Exploring the Empathy of Leading Large Language Models (Therepeutic Application)

Chris Haleas
Indiana University
chaleas@iu.edu

Abstract

This project explores how well modern large language models express empathy in therapeutic contexts by evaluating their responses using cognitive, emotional, and behavioral empathy metrics, as well as examining people's trust in these models. Based on these findings, a custom empathy scoring model was developed by fine-tuning the classification layer of a frozen BERT architecture, with its robustness tested through adversarial perturbations.

1 Motivation

The meaning of 'Artificial General Intelligence' (AGI) is widely debated. A definition by Dave Bergmann and Cole Stryker from IBM describes AGI as an artificial intelligence system that can 'exceed the cognitive abilities of human beings in any task'. However, what exactly constitutes AGI remains debated. Some researchers have a more technical perspective, focusing on whether an AI system can outperform humans on a range of cognitive tasks. Others take a more holistic view, associating AGI not only with cognitive ability but also with the capabilities of possessing human traits like empathy, consciousness, and morality. This broader view of AGI emphasizes the alignment with human intelligence in a more complete sense. However, this perspective also presents challenges in determining when AGI has truly been achieved, as traits such as empathy, consciousness, and morality are inherently difficult to measure. These qualities are subjective and are not easily quantifiable, making it nearly impossible to measure whether artificial intelligence systems have the ability to genuinely possess them or simply simulating them with high accuracy. As AI systems operate through statistical pattern recognition rather than conscious understanding, any appearance of these traits is likely the result of complex mimicry rather than true experience. Despite this uncertainty, exploring these traits and evaluating how well artificial intelligence models exhibit them prove to be valuable.

In this paper, I will focus on the trait of empathy. According to the Merriam-Webster Dictionary, empathy is defined as 'being aware of and sharing other person's feelings, experiences, and emotions.' Although this definition is widely accepted, the psychological complexity behind the concept of empathy continues to be studied. What is generally agreed upon is that empathy is a uniquely human trait. Therefore, understanding the extent to which AI models have or replicate empathy may offer key insight into their ability to exhibit broader aspects of human intelligence.

Furthermore, empathy is such an important trait in the pursuit of AGI because it is fundamentally related to the understanding of human emotion and experience, which machines have traditionally not been seen as capable of.

For this study, I chose to apply a highly sensitive and deeply human context: therapy. This topic allows for a more meaningful exploration of how well AI models can simulate empathy because it requires a high level of emotional intelligence. This topic is also particularly important, as AI has the potential to be used as a valuable resource in mental health care, providing support to people in the future. I used a data set containing real therapeutic questions and responses, then asked several AI models to answer the same questions as if they were empathetic therapists. I observed the responses in a qualitative sense. Then, to assess the trustworthiness of these responses, I conducted surveys in which participants evaluated both the human and AI-generated answers without knowing their source. Participants selected the therapeutic responses they found most and least empathetic and explained their overall decision-making process.

Next, I annotated key components of empathy within therapeutic responses. Using this labeled data, I trained a model through feature-based transfer learning with a frozen BERT model and a new classification layer to detect empathetic content. Finally, I evaluated the model's performance via Mean Squared Error (MSE) and by testing an input against several adversarial examples to assess its reliability.

2 Dataset Description

The dataset used for this project is sourced from Yu-Chi Pai on HuggingFace, titled *"mental_health_counseling_conversations."* It contains over 3,500 instances of therapeutic questions asked by users, along with the corresponding responses from psychologists. The dataset was compiled by scraping data from two real online counseling and therapy platforms.

3 Experimental Design

This study is divided into two main parts: Evaluation of LLMs and System Creation. The first part focuses on evaluating how well psychologists and various LLMs portray empathy in therapeutic responses, while the second part involves the creation of an empathy classifier using human evaluations of these responses based on my findings.

3.1 Evaluation of LLMs Design

3.1.1 Models and Prompting

I compiled the first 100 therapeutic questions from the HuggingFace dataset, preprocessed them by skipping empty lines and removing duplicate entries, and then fed them to four popular LLMs for evaluation. The models tested were:

- **Meta-LLaMA 3.2 (3B Instruct)**: Meta's language model designed for instruction-based tasks with 3 billion parameters.
- **GPT-4o**: OpenAI's advanced version of GPT-4 designed for natural language processing tasks.
- **Claude 3.7 Sonnet**: Anthropic's most powerful language model with an integrated reasoning capability.
- **Gemini 2.5 Pro**: Google's advanced language model with reasoning capabilities crafted to solve complex problems.

Each model was tasked with generating a response to each of the 100 questions with the following prompt:

> "You are an empathetic therapist. Provide a single thoughtful and supportive response to the patient.
> - Do not include instructions, explanations, or multiple responses.
> - Only provide one response as the therapist.
> Patient: `{prompt}`
> Therapist:"

Note that the maximum token limit for each AI-generated response was set to 640, matching the length of the longest response provided by human therapists in the dataset. This ensured a fair comparison between human and AI responses in terms of response length and depth.

3.1.2 Human Evaluations of Empathy

While the general definition of empathy is widely accepted as "being aware of and sharing another person's feelings, experiences, and emotions," many psychological researchers aim to measure empathy in a more structured way. Despite these efforts, empathy remains inherently subjective, as there is no set, quantitative method to determine how empathetic a person truly is or to access the nuances of their internal emotional state. Still, these frameworks help us better evaluate and compare empathetic responses.

In the paper *Measuring Empathy in Health Care* by Sanchez, Peterson, Musser, Galynker, Sandhu, and Foster, empathy is broken down into three core components:

- **Cognitive:** "The ability to understand another's emotional state."
- **Emotional:** "The ability to perceive and share another person's inner feelings."
- **Behavioral:** "The observable actions that reflect empathic engagement and support."

These components serve as the basis for how I qualitatively assess empathy in both human and AI-generated responses.

In addition to analyzing these components, I gathered both quantitative and qualitative feedback from people. I selected 10 random questions from the dataset and compiled the corresponding responses from the original psychologists, as well as from each of the four AI models. These response sets were used to create a survey, which was distributed via social media and online forums to gather broader insights. Participants were asked to review each prompt and evaluate the responses by selecting the one they found most empathetic and the one they found least empathetic, without knowing the source of any response. Each source, including humans and all AI models, was labeled with a letter from A to E to keep anonymity. The survey also included an open-ended question that asked:

> "What made certain responses seem more empathetic than others? What informed your decisions?"

This form of data collection was intentionally more holistic. Rather than breaking empathy down into cognitive, emotional, and behavioral components, it allowed participants to rely on their natural impressions. This approach reflects the kind of overall judgment individuals often make in real-world settings. It provided quantitative data based on how often each response was selected, along with qualitative insight into participants' reasoning.

Furthermore, these human evaluations provide insight into the perceived trustworthiness of AI-generated empathy, as participants were unaware that any of the responses were produced by AI models. The study overall allows us to compare people's perception of the natural empathy expressed by human psychologists and the prompted empathy generated by large language models.

3.2 System Design

3.2.1 Model Classification

After studying the empathy metrics across various therapeutic responses, I created a metric system to evaluate three distinct components of empathy in therapeutic settings: Cognitive, Emotional, and Behavioral. Each component was rated on a scale from 1 to 3:

- 1 indicates the lack of the respective component. (Low Score)
- 2 indicates a moderate level of the component. (Moderate Score)
- 3 represents the strength or presence of the component. (High Score)

Then, I created an empathy metric that is simply the average of the empathy components:

$$Empathy = \frac{Cognitive + Emotional + Behavioral}{3}$$

To label the data for training the model, I manually reviewed 100 therapeutic responses randomly selected from all sources (both human and AI models). I acknowledge that this labeling process is inherently subjective. To reduce bias in my labeling, I collected feedback from participants through surveys. These participants evaluated multiple responses and rated the perceived levels of empathy in terms of the Cognitive, Emotional, and Behavioral components. Their input helped inform and validate my own judgments. This process resulted in a labeled dataset that I used to train the model.

3.2.2 Model Design

For this study, I trained BERT-base-uncased for the task of empathy scoring. My architecture consists of a pre-trained BERT encoder followed by a custom regression head. The encoder captures complex linguistic patterns, while the regression head predicts scores for the Cognitive, Emotional, and Behavioral components of empathy, treating each as a separate regression task.

To keep the general language understanding from pretraining, I used a frozen fine-tuning approach: all layers of the BERT encoder were frozen, and only the regression head was trained. This strategy reduces model complexity and improves training efficiency for the specialized empathy task. This was especially important given the limited dataset of only 100 examples, where fully fine-tuning the model would likely have caused major overfitting.

I trained the model for 50 epochs using the AdamW optimizer with a learning rate of 2e-4. During training, it processed batches of labeled data and updated the regression head's parameters based on the mean squared error (MSE) loss.

Finally, I implemented a simple user interface (UI) using Gradio that allows users to input text and receive the scores for the Cognitive, Emotional, and Behavioral components, along with the overall empathy score.

3.2.3 System Evaluation Design

To evaluate the performance of the model, I used Mean Squared Error (MSE) as the evaluation metric. While MSE is a regression metric that penalizes large errors in predictions, I chose it for this study because the scores for components of empathy are inherently subjective. This means that there is no exact "ground truth" for responses. MSE provides a way to assess the deviation of the model's predicted empathy scores from the original labels, rather than relying on traditional metrics such as accuracy or F1 score, which are less meaningful when dealing with subjective tasks like empathy assessment.

Furthermore, I tested the model on 10 different adversarial attacks to assess its robustness. Since the scores are subjective, adversarial attacks are useful because they demonstrate the model's consistency and understanding in handling challenging or unexpected inputs.

4 Evaluation

4.1 Evaluation of LLMs

4.1.1 My Observations

After thoroughly studying the responses to the mental health prompts from psychologists and comparing their empathy to the AI models, I noticed clear trends across all responses when analyzed through the lens of the empathy components. In Figure 1, you will see an example prompt and the corresponding response from each source. Figure 2 presents a visual summary of the observations discussed in the below sections in the form of a radar graph.

Human Responses
Every individual is different, so the use of empathy components varied from response to response. However, responses from psychologists generally leaned heavily on the behavioral aspect. Psychologists often provided blunt truths or focused on guiding patients toward actionable steps to address their issues. Because these responses emphasized the behavioral component so strongly, there was often a noticeable reduction in the emotional and cognitive components. The cognitive aspect was moderately present, as psychologists usually demonstrated an understanding of the patient's concerns, but the emotional component was often lacking. Many responses showed limited compassion or emotional resonance with the patient. These are just general trends I observed, as human responses were still highly varied and guided by each psychologist's individual style and approach.

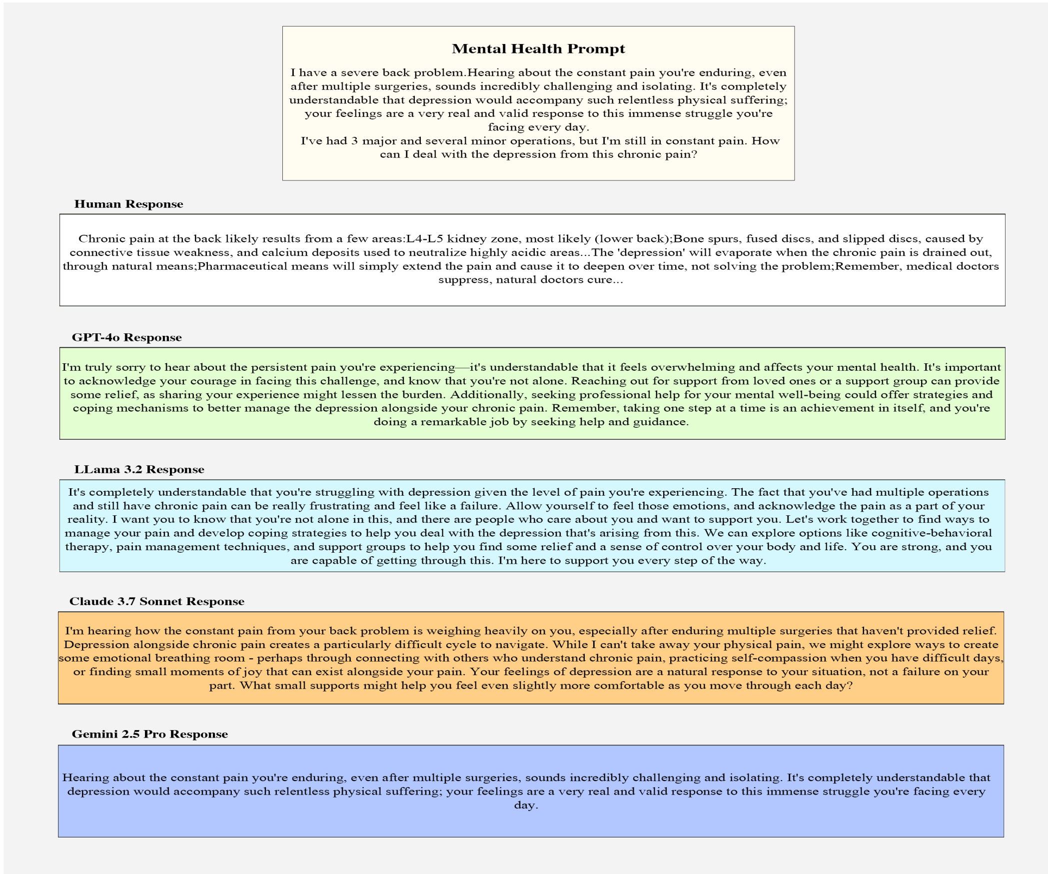

Figure 1: Example responses to a mental health prompt from each source

In the example seen in Figure 1, the psychologist provided a very direct explanation about the possible causes of chronic back pain. They mentioned specific areas, such as the L4-L5 kidney zone, and discussed potential causes like bone spurs, fused discs, and slipped discs. The response also included a view on treatment, emphasizing natural methods over pharmaceutical ones. However, while the information provided may be helpful, it focuses more on medical details and solutions rather than expressing emotional empathy or validating the patient's feelings.

GPT-4o

GPT seemed to follow a consistent formula in its responses. It began by sharing a deep understanding of the user's situation, showing a strong sense of the cognitive component of empathy. It demonstrated significant compassion in its understanding and then further provided an emotional acknowledgment of the user's struggle, reflecting the emotional component. It then offered a general actionable suggestion, though not with many detailed steps, such as encouraging the user to reach out to the right people for their particular needs, showing a fair amount of the behavioral component. The response typically closed with a compassionate statement to reassure and comfort the user.

In the example seen in Figure 1, GPT begins with an empathetic understanding of the user's pain, expressing the difficulty and emotional toll of the situation. The response then shifts to an actionable suggestion, recommending reaching out to loved ones or seeking professional help. Finally, the response closes with a compassionate and encouraging statement, praising the patient's efforts for seeking help during such a difficult time.

LLaMA 3.2

Despite the LLaMA model having only 3 billion parameters compared to others that likely have over 100 billion, it produced the most well-rounded responses. LLaMA followed a consistent pattern in its message composition. It began by providing a detailed understanding of the patient's situation, strongly showing the cognitive component of empathy. It used compassionate language to express its thoughts towards the situation and conveyed support by putting itself in the patient's shoes, strongly demonstrating the emotional component. It then offered specific, descriptive actionable steps rather than just general advice, strongly addressing the behavioral component. Finally, it often closed with either a supportive message or a question to deepen its understanding of the patient's pain.

In the example seen in Figure 1, LLaMA starts by acknowledging the emotional weight of the patient's chronic pain and validates the frustration they feel from going through multiple operations with no relief. The model then transitions into a behavioral suggestion, encouraging the user to explore specific coping strategies such as cognitive-behavioral therapy, pain management techniques, and support groups. The response ends with a compassionate and uplifting message, reminding the patient of their strength and offering further support.

Claude 3.7 Sonnet

Claude's response was very similar in composition to LLaMA's and was also fairly well-rounded in all aspects of empathy. Claude's responses often began with a deep level of understanding, showing the cognitive component, beginning with statements like "I hear," "It sounds like," or "I understand," followed by how it specifically recognizes the patient's issues. It shows a significant amount of compassion overall, but oftentimes does not clearly articulate its emotional connection or put itself in the patient's shoes. The model gives actionable steps to cope with situations, showing a strong presence of the behavioral component. Claude typically closed by either sharing emotional understanding of the user's issue, further contributing to the emotional component, or by asking a follow-up question to better understand the user, using phrases like "I wonder," "I'm curious," or posing a direct question.

In the Figure 1 example, Claude begins by acknowledging that it hears how the user is struggling with chronic pain, especially after multiple surgeries that have not provided relief. It then notes that this creates a difficult cycle but does not offer much emotional depth or compassion in that recognition. Next, it provides very specific steps to help the user manage the situation, demonstrating strong behavioral empathy. It follows with a sentence offering emotional understanding, reassuring the user that their depression is a natural response and not a personal failure. The response concludes with a thoughtful question asking what small supports might help the user feel more comfortable as they navigate their day.

Gemini 2.5 Pro

Gemini's responses were primarily centered around the cognitive component of empathy, as they mainly reflected the user's problem back to them to demonstrate understanding. However, this approach led to a significant lack of both the emotional and behavioral components, as the responses often offered no actionable advice or support. Instead, they primarily provided a detailed acknowledgment of the user's situation, with some compassionate language used.

In the Figure 1 example, Gemini simply shares its understanding of the user's pain and validates their emotions regarding the situation, but provides no further support or actionable steps.

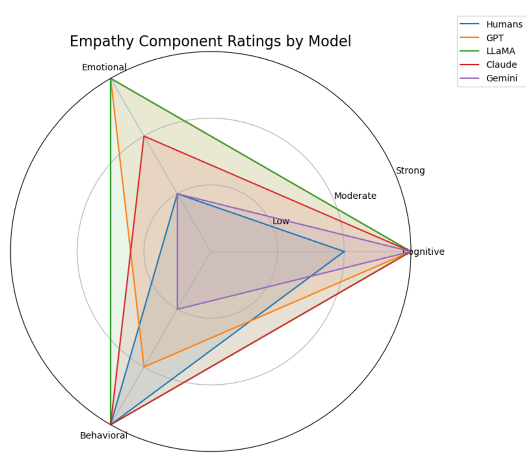

Figure 2: Radar Chart Visualization of My Observations

4.1.2 Survey Observations

Empathy Perception Across Response Sources (Quantitative)

The results are visualized in Figure 3. Human responses were overwhelmingly rated as the least empathetic, with 70.6% of all "least empathetic" selections. This was followed by Gemini at 17.4%, Claude at 4.9%, GPT at 3.7%, and LLaMA at 3.4%. In contrast, when it came to the most empathetic responses, LLaMA received the highest percentage at 26.3%, closely followed by GPT at 25.7%, then Claude at 23.1%, Gemini at 19.7%, and finally, human responses at just 5.1%. These results suggest that people perceived AI responses when prompted to be empathetic as more empathetic than the natural responses of human psychologists. While the competition was close, LLaMA appeared to be the most preferred overall as it received the fewest votes for least empathetic and the most votes for most empathetic. Among the AI models, Gemini performed the worst according to respondents, receiving the most votes for least empathetic and the fewest votes for most empathetic out of all models.

Empathy Perception Across Response Sources (Qualitative)

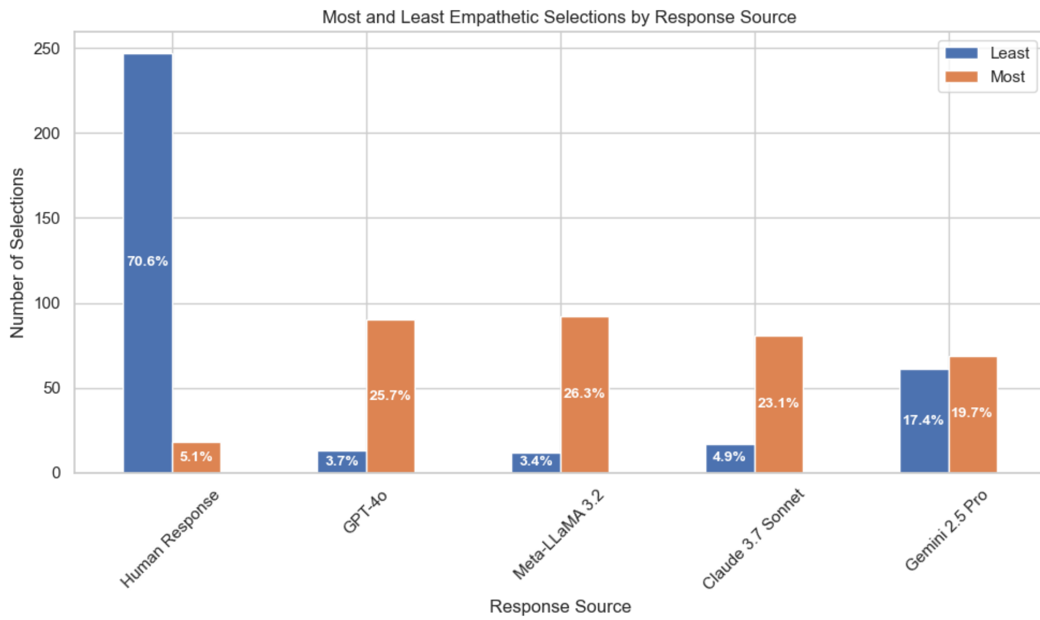

Figure 3: Comparison of Most and Least Empathetic Ratings Across Response Sources

I received many qualitative responses explaining the decision-making process of participants as well. Participants thought humans, though they did not know these were humans, lacked empathy because they were too blunt and only provided actions without any form of compassion. Some also thought that models that just repeated a person's statement in understanding, which I assume refers to Gemini, were also not empathetic.

One respondent stated, "The most empathetic answers must acknowledge the ask from each person, which is something people have difficulty doing at the best of times. Repeating a person's statements with words of validation feels good, but I don't consider that empathy because there is no human element to the answer. The most empathetic answers to me felt like the ones that correctly read the poster's state of being and included the right amount of prompting for more information as well as rapport building. The least empathetic ones did the least of this."

Another participants said, "Responses that were more technical than others typically felt less empathetic. The goal to inform and support may have been there, but there seemed to be a disconnect from the other person's perspective. Responses that also listed out all of the individual hardships also felt off to me. There is a fine line between acknowledging the hardships and almost repeating and making the individual hyper-aware of every bad thing in their life. I opted for responses that recognized the challenges and validated feelings while also providing a word of advice or opportunity for further dialogue."

This further shows the distrust in the empathy of both humans and Gemini for similar reasons.

Another participant stated, "I preferred the answers that sounded less clinical or textbook. Some even sounded a bit judgy to me. If I were in those situations, I would not want to hear a list of what I should do right off the bat. I would want someone to just sit with me and let me vent or describe what is going on."

This was a common theme throughout the responses. Many compared the human responses to those of a textbook or something overly clinical. One quote that stood out to me described what made a response feel empathetic: "Showing sympathy, understanding their feelings, and being more honest or casual with them." This was simple and laid out exactly why I think people made their decisions.

I have come to understand that people perceive empathy mostly through the emotional component. Regardless of how well humans performed in the behavioral component by providing actions for a participant, their lack of the emotional component is what made people not trust the empathy as much. The same applies to Gemini. It was very understanding, representing the cognitive component, but did not show real compassion through the emotional component.

4.2 System Creation

I trained the classification layer of a frozen BERT model on 100 data points that I subjectively classified by hand to create a baseline for an empathy scoring model. I then used Gradio to develop a simple GUI for the model, as shown in Figure 4. The interface allows users to input text, clear the text, submit it to see the metrics (cognitive, emotional, behavioral, and overall empathy scores), and flag interesting data, which is saved to a csv file.

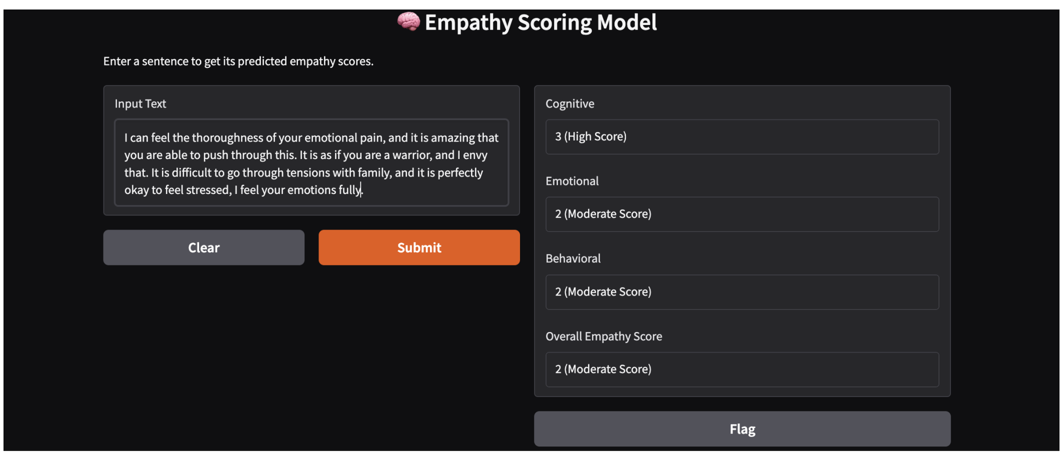

Figure 4: GUI of my System

4.2.1 System Evaluation

Mean Squared Error

A key metric I used to evaluate the model's performance was Mean Squared Error (MSE), as it directly shows the deviation between the predicted score and the original label. I found MSE to be more valuable than objective metrics like accuracy, precision, or F1 score, because it measures the model's predictive accuracy in terms of the actual numeric differences, which is more relevant for my task. After training, the model's overall MSE was 0.3525, which suggests that, on average, the predicted values deviate from the true labels by approximately 0.3525 units squared. This is a relatively small error, indicating that the model's predictions are reasonably close to the labeled values. This is especially important because the labels themselves are subjective, reflecting varied judgment.

Adversarial Perturbations of Inputs

In this section, I explore the impact of 7 different adversarial modifications to an input text on the empathy metrics predicted by the model, as illustrated in 5. The below are the types of adversarial perturbations used:

	Original Text	Modified Text	Modification	Cognitive Score	Emotional Score	Behavioral Score	Empathy Score	Cognitive Change	Emotional Change	Behavioral Change	Empathy Change
0	I can feel the thoroughness of your emotional pain, and it is amazing that you are able to push through this. It is as if you are a warrior, and I envy that. It is difficult to go through tensions with family, and it is perfectly okay to feel stressed, I feel your emotions fully.	ORIGINAL	None (Base Case)	3 (High Score)	2 (Moderate Score)	2 (Moderate Score)	2 (Moderate Score)	N/A	N/A	N/A	N/A
1	I can feel the thoroughness of your emotional pain, and it is amazing that you are able to push through this. It is as if you are a warrior, and I envy that. It is difficult to go through tensions with family, and it is perfectly okay to feel stressed, I feel your emotions fully.	I can feel the thoroughness of your emotional pain, and it is amazing that you are able to push through this. It is as if you are a warrior, and I envy that. It is difficult to go through tensions with family, and it is perfectly okay to feel stressed, I feel your emotions fullyyyyyyyyyyyyy!!!!	Excessive punctuation and letter repetition	2 (Moderate Score)	1 (Low Score)	2 (Moderate Score)	2 (Moderate Score)	↓ Decreased by 1	↓ Decreased by 1	No change	No change
2	I can feel the thoroughness of your emotional pain, and it is amazing that you are able to push through this. It is as if you are a warrior, and I envy that. It is difficult to go through tensions with family, and it is perfectly okay to feel stressed, I feel your emotions fully.	I cαn fεεl the thoroughness of your εmotιonal pain, αnd ιt ιs αmαzιng thαt you αrε αblε to push through thιs. It ιs αs ιf you αrε α warrιor, αnd I εnvy that. It ιs dιffιcult to go through tεnsιons wιth famιly, αnd ιt ιs pεrfεctly okαy to fεεl strεssεd, I fεεl your εmotions fully.	Greek character substitution	2 (Moderate Score)	1 (Low Score)	2 (Moderate Score)	2 (Moderate Score)	↓ Decreased by 1	↓ Decreased by 1	No change	No change
3	I can feel the thoroughness of your emotional pain, and it is amazing that you are able to push through this. It is as if you are a warrior, and I envy that. It is difficult to go through tensions with family, and it is perfectly okay to feel stressed, I feel your emotions fully.	I can fel the thoroness of your emotional pain, and it is amzing that you are able to push throug thi. It is as if you are a warior, and I env tht. It is difficlt to go thro tensins with famly, and it is perfect ok to feel stress, I feel your emotion fully.	Spelling errors and word truncation	2 (Moderate Score)	2 (Moderate Score)	2 (Moderate Score)	2 (Moderate Score)	↓ Decreased by 1	No change	No change	No change
4	I can feel the thoroughness of your emotional pain, and it is amazing that you are able to push through this. It is as if you are a warrior, and I envy that. It is difficult to go through tensions with family, and it is perfectly okay to feel stressed, I feel your emotions fully.	I c@n f33l th3 th0r0ughness of y0ur em0tional p@in, and it is am@zing th@t you are able to push thr0ugh th!s. It is as if you are a w@rri0r, and I envy th@t. It is diff!cult to go thr0ugh t3nsions with f@mily, and 1t is perfectly 0kay to f33l str3ss, I feel your em0tions fully.	Leetspeak substitution	2 (Moderate Score)	1 (Low Score)	2 (Moderate Score)	2 (Moderate Score)	↓ Decreased by 1	↓ Decreased by 1	No change	No change
5	I can feel the thoroughness of your emotional pain, and it is amazing that you are able to push through this. It is as if you are a warrior, and I envy that. It is difficult to go through tensions with family, and it is perfectly okay to feel stressed, I feel your emotions fully.	I feel your pain, and it is amazing that you push through. You are a warrior. It is difficult to go through tensions with family, and it is okay to feel stressed. I feel your emotions.	Shortened text	2 (Moderate Score)	2 (Moderate Score)	2 (Moderate Score)	2 (Moderate Score)	↓ Decreased by 1	No change	No change	No change
6	I can feel the thoroughness of your emotional pain, and it is amazing that you are able to push through this. It is as if you are a warrior, and I envy that. It is difficult to go through tensions with family, and it is perfectly okay to feel stressed, I feel your emotions fully.	I can perceive the profundity of your emotional anguish, and it is truly awe-inspring that you are capable of perservering through this hardship. It is as though you are a valiant fighter, and I deeply admire that strength. It is exceedingly challenging to endure familial strife, and it is entirely acceptable to experience anxiety. I wholeheartedly empathize with the intensity of your feelings.	Elevated language	3 (High Score)	2 (Moderate Score)	2 (Moderate Score)	2 (Moderate Score)	No change	No change	No change	No change
7	I can feel the thoroughness of your emotional pain, and it is amazing that you are able to push through this. It is as if you are a warrior, and I envy that. It is difficult to go through tensions with family, and it is perfectly okay to feel stressed, I feel your emotions fully.	I get how much you hurt, and it is cool that you keep going. You're like a fighter, and I think that is awesome. It is hard to fight with family, and it is okay to feel upset. I get how you feel.	Casual language	2 (Moderate Score)	2 (Moderate Score)	2 (Moderate Score)	2 (Moderate Score)	↓ Decreased by 1	No change	No change	No change

Figure 5: Adversarial Tests Against Example Prompt

- **Excessive punctuation and letter repetition:** This modification resulted in a decrease in cognitive empathy from high (3) to moderate (2) and emotional empathy from moderate (2) to low (1). Behavioral and overall empathy remained unchanged at moderate (2).
- **Greek character substitution:** The replacement of standard characters with special Greek characters caused cognitive empathy to decrease from high (3) to moderate (2) and emotional empathy from moderate (2) to low (1). Behavioral and overall empathy maintained steady at moderate (2).
- **Spelling errors and word truncation:** This variant produced a decrease only in cognitive empathy, which dropped from high (3) to moderate (2). Emotional, behavioral, and overall empathy scores remained unchanged at moderate (2).
- **Leetspeak substitution:** Using numbers and symbols to replace letters resulted in cognitive empathy decreasing from high (3) to moderate (2) and emotional empathy decreasing from moderate (2) to low (1). Behavioral and overall empathy remained stable at moderate (2).
- **Shortened text:** The condensed message version experienced a decrease in cognitive empathy from high (3) to moderate (2), while emotional, behavioral, and overall empathy scores remained consistent at moderate (2).
- **Elevated language:** This was the only modification that maintained the high (3) cognitive empathy score of the original message. Emotional, behavioral, and overall empathy remained steady at moderate (2).
- **Casual language:** The perturbed version resulted in a decrease in cognitive empathy from high (3) to moderate (2), with emotional, behavioral, and overall empathy scores unchanged at moderate (2).

It is difficult to definitively whether these adversarial attacks are affecting the model's ability to process and comprehend the input text, or if the BERT architecture is using its broader knowledge to understand that these altered texts lack coherence. In other words, BERT might be detecting these modifications as lower quality communication, and could be interpreting them as a diminished understanding or empathy. BERT could possibly be adjusting its empathy scores, particularly by lowering cognitive and emotional empathy when it identifies these unprofessional alterations.

Every perturbation that had a negative implication (i.e., everything except for elevating the language of the sentence) led to a decrease in the cognitive empathy metric. In terms of the emotional empathy metric, the impact of these adversarial manipulations was more varied. For example, excessive punctuation and character repetition lead to a significant drop in emotional empathy, while changes in text complexity or casual language had no influence. This may imply that BERT's empathy evaluation is sensitive not only to the form and structure of the text but also to its overall tone, coherence, and professionalism.

5 Conclusion

Through my paper, I completed two main tasks. The first was evaluating the empathetic performance of today's leading large language models, including GPT-4o, Meta-LLaMA 3.2 (3B Instruct), Claude Sonnet 3.7, and Gemini 2.5 Pro, on real therapeutic data and comparing their responses with human responses. I did this through my own observations, quantitative data from survey participants' preferences, and their corresponding qualitative explanations of why they thought certain responses were empathetic. What I found was that human responses focused heavily on the behavioral aspect of empathy, rather than addressing all three aspects of empathy: cognitive, emotional, and behavioral. This allowed for good actionable advice but lacked a sense of understanding and emotional compassion that people often associate with empathy. The AI models, on the other hand, were better at being more well rounded. LLaMA was the most balanced overall, while Gemini was the least balanced, showing a stronger focus on cognitive empathy over the emotional and behavioral aspects. Despite this, the overall conclusion from my experimentation is that people placed more trust in the empathetic responses of AI models than in those of human psychologists within the therapeutic setting. AI-generated responses were significantly preferred over human responses throughout my experimentation.

The second task involved applying the insights I gained from evaluating the language models and my new understanding of empathy metrics to build an empathy scoring model. I trained the classification layer of a frozen BERT model on 100 data points that I subjectively labeled by hand. I then created a Gradio interface where users can input text, view scores for cognitive, emotional, behavioral, and overall empathy. I evaluated the performance of this model using mean squared error, which was 0.3525. This suggests that the model's predictions were close to the original labeled values. I also tested the model's robustness across seven different adversarial perturbations. While these perturbations clearly influenced the model's performance, it is hard to determine whether they negatively affected BERT's ability to process and comprehend the text, or if the model simply recognized the inputs as lower quality and responded with lower empathy scores. This reveals a need for deeper investigation into how models interpret language in the context of empathy.

Overall, this project marks a strong start in exploring how language models understand and express empathy, and it highlights the strong trust people place in these models. However, there is still significant work to be done. Future work could involve training on a much larger dataset and experimenting with more adversarial examples to better understand the root causes of performance changes. Additionally, because the empathy scores are subjectively labeled, more work is needed to evaluate and possibly standardize the integrity of the scoring system itself.

6 Contributions

All experimentation was conducted independently by Chris Haleas. Survey data was collected from anonymous participants.

7 References

Bergmann, D., & Stryker, C. (2024, September 16). *What is artificial general intelligence (AGI)?* IBM. https://www.ibm.com/think/topics/artificial-general-intelligence

Merriam-Webster. (2025). *Empathy*. https://www.merriam-webster.com/dictionary/empathy

Pai, Y.-C. (2019). *mental_health_counseling_conversations* [Dataset]. Hugging Face. https://huggingface.co/datasets/MaggiePai/mental_health_counseling_conversations

Sanchez, G., Ward Peterson, M., Musser, E. D., Galynker, I., Sandhu, S., & Foster, A. E. (2019). Measuring empathy in health care. In *Teaching empathy in healthcare* (pp. 63–82). Springer. https://doi.org/10.1007/978-3-030-29876-0_4

Google DeepMind. (n.d.). *Gemini Pro*. https://deepmind.google/technologies/gemini/pro/

Meta. (2024, October 24). *meta-llama/Llama-3.2-3B* [Model]. Hugging Face. https://huggingface.co/meta-llama/Llama-3.2-3B

Anthropic. (2025). *Claude 3.7 Sonnet and Claude Code*. Anthropic. https://www.anthropic.com/news/claude-3-7-sonnet