EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

GPT-4.0 as Mental Health Counselor

Rojan Abolhassani

Supervised by Ali Bahranian

April 6, 2025

Research Project Winter 2024/2025

GPT-4.0 as Mental Health Counselor

Rojan Abolhassani

Supervised by Ali Bahranian

1 Introduction

With the growing success of Large Language Models (LLMs) in a wide range of general tasks and their increasing accessibility, it is likely that many will attempt to utilize them as mental health counselors. However, given the sensitivity of mental health issues, a cautious and critical approach is essential. The deployment of Large Language Models in this domain requires rigorous evaluation to ensure they provide safe, empathetic, and unbiased support, while avoiding potential harm or misuse. In this project, we adopt a careful, systematic, and empirical approach to evaluate the performance of LLMs as emotional support agents. We aim to analyze the distribution of Emotional Support Conversation (ESC) strategies—a framework designed to address individuals' emotional distress through structured conversational techniques [3]—chosen by GPT-4.0 to identify potential biases and in the model's approach. Additionally, we seek to evaluate its proficiency and performance in providing effective and empathetic emotional support. To achieve this, we employ a classifier trained using the Bidirectional Encoder Representations from Transformers (BERT) model [1], which categorizes the support strategies embedded in the model's responses. Our results indicate that GPT-4.0 demonstrates a significant bias in its strategy selection, suggesting that it lacks proficiency as a well-rounded mental health counselor. These findings highlight the need for further refinement of LLMs before they can be reliably integrated into mental health support applications.

The remainder of this paper is structured as follows: In Section Preliminaries & Related Work 2, we provide an overview of related work, including studies evaluating the diagnostic capabilities of LLMs in mental health contexts [4] and their effectiveness as emotional support agents [2]. Section Dataset 3 introduces the dataset used in this study, detailing its composition, annotation process, and the distribution of counseling strategies. Section Methods 4 outlines our methodology, including the development of a BERT-based strategy classification model and the analysis of GPT-4.0's strategy selection. In Section Results and Observations 4.4, we present our findings, highlighting the classifier's performance and the biases observed in GPT-4.0's strategy distribution. Finally, Section Conclusion 5 summarizes our conclusions and Section Future Works 6 discusses potential future directions for improving the use of LLMs in mental health counseling.

2 Preliminaries & Related Work

2.1 Evaluating ChatGPT's Diagnostic Capabilities for Mental Health Disorders

Wishnia (2024) [4] explores the efficiency and validity of ChatGPT as a tool for mental health professionals and a self-diagnostic instrument for individuals. The study investigates ChatGPT's ability to diagnose various mental health conditions, including rare disorders, and examines whether it considers gender-specific symptomatology, interprets abbreviated symptom descriptions, and understands human-like quotes describing symptoms. By analyzing these factors, the research aims to provide insights into ChatGPT's effectiveness and reliability in mental health evaluations. The paper discusses three key limitations. First, the small sample size of 25 vignettes limits the study's statistical power and the ability to draw definitive conclusions about ChatGPT's effectiveness in mental health scenarios. Second, the use of ChatGPT for self-diagnosis poses risks, as self-diagnosis can lead to misdiagnosis, inappropriate treatment, and potential worsening of conditions. Third, while the study aimed to authentically represent patient experiences, the use of vignettes from research articles and textbooks may introduce biases or inconsistencies. In conclu-

sion, while the findings highlight ChatGPT's potential as a mental health diagnostic tool, the study emphasizes the need to address its limitations.

2.2 LLMs as Emotional Support Agents

Kang (2024) [2] explores the use of LLMs in the emotional support domain through an in-depth analysis of the ESConv dataset. The paper highlights several challenges faced by LLMs, such as difficulties in strategy selection and a strong, often detrimental, preference for specific strategies. Through a strategy-centric analysis, the study aims to understand why LLMs struggle with emotional support tasks and emphasizes the critical role that strategy selection plays in providing effective emotional assistance.

The findings reveal that LLMs lack robustness in predicting appropriate strategies across the three stages of emotional support: initial engagement, emotional validation, and problem-solving. In particular, struggles in one stage often hinder progress to the next, ultimately compromising the quality of the emotional support provided. These limitations highlight that LLMs, in their current form, are not inherently equipped to navigate the complexities of emotional support without external guidance or intervention.

The paper further demonstrates that LLMs' preference bias—where certain strategies are favored over others—aligns with the psychological Contact Hypothesis. This suggests that incorporating external assistance or feedback can help mitigate this bias, leading to improved strategy selection and overall emotional support quality. By reducing this bias, LLMs become more capable of providing consistent and appropriate emotional responses, thereby decreasing the likelihood of producing poor-quality or ineffective interactions. These findings provide a promising direction for enhancing the emotional intelligence of LLMs and highlight the potential benefits of combining AI with human oversight to better address the nuanced needs of individuals seeking emotional support.

3 Dataset

We have selected the *Amod Mental Health Counseling Conversations* dataset [1] for our training, testing and analysis. This dataset is publicly accessible and comprises 3,512 question-and-answer pairs collected from two online platforms dedicated to mental health counseling and therapy. The dataset does not include the ESC strategy for each response. To overcome this limitation, we manually annotated the support strategies in the responses and provided these annotations to GPT-4.0, prompting it to classify the strategies. The ESC strategies, along with their definitions and examples, were developed in consultation with a professional psychiatrist to ensure their validity and relevance. Below are the ESC strategies and definitions used for annotation in GPT-4.0.

- **Psychodynamic Therapy:** Aims to explore unconscious processes to achieve deep insight into personal issues, leading to long-term behavioral and personality changes.
- **Behavioral Therapy:** Focuses on modifying behavior through reinforcement, communication, and modeling to improve adherence to treatment plans.
- **Cognitive Behavioral Therapy (CBT):** Challenges and changes cognitive distortions to enhance emotional regulation and develop coping strategies.
- **Gestalt Therapy:** Emphasizes personal responsibility and present-moment awareness, integrating environmental and social contexts.
- **Humanistic Therapy:** Encourages self-awareness and reflection to foster healthier, more productive behaviors, often merging mindfulness with behavioral techniques.
- **Existential Therapy:** Centers on universal human truths such as death, freedom, isolation, and the search for meaning.

Figure 2 presents the results of this process along with an overview of the dataset, while Figure 1 showcases examples of the previously mentioned strategies.

Given the apparent imbalance in the dataset distribution, we applied data balancing techniques to ensure a fair and unbiased selection for training, improving the robustness of our model.

[1] https://huggingface.co/datasets/Amod/mental_health_counseling_conversations?doi=true

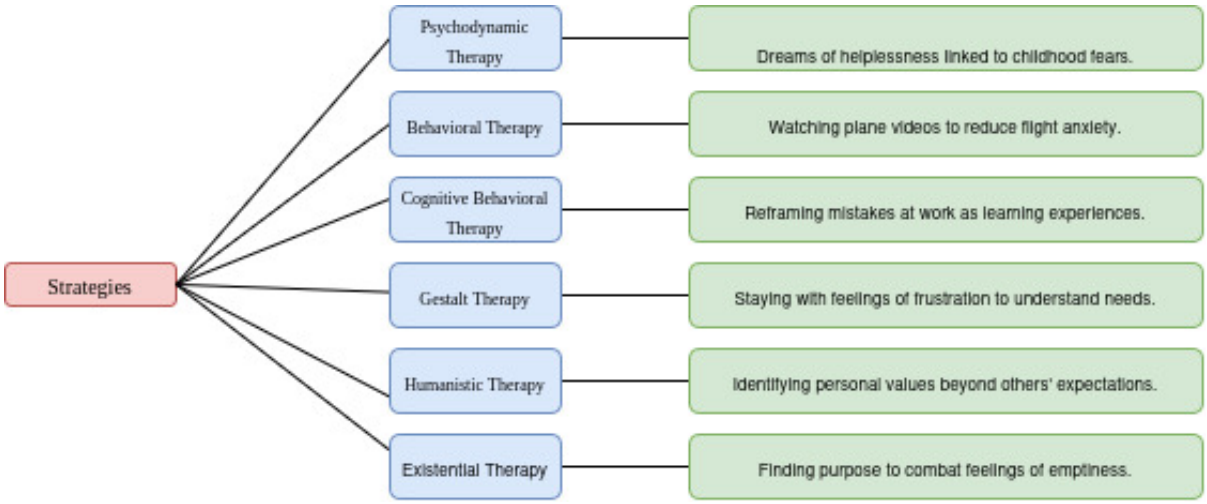

Figure 1: The following figure illustrates six primary ESC methods used in psychotherapy. Each method is tailored to address different aspects of emotional distress, behavioral patterns, and cognitive processes. The examples demonstrate how these strategies are applied in real-life counseling scenarios.

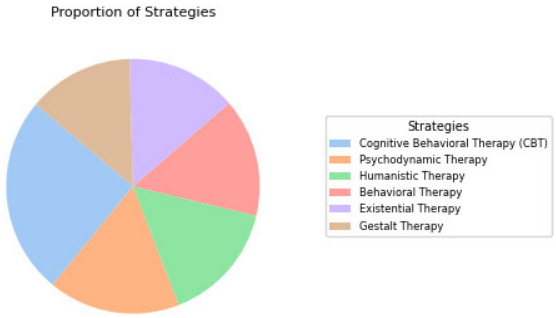

Figure 2: The figure illustrates the distribution of different strategies. It provides insights into the prevalence of each strategy.

4 Methods

4.1 Overview

To investigate the distribution and potential biases in strategy selection by GPT-4.0, we developed a classification framework capable of identifying counseling strategies from counselor responses. Our approach involved training a classifier to categorize responses into predefined strategy labels, leveraging a fine-tuned BERT-based model. We then applied this classifier to analyze the strategies employed by GPT-4.0 when generating responses within annotated contexts. Afterwards, we analyzed the strategy classifications generated by the classifier to create a distribution of the strategies employed by GPT-4.0, allowing us to examine the frequency and patterns of its strategy usage.

4.2 Strategy Classification Model

For the classification task, we utilized `BertForSequenceClassification` from the Hugging Face `transformers` library, initializing it with the pre-trained `bert-base-uncased` model. The model was fine-tuned to classify six distinct strategy labels. Additionally, we experimented with several other models, including DistilBERT, RoBERTa, and other variants. However, `bert-base-uncased` consistently outperformed all alternatives. To optimize performance, we carefully tuned the hyperparameters, adjusting learning rates, batch sizes, and dropout rates to achieve the best possible results. The final configuration of these hyperparameters is detailed in Table 1. Additionally, we employed techniques such as early stopping and model checkpointing to prevent overfitting and ensure robust generalization across different data splits.

4.3 GPT-4.0 Counseling Strategy Analysis

GPT-4.0 was first introduced to the different ESC strategies through annotations and explanations, along with several examples of each strategy. It

Table 1: Hyperparameters used for fine-tuning BERT

Hyperparameter	Value
Batch Size (Train/Eval)	32
Number of Epochs	15
Learning Rate	5×10^{-5}
Weight Decay	0.05
Warmup Ratio	0.1
LR Scheduler Type	Linear
Evaluation Strategy	Per Epoch
Save Strategy	Per Epoch
Mixed-Precision (fp16)	Enabled
Load Best Model at End	True
Best Model Selection Metric	Evaluation Loss

was then queried for responses, which were subsequently classified into one of the ESC strategies—meaning GPT-4.0 did not explicitly choose a strategy itself. This approach enabled us to create a distribution of GPT-4.0's ESC strategy selection, allowing for a direct comparison with the distribution of strategies used by human counselors. This comparison provides insights into potential differences in strategy selection, highlighting areas where GPT-4.0 aligns with or deviates from human counseling practices. Understanding these variations can help assess the effectiveness and biases of GPT-4.0 counseling responses.

4.4 Results and Observations

4.4.1 Classifier Performance

Our classifier achieved strong performance in accurately identifying the correct counseling strategy based on the given responses. Trained exclusively on the responses themselves—without access to any contextual information—the model effectively classified the applied counseling method. This demonstrates that the responses alone provide sufficient information for accurately identifying the applied counseling strategy. The detailed classification results, including performance metrics and strategy distribution, are presented in Figure 3, offering a comprehensive view of the model's effectiveness and areas for potential improvement.

4.4.2 Bias in GPT-4.0 Strategy Selection

Analysis of GPT-4.0-generated responses revealed a significant preference imbalance. As depicted in Figure 3, GPT-4.0 exhibited a strong bias toward the Humanistic Therapy method, selecting it in over 90% of cases. Furthermore, it utilized only four out of the six available strategies, neglecting the remaining two. This indicates a notable skew in GPT-4.0's approach, suggesting an inherent bias in its selection of support strategies.

These results highlight the need for further refinements in GPT-4.0's strategy generation process to ensure a more diverse and balanced application of support strategies. Addressing these biases can improve the model's effectiveness in delivering varied and contextually appropriate counseling responses.

5 Conclusion

Our results indicate that the standard GPT-4.0 model exhibits a significant bias toward certain strategies, specifically Humanistic Therapy, while neglecting Gestalt Therapy and Psychodynamic Therapy. This bias makes it unsuitable as a standalone replacement for a professional mental health counselor. The inherent uncertainty in identifying and evaluating the most appropriate therapeutic strategy—due to the complexities and subjectivity involved in psychological practice—further complicates the task of training a model to effectively handle such scenarios. Since mental health is a highly sensitive domain where incorrect responses can have serious consequences, we believe that the plain GPT-4.0 model is not suitable for this task in its current form.

6 Future Works

We believe that using GPT-4.0 to identify the strategy employed in creating our dataset may introduce inherent bias, potentially leading to incorrect strategy detection and the propagation of errors throughout the dataset. To ensure data integrity, we propose manually selecting the strategies with the assistance of an expert in the field. Additionally, expanding the dataset to include a more diverse range of human counselors would provide different perspectives and help broaden the overall understanding of strategy selection.

Furthermore, since the dataset contains multiple responses to a single prompt, there is no definitive way to determine which response is the most appropriate in a given context. This introduces uncertainty and randomness into our classification pro-

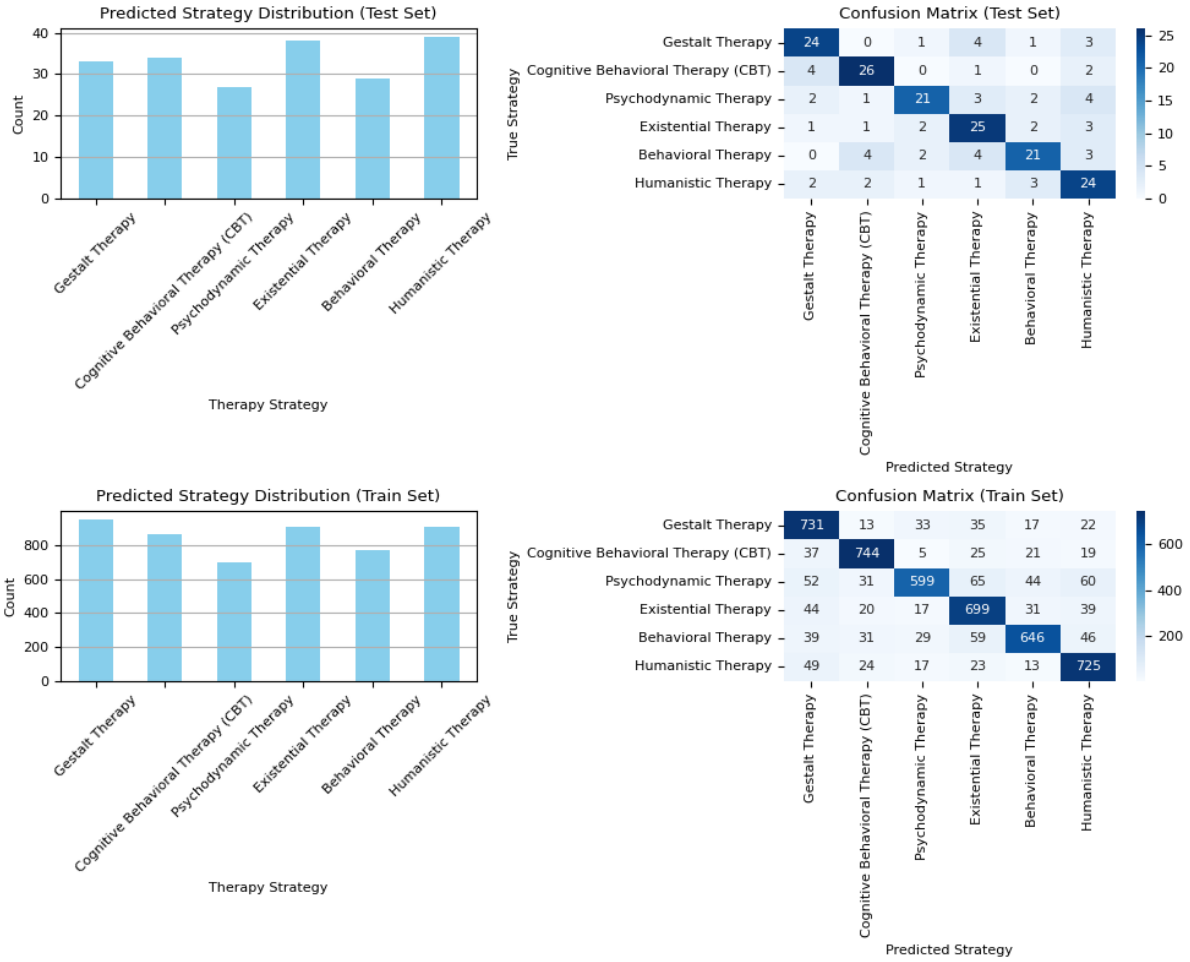

Figure 3: This figure illustrates the confusion matrix and the ESC method distribution of our classifier on both the training and test datasets separately.

cess. To mitigate this issue, it is crucial to incorporate performance metrics for evaluating strategies, such as measuring alignment with expert-labeled responses or assessing consistency across different classification models. Doing so would enhance the reliability and robustness of our classification framework, ensuring a more accurate analysis of strategy distribution.

Conducting the same experiment on different large language models (LLMs) could provide valuable insights. Since different models vary in their training data, architectures, and fine-tuning approaches, analyzing their strategy selection patterns could help determine whether certain biases or tendencies are unique to GPT-4.0 or prevalent across multiple models.

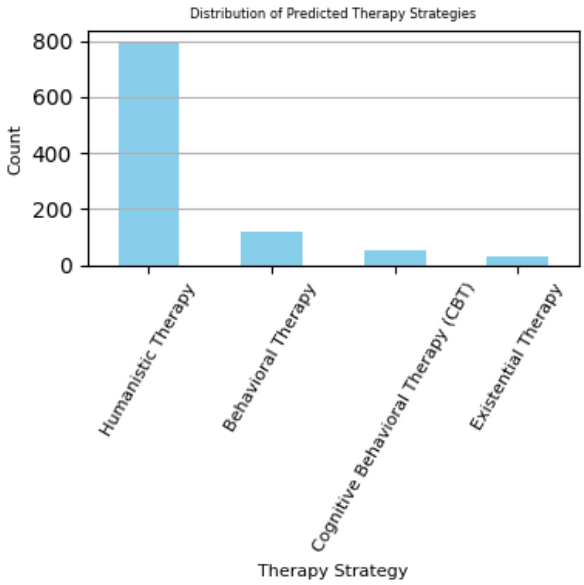

Figure 4: This chart illustrates the distribution of strategies GPT-4.0 selected when prompted.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019.

[2] D. Kang, S. Kim, T. Kwon, S. Moon, H. Cho, Y. Yu, D. Lee, and J. Yeo. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation, 2024.

[3] S. Liu, C. Zheng, O. Demasi, S. Sabour, Y. Li, Z. Yu, Y. Jiang, and M. Huang. Towards emotional support dialog systems, 2021.

[4] A. Wishnia, E. Rosenstreich, and U. Levi. Evaluating chatgpt's diagnostic capabilities for mental health disorders. *Clinical Reviews and Case Reports*, 3(6), 2024.