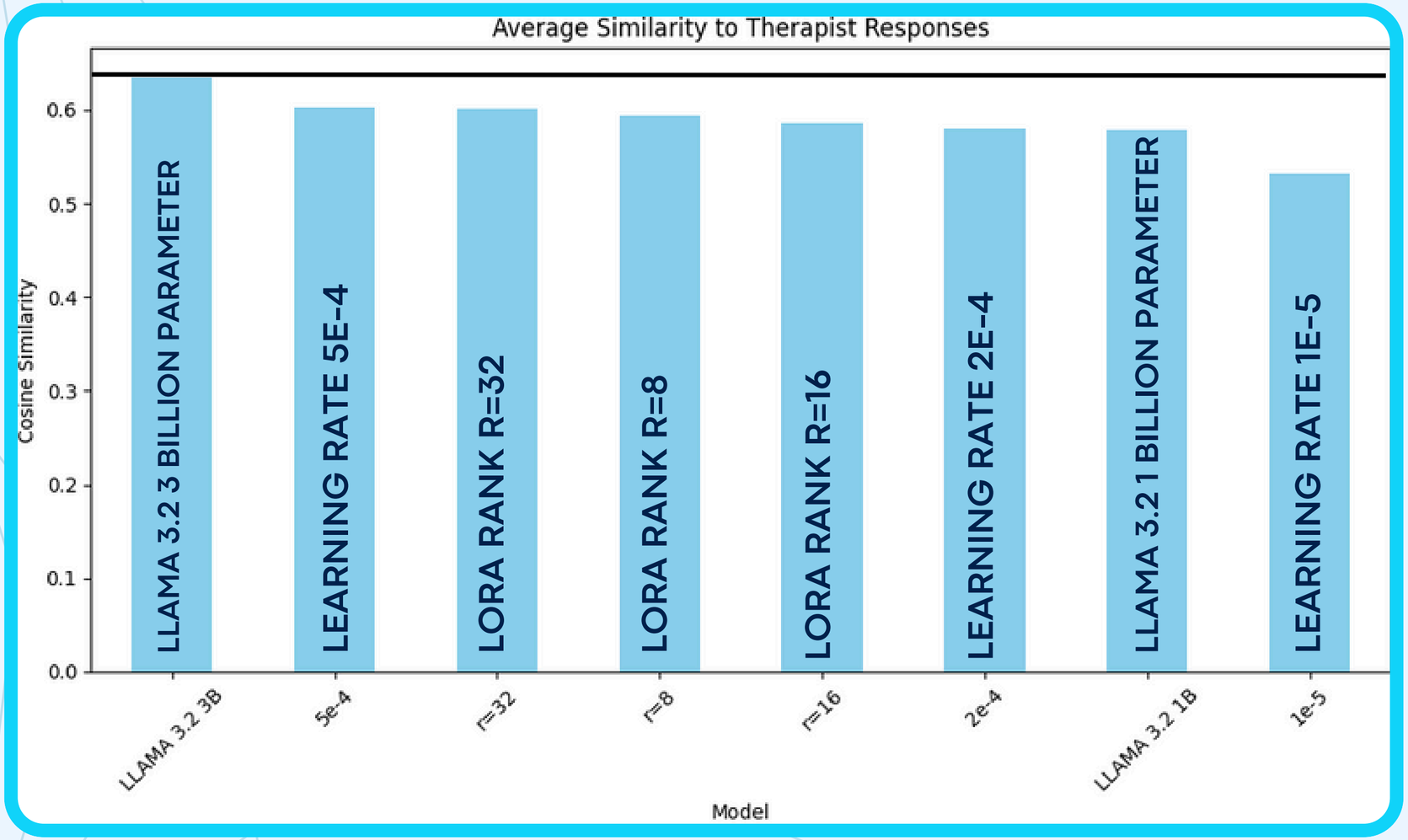
Evaluation Metrics

- Sematic Similarity
- Empathy scores

Performance Comparison

- **Model Size**
  - 1B Parameter
  - 3B Parameter
- **LORA Rank**
  - r=8
  - r=16
  - r=32
- **Learning Rate**
  - 1e-5
  - 2e-4
  - 5e-4

LLMs Responses Vs Therapist Responses



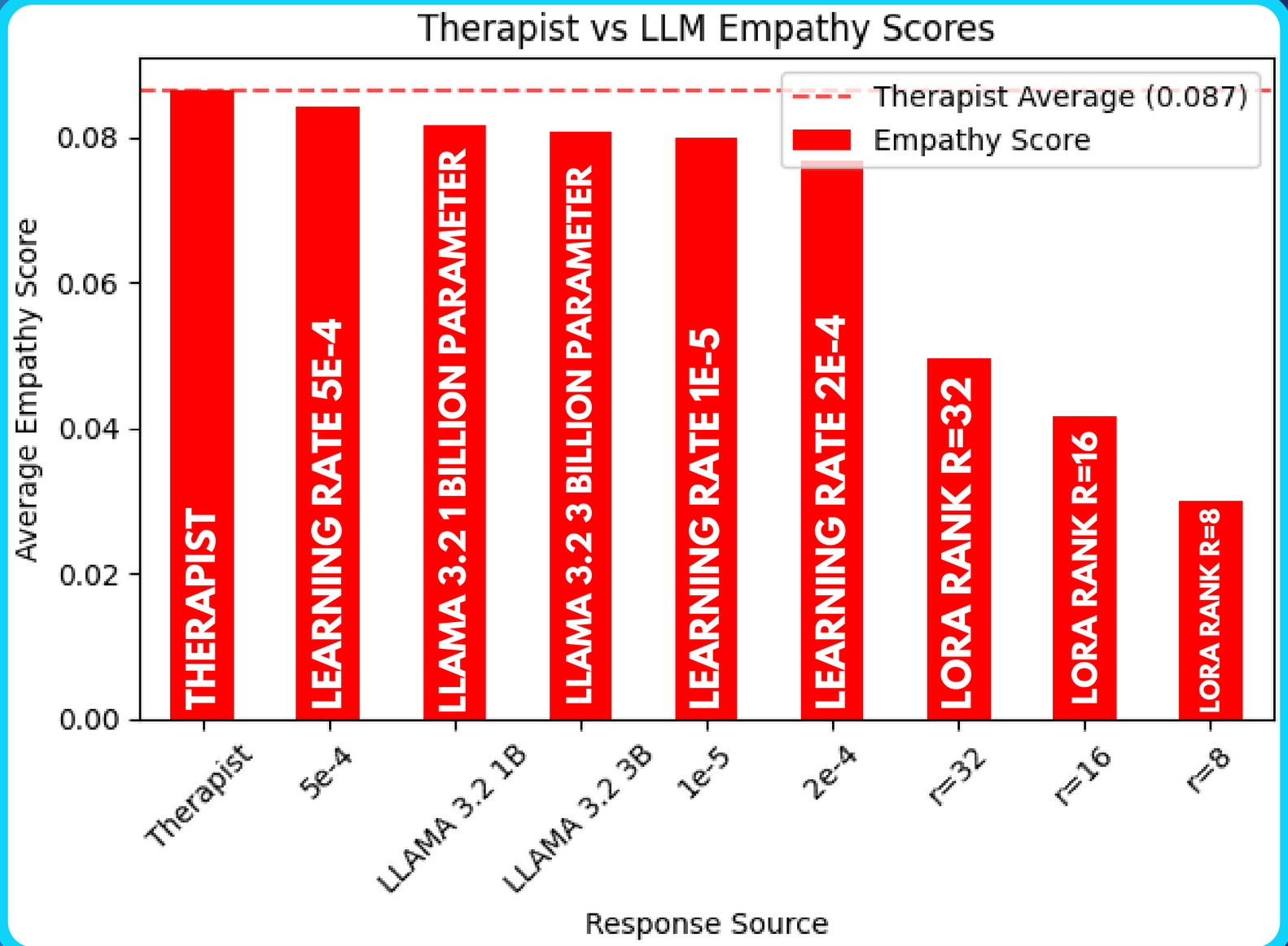
INSIGHTS

Larger models perform better

Higher learning rates (5e-4) produced better similarity scores

Diminishing returns with parameter adjustments

Empathy Scores



INSIGHTS

Human therapists maintain the highest empathy score

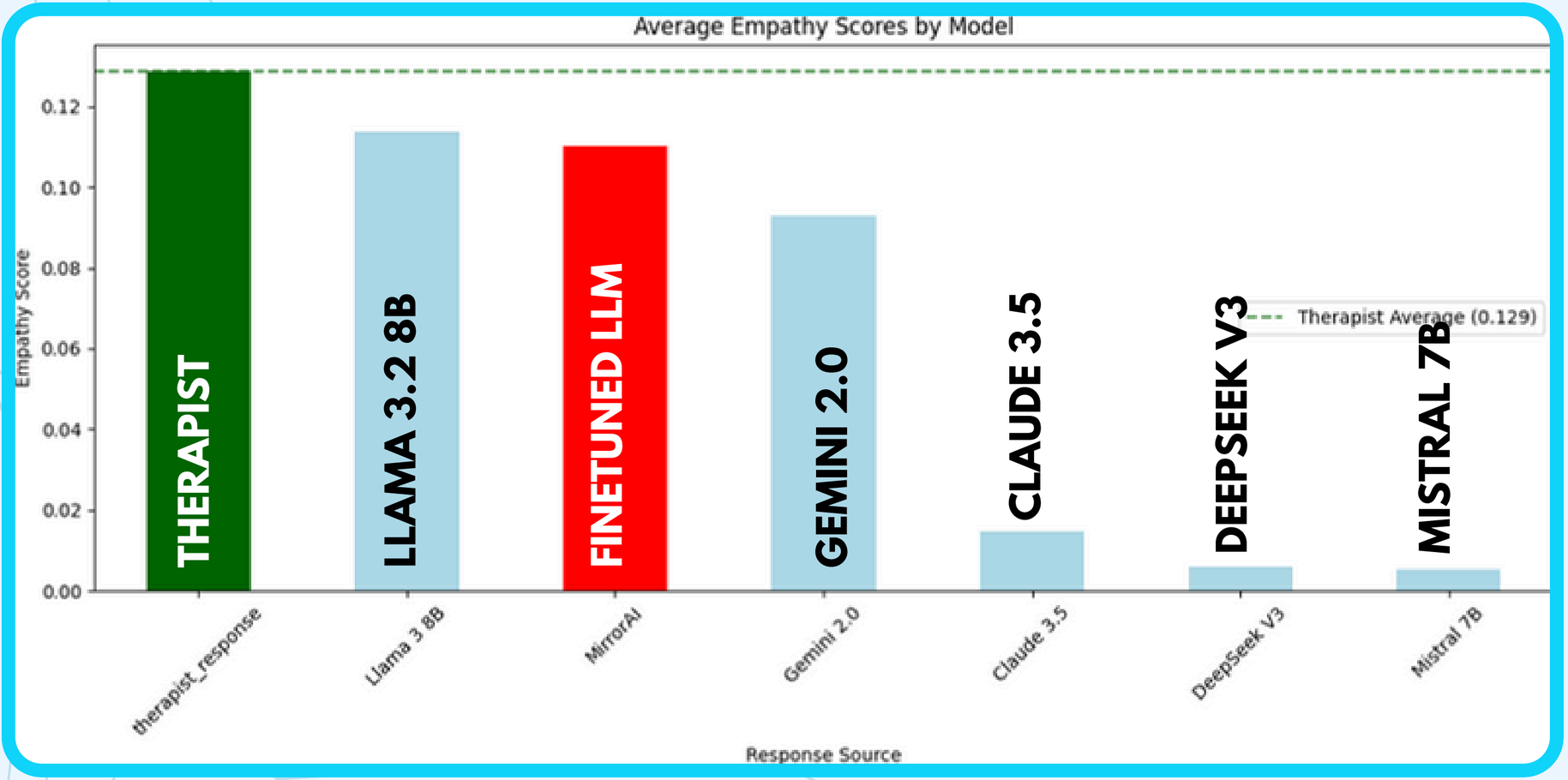
Higher learning rates (5e-4: 0.0841) better preserved empathetic responses.

Lower rank parameters dramatically damaged empathy

Smaller models performed slightly better than larger models on empathy scores

Empathy scores decline almost linearly as rank decreases.

Fine Tuned LLM Vs Other Open Source LLMs Empathy Scores



INSIGHTS

Fine Tuned LLM generated better empathetic reponses compared to other open source models that are not fine tuned.

Human therapist still outperformed LLMs

There's a notable trade-off between semantic similarity and empathy

Results and Discussion

KEY FINDINGS

🧠 Model size matters: 3B model outperformed 1B in similarity (+6%)

🧠 Higher learning rates preserved empathy better than lower ones

🧠 Rank parameter critical for empathy preservation (higher rank = better empathy)

CHALLENGES & LIMITATIONS

🧽 Trade-offs between model similarity and empathy scores

🧽 Parameter configurations affect different aspects of therapeutic quality

🧽 Fine-tuning techniques can significantly degrade emotional understanding

INTERPRETATION

❑ Large baseline models retain therapeutic capabilities better

❑ Careful parameter tuning necessary to maintain empathetic responses

CONCLUSION & CONTRIBUTIONS

KEY CONTRIBUTIONS

🧠 Demonstrated viability of open-source LLMs for mental health applications

🧠 Established evaluation framework for therapeutic responses (dual metric approach)

🧠 Identified optimal fine-tuning parameters for therapeutic language models

FUTURE DIRECTION

🧠 Multi-Modal LLMs for mental health support.

🧠 Develop standard benchmarks for evaluating mental health LLMs in empathy, safety, and helpfulness.

🧠 Expand to non-English datasets and culturally sensitive therapy styles.

THANK YOU!

Demo & QA