Evaluating Large Language Models for Enhancing Live Chat Therapy: A Comparative Study with Psychotherapists

Neha Deshpande, Stefan Hillmann, Sebastian Möller
Technische Universität Berlin
{n.deshpande, stefan.hillmann, sebastian.moeller}@tu-berlin.de

Abstract

Large Language Models (LLMs) hold promise for addressing the shortage of qualified therapists in mental health care. While chatbot-based Cognitive Behavioral Therapy (CBT) tools exist, their efficacy in sensitive contexts remains underexplored. This study examines the potential of LLMs to support therapy sessions aimed at reducing Child Sexual Abuse Material (CSAM) consumption. We propose a Retrieval-Augmented Generation (RAG) framework that leverages a fine-tuned BERT-based retriever to guide LLM-generated responses, better capturing the multi-turn, context-specific dynamics of therapy. Four LLMs—Qwen2-7B-Instruct, Mistral-7B-Instruct-v0.3, Orca-2-13B, and Zephyr-7B-Alpha—were evaluated in a small-scale study with 14 domain-expert psychotherapists. Our comparative analysis reveals that, in certain scenarios, LLMs like Mistral-7B-Instruct-v0.3 and Orca-2-13B were preferred over human therapist responses. While limited by sample size, these findings suggest that LLMs can perform at a level comparable to or even exceeding that of human therapists, especially in therapy focused on reducing CSAM consumption. Our code is available online. [1]

1 Introduction

Over the last few years, the rise of Large Language Models (LLMs) has led to significant advancements in different and varied applications of online therapy. Within the domain of therapy, Motivational Interviewing (Bundy, 2004) which is aimed at enhancing motivation to change problematic behaviors, and Cognitive Behavioral Therapy (CBT) (Beck, 1963), which focuses on altering negative thought patterns, are two notable techniques explored by therapy chatbots (Na, 2024; Bill and Eriksson, 2023). Some CBT-based chatbots, in particular, use LLMs to address user queries on online forums or through therapy platforms that connect users with mental health professionals (Na, 2024; Lai et al., 2023). While LLM-powered chatbots can deliver helpful, empathetic, and structured responses based on CBT principles, they tend to be limited to pre-defined queries and struggle with dynamic, multi-turn conversations as highlighted by (Berdowska and Zdanowicz-Cyganiak, 2024; Bell et al., 2019). These chatbots sometimes lack the empathy required for more specialized assistance, particularly for patients with severe mental health issues. The deployment of such fully automated mental health chatbots also raises significant ethical concerns, including risks of inadequate or harmful advice, exploitation of vulnerable users, and potential biases in recommendations (Khawaja and Bélisle-Pipon, 2023; Ke et al., 2024; Shen et al., 2024).

2 The Present Study

This study is part of a broader initiative exploring online psychological intervention via chat for individuals seeking to reduce or stop consuming Child Sexual Abuse Material (CSAM) (Gannon et al., 2023). In this intervention, conducted over four weeks, mental health professionals (referred to as "Therapists") guide individuals (referred to as "Patients") through weekly 50-minute sessions while monitoring CSAM use. This approach presents unique challenges, including ethical and legal concerns, stigma, and high dropout rates.

In this paper, we present an important aspect of the study: testing the use of an AI assistant to support therapists in chat-based therapy sessions for consumers of CSAM. The goal is not to replace therapists but to use AI as a supportive tool to reduce their cognitive load and enhance the overall therapy experience. Specifically, we

[1] https://git.tu-berlin.de/neha.deshpande/therapy_responses/-/tree/main

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 800–812
Aug 25–27, 2025.

focus on improving AI's human-like qualities by using ideal dialogue examples to generate therapist-like responses. Our approach aims to replicate key therapist behaviors—such as appropriate response length, open-endedness, balanced empathy, and effective emotional support—critical for maintaining patient engagement in live-chat sessions especially in the domain of CSAM prevention.

3 Related Work

3.1 LLMs for Mental Health Support

LLMs have been explored for mental health support, with studies demonstrating their ability to engage in therapeutic tasks. For example, (Cho et al., 2023) highlights LLMs' effectiveness in therapy for high-functioning autistic adolescents, while (Nie et al., 2024) presents a Conversational AI Therapist combining LLMs with smart devices for personalized mental health interventions like Cognitive Behavioral Therapy (CBT). Additionally, LLMs have been tested to simulate full therapy sessions, showing potential in delivering CBT-based responses (Xie et al., 2024; Lee et al., 2024). However, challenges remain, as LLMs often lack human-like empathy, collaboration, and cultural sensitivity (Iftikhar et al., 2024), which are essential for effective therapy. These limitations emphasize the need for human-AI collaboration, particularly to avoid "deceptive empathy" in AI-driven therapies (Syed et al., 2024). Efforts to address these shortcomings include enhancing LLMs with domain-specific knowledge, such as a CBT knowledge base implemented by (Yeom et al., 2024), which significantly improved the therapeutic utility of LLMs.

3.2 LLMs for Role-Playing and User Simulation

An emerging area of research involves using LLMs to simulate user personas, such as therapists or patients, to support mental health interventions. Prior studies have demonstrated how LLMs can role-play individuals with complex needs, enabling more tailored interactions in therapy settings (Sun et al., 2024). Similarly, research has explored LLMs as AI-driven patients for training novice therapists, introducing a principle-adherence prompting pipeline that improves response quality by 30% (Louie et al., 2024). Additionally, LLMs have been designed to simulate both doctor and patient personas, facilitating controlled experiments in healthcare dialogues (Chen et al., 2023).

Beyond healthcare, structured approaches to LLM-driven role-playing have been proposed to enhance persona consistency and realism. RoleLLM (Wang et al., 2023) introduces a multi-stage framework for role simulation, comprising role profile construction, context-based instruction generation, role-specific prompting, and fine-tuning. This method yields RoleLLaMA, an open-source model fine-tuned for role-playing, achieving performance comparable to GPT-4 in character-level benchmarks. Such advancements reinforce the potential of persona-based AI interactions for both therapist augmentation and therapist training.

3.3 Prompt Engineering for LLM-Controlled Dialogue

Prompt engineering plays a crucial role in shaping LLM responses for various applications, including mental health and therapy. Effective prompting strategies define the desired response structure, tone, and conversational flow, ensuring the AI adheres to professional and ethical guidelines (Marvin et al., 2023). Studies show that well-structured prompts can outperform fine-tuned models in tasks like conversational support and comment generation (Shin et al., 2023), although fine-tuning remains preferable for highly specialized tasks like clinical diagnoses.

Recent advancements include zero-shot, one-shot, and few-shot prompting to improve generalization across conversational tasks (Mann et al., 2020), as well as Chain-of-Thought (CoT) prompting, which enhances reasoning and coherence in multi-turn interactions (Wei et al., 2022). Furthermore, instruction tuning has been applied to optimize response control, particularly for empathy-driven tasks like mental health support (Ranaldi and Freitas, 2024).

Our work expands on the above mentioned techniques of carefully choosing a prompt that align LLM responses with professional therapist behaviors in this specialized therapy intervention, improving the effectiveness of AI-assisted interventions.

4 Methods

4.1 Overview

To preserve data privacy during therapy sessions, we restricted our use to LLMs that could run locally on our organization's infrastructure. Our

Dialog property	Value
Total utterances (turns)	15,918
Total words	477,232
Minimum turns per dialogue	10
Average turns per dialogue	49.74
Average words per utterance	30.60
Median conversation length (min)	57.42

Table 1: Overview of the properties for the 320 (after cleaning) therapist-patient dialogues in the dataset used in the study.

Property	Count
Number of context-response pairs	102362
Avg context word count	34.48
Max context word count	2,347
Min context word count	1
Avg response word count	20.07
Max response word count	189
Min response word count	1
Number of true labels (1)	10281
Number of false labels (0)	92081

Table 2: Summary of properties for the pairs of context-repsonses from the cleaned dialogue dataset with a label of either 0 or 1.

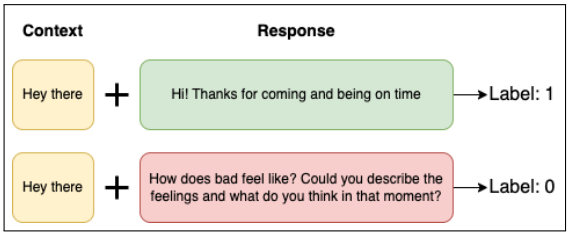

Figure 1: Example from the cleaned dataset after creating pairs of the contexts (1-5 utterances) and responses with their labels (0 or 1)

objective was to generate responses that emulate human-like therapeutic dialogue. However, we found that LLMs frequently struggled with multi-turn conversations, often producing close-ended replies that disrupted the natural flow of interaction. To mitigate this issue, we adopted a three-step retrieval-augmented approach inspired by Retrieval-Augmented Generation (RAG) (Lewis et al., 2020). First, a fine-tuned BERT-based model (Kenton and Toutanova, 2019) was used to retrieve relevant responses from a database of historical therapist–patient interactions. These responses were ranked using similarity, and the top 10 were provided as contextual input to the LLMs. This retrieval step helped ground the model's outputs, resulting in more coherent and context-aware responses. The BERT-based retriever was fine-tuned on a corpus of domain-specific therapist–patient conversations collected over a seven-month period. The following sections describe the dataset and pre-processing pipeline in detail.

4.2 Dataset and Pre-processing

The dataset consisted of 567 therapist-patient chat sessions, collected to monitor and reduce CSAM consumption. Sessions with fewer than 15 messages were excluded (indicating patient dropout), leaving 320 unique sessions, each lasting 50 minutes to 1 hour. Patient usernames were removed, and messages were rearranged to address overlapping responses. These multi-turn interactions were conducted via an anonymous online chat service. Table 1 summarizes the dataset properties.

The structure of our dataset is similar to the Ubuntu dialogue corpus (Lowe et al., 2015), both consisting of goal-oriented conversations and casual chit-chat, making them equally challenging to model (Muise et al., 2019). Both datasets are extracted from text-based conversations, further aligning their characteristics.

For model fine-tuning, dialogues were split into context-response pairs, with the context consisting of 1 to 5 randomly selected messages, and the subsequent therapist message as the response. From this, we created 102,362 context-response pairs, labeled as 0 (random responses) or 1 (gold standard responses). The dataset was split into training (92,126 pairs), testing, and validation sets (5% each). Detailed statistics are in Table 2, with an example in Figure 1.

4.3 Dialogue Retrieval and Ranking

For retrieval, we selected a BERT-based model (Kenton and Toutanova, 2019) called bert-base-uncased[2] that has already been fine-tuned and tested on dialogue datasets such as the Ubuntu dialogue corpus using different model architectures and fine-tuning techniques. We evaluated two such state-of-the-art approaches: Dial-MAE (Su et al., 2023) and Uni-Encoder (Song et al., 2021), both of which demonstrated strong performance in multi-turn dialogue retrieval when tested on the

[2] https://huggingface.co/google-bert/bert-base-uncased

Model	R10@1	R10@2	R10@5
Dial-MAE	0.918	0.964	0.993
Uni-Encoder	0.916	0.965	0.994

Table 3: Performance comparison of Dial-MAE (Su et al., 2023) and Uni-Encoder (Song et al., 2021) on the Ubuntu Dialogue Corpus

Model	R10@1	R10@2	R10@5	MRR
Dial-MAE (fine-tuned)	84.05	91.67	98.1	89.47
Dial-MAE (post-trained + fine-tuned)	**92.14**	**97.14**	**98.93**	**96.04**
Uni-Encoder (fine-tuned)	39.28	60.71	76.78	57.47
Uni-Encoder (post-trained + fine-tuned)	46.42	58.92	78.57	60.76

Table 4: Performance metrics of different models based on R10@1, R10@2, R10@5, and MRR.

Ubuntu dialogue corpus. Both approaches involve two stages: post-training and fine-tuning. Table 3 presents the retrieval accuracy of the models, as reported in their original papers, on the Ubuntu Dialogue Corpus. Performance is measured using R10@1, R10@2, and R10@5 metrics, which evaluate how often a relevant response appears among the top 1, 2, or 5 of the top 10 retrieved candidates. For instance, R10@1 indicates the percentage of times the correct response is ranked first. The two retrieval approaches used in our study are detailed below:

- The first approach, Dialogue Contextual Masking Auto-Encoder (Dial-MAE), a post-trained BERT-based model (bert-base-uncased) for dialogue retrieval (Su et al., 2023). We fine-tuned it using therapist-patient dialogue data, following the original methodology with a learning rate of 5e-5 for 5 epochs and a batch size of 64.

- The second approach, Uni-Encoder (Song et al., 2021), introduced Arrow Attention to optimize context-response interaction. We used the same BERT model (bert-base-uncased), post-trained and fine-tuned for 20 epochs with a batch size of 8 and a learning rate of 2e-4.

Both approaches were assessed using key metrics including R10@1, R10@2, R10@5, and Mean Reciprocal Rank (MRR) (Wu et al., 2011). The models were tested with a modified test set, where each context was paired with 10 responses (9 false and 1 true) to evaluate their ranking performance. The results are summarized in Table 4, highlighting the performance improvement with post-training + fine-tuning of the BERT model compared to only fine-tuning.

Due to its superior performance, the post-trained + fine-tuned Dial-MAE model was selected for the process of retrieval and ranking responses before response generation by LLMs. This model uses dot product (a standard measure of similarity between vectors) (Su et al., 2023) to measure the similarity between the context vector and the response vector. Based on these dot product scores, the responses from a bank of cleaned therapist messages were ranked in a descending order. The top 10 ranked responses were then used as examples in the final prompt provided to the LLMs.

4.4 LLMs and Prompting Techniques

We selected Large Language Models (LLMs) based on two primary criteria: their task-specific fine-tuning and supported context lengths, ensuring alignment with our experimental objectives. A major constraint was the inability to use large-scale models such as GPT-3.5 or GPT-4, due to strict data privacy and security requirements. This was particularly important given the highly sensitive and anonymized nature of the data, which involved individuals consuming Child Sexual Abuse Material (CSAM). As a result, we chose smaller models—up to 13 billion parameters, that could be run locally on our secure infrastructure, despite some computational limitations. All models were accessed via the Huggiång Face library[3], and are listed below:

- **Qwen2-7B-Instruct**: This 7-billion-parameter model [4] from the Qwen series (Bai et al., 2023; Yang et al., 2024) is fine-tuned on diverse instruction-based datasets, enhancing accuracy and contextual understanding. The Qwen2 series offers models ranging from 0.5B to 72B parameters and consistently outperforms many previous open-weight models.

- **Mistral-7B-Instruct-v0.3**: This instruction fine-tuned variant of Mistral-7B-v0.3[5] excels in reasoning, mathematics, and code generation, outperforming Llama 2 13b—Chat model in both human and automated benchmarks (Jiang et al., 2023).

[3] https://huggingface.co/

[4] https://huggingface.co/Qwen/Qwen2-7B-Instruct

[5] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

- **Orca-2-13b**: A 13-billion-parameter model optimized for complex reasoning tasks[6]. It matches or surpasses the performance of much larger models, excelling in zero-shot tasks and diverse contexts (Mitra et al., 2023).

- **Zephyr-7b-alpha**: A fine-tuned variant of Mistral-7B-v0.1 optimized for dialogue[7]. Trained on Ultrachat and Ultra Feedback, it sets a new standard for 7B models in chat benchmarks, outperforming Llama2-Chat-70B without human annotation (Tunstall et al., 2023).

4.5 Retrieval, Ranking and Generation Pipeline

The pipeline (baseline setup) involved retrieving and ranking the top 10 responses using the Dial-MAE model (post-trained + fine-tuned), based on their dot scores as mentioned in the subsection 4.3. Following this, a prompt was crafted for the LLMs for generation of a response. This prompt was iteratively refined to align with therapists' conversational style, ensuring responses were open-ended, concise, and therapeutic. To ensure consistency, the same prompt was used across all four selected LLMs. Thirty randomly chosen dialogue contexts were employed to compare model performance under identical conditions, using a temperature setting of 0.7 to limit response novelty (Peeperkorn et al., 2024). Each model generated responses for all the dialogues being tested, which were then used in the user study (see next subsection). Full details of the final prompt are provided in Figure 2 in Appendix A. Additionally, an example dialogue history with subsequent responses from a therapist and an LLM is provided in Appendix B.

4.6 Experimental Setup

For our experiments, we randomly selected 30 dialogue snippets from the test set of the therapist–patient dataset described in Subsection 4.2. Each dialogue consisted of approximately 5–10 utterances exchanged between therapist and patient, always concluding with a patient message. To ensure diversity, the dialogues were drawn from different patients and sessions. The selection also captured a range of conversation types, including dialogues from the beginning, middle, and end of sessions. These dialogues served as input to our retrieval, ranking, and generation pipeline, which produced four responses—one from each of the LLMs described in Subsection 4.4 and one response from the original human therapist, referred to as the "Therapist Response." This "Therapist Response" was the message immediately following the dialogue context. In total, we created a database of 30 dialogues, each with 5 responses, for evaluation.

4.7 Evaluation Setup

The evaluation comprised two stages: automatic (through an ablation study) and manual, explained in the following subsections.

4.7.1 Automatic Evaluation: Ablation Study

To assess the contribution of different components in the RAG framework, we conducted an ablation study by systematically removing or simplifying key elements. The study focused on evaluating the retrieval mechanism and the generative model through the following configurations:

- **Baseline Setup**: The complete RAG framework (Section 4.5), consisting of the BERT-based retrieval model, DIAL-MAE for ranking responses, and an LLM for response generation.

- **Ablation 1 - No Retrieval**: The retrieval step was removed, leaving only the LLM to generate responses. To compensate, the dialogue context was directly inserted into the prompt (Figure 3 in Appendix A), allowing the LLM to generate responses without retrieved examples. This setup isolates the retrieval mechanism's impact on performance.

- **Ablation 2 - No Generator**: The LLM was removed, and responses were derived solely from retrieved documents. The post-trained and fine-tuned DIAL-MAE retriever selected the top-ranked response, which was then evaluated against other configurations.

- **Ablation 3 - Simplified Retriever**: The DIAL-MAE retriever was replaced with all-MiniLM-L6-v2[8], a sentence-transformers model without fine-tuning. Unlike DIAL-MAE, which ranks responses using dot product scores, this model employed cosine similarity (a common metric that measures the

[6] https://huggingface.co/microsoft/Orca-2-13b

[7] https://huggingface.co/HuggingFaceH4/Zephyr-7b-alpha

[8] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

angle between two vectors) (Rahutomo et al., 2012) to compare dialogue history embeddings with therapist responses, assigning similarity scores between 0 (no similarity) and 1 (highest similarity).

For each configuration, the models were evaluated on the same task of response generation using the same dataset (explained in Subsection 4.6). The performance of each setup was measured using BERTScore (Zhang et al., 2019), which uses contextual embeddings to capture semantic similarity between generated and reference texts. Unlike ROUGE, which relies on exact word overlap, BERTScore better evaluates the nuanced and diverse responses typical in therapeutic dialogue. By conducting an ablation study, we were able to assess the contribution of each part of the RAG framework to its overall performance.

4.7.2 Manual Evaluation

The 14 participants included 11 professional therapists with varying levels of experience in CSAM therapy, all with at least one year of experience. Additionally, 3 students specializing in sexual medicine also took part in the study. The funding for this study was provided through as part of the larger project mentioned in Section 2.

Each of the 30 selected test dialogues was paired with five responses for evaluation: four generated by the four selected LLMs (Section 4.4) and one provided by a human therapist (Therapist Response). The participants assessed the appropriateness of each response using a 5-point Likert scale, with the labels "Least appropriate," "Inappropriate," "Neutral," "Appropriate," and "Most appropriate." The therapists were not aware of the hidden "Therapist Response" among the five options as the chats came from another therapist who was not a participant in this study. In addition to rating each individual response on a scale of appropriateness, the participants also selected their most preferred response out the five options provided. A screenshot of the questions used in the study can be found in Appendix C. Each participant rated responses for 10 randomly selected dialogues from a pool of 30, resulting in 50 ratings per participant.

To analyze the ratings for each model and the Therapist Response, we calculated the mean and standard deviation. We first assessed the normality of the distribution using the Shapiro-Wilk test (Shapiro and Wilk, 1965). Since the distribution did not meet the normality assumption, we employed the Kruskal-Wallis test (Kruskal and Wallis, 1952) to determine significant differences among the models. For any significant findings ($p < 0.05$), we conducted Dunn's post-hoc test with Bonferroni correction (Dunn, 1964) to identify specific model pairs with significant differences.

5 Results

5.1 Automatic Evaluation: Ablation Study

Table 5 presents the BERTScore (F1, Precision, and Recall) for each model across the baseline setup and two ablation configurations: *Ablation 1* (No Retrieval) and *Ablation 3* (Simplified Retriever). Table 6 separately reports results for *Ablation 2* (No Generator).

Mistral-7B-Instruct-v0.3 achieves the highest F1 in the baseline, while Orca-2-13b used in the baseline setup leads in Recall across all models. For the Zephyr-7b-alpha model, the setup used in Ablation 1 improves the F1 score unlike in other models. This shows that Zephyr-7b-alpha model generates similar responses as in the baseline setup even without a retriever. In Ablation 3 (Simplified Retriever), performance differences across models are generally small and variable, with some decreases in F1, Precision, and Recall. These variations suggest a subtle effect of the retriever component, though Dunn's post-hoc tests show no statistically significant differences ($p > 0.05$), suggesting that while performance variations exist, they are not statistically meaningful. Table 5 summarizes the results, highlighting the complementary roles of both retrieval and generation for optimal performance. On the other hand, Ablation 2 (Retriever-only, DialMAE model) shows a notable decrease in F1, Precision, and Recall (see Table 6), emphasizing the importance of the generator component in achieving balanced performance.

5.2 Results on a similar dataset

To assess generalizability, we evaluated our baseline setup (see 4.5) on the CACTUS dataset (Lee et al., 2024), a GPT-4o-generated multi-turn counseling dataset in English. While synthetic and contextually different, it shares structural traits (e.g., turn count, message length) with our primary data. Only automatic evaluation was performed using BERTScore, following a similar approach as described in Section 4.6. As shown in Appendix D, overall BERTScore values on CACTUS are notably

Model	Setup	F1	Precision	Recall
Mistral-7B-Instruct-v0.3	Baseline	**0.8552**	0.8568	0.8540
	Ablation 1	0.8544	0.8551	0.8538
	Ablation 3	0.8549	0.8579	0.8523
Zephyr-7b-alpha	Baseline	0.8515	0.8519	0.8514
	Ablation 1	0.8528	0.8546	0.8513
	Ablation 3	0.8428	0.8351	0.8512
Orca-2-13b	Baseline	0.8541	0.8517	**0.8567**
	Ablation 1	0.8485	0.8469	0.8502
	Ablation 3	0.8448	0.8413	0.8485
Qwen2-7B-Instruct	Baseline	0.8544	**0.8586**	0.8506
	Ablation 1	0.8490	0.8516	0.8466
	Ablation 3	0.8540	0.8563	0.8519

Table 5: Summary of BERTScore (F1, Precision, Recall) for all models across the baseline and ablation configurations 1 (No retrieval) and 3 (Simplified retriever).

Model	F1	Precision	Recall
Dial-MAE	0.8100	0.8022	0.8275

Table 6: Performance metrics for the retriever-only (post-trained+fine-tuned DIAL-MAE model) model (Ablation 2: No Generator).

lower than our baseline results in Table 5, highlighting the challenge of generalizing across datasets. However, consistent trends persist—Mistral-7B-Instruct-v0.3 achieves the highest F1 and recall (tied with Orca-2-13b), and Qwen2-7B-Instruct retains the highest precision.

5.3 Manual Evaluation

Manual evaluation was conducted using user ratings for four LLMs: Mistral-7B-Instruct-v0.3, Orca-2-13b, Qwen2-7B-Instruct, Zephyr-7b-alpha, and the "Therapist Response" where the responses evaluated were generated using the baseline setup (section 4.5). The statistical test results are summarized below:

Table 7 displays the mean and standard deviation of ratings. Mistral-7B-Instruct-v0.3 received the highest average rating (3.34), while Zephyr-7b-alpha had the lowest (2.76). The standard deviations indicate variability in user evaluations. This variability is also depicted in Appendix E, which illustrates the distribution of user ratings. Mistral-7B-Instruct-v0.3 and Orca-2-13b received higher ratings (4.0 and 5.0) more frequently, while Zephyr-7b-alpha and Qwen2-7B-Instruct had a greater proportion of lower ratings (1.0 and 2.0).

A Shapiro-Wilk test, as shown in Table 8, revealed that the ratings for all models, including the "Therapist Response", deviate from normality ($p < 0.05$). Therefore, non-parametric testing was used. A Kruskal-Wallis test was conducted,

Model	Mean	Std Dev
Therapist Response	3.06	1.24
Mistral-7B-Instruct-v0.3	**3.34**	1.16
Orca-2-13b	3.26	1.25
Qwen2-7B-Instruct	2.91	1.15
Zephyr-7b-alpha	2.76	1.19

Table 7: Model rating summary showing the mean and standard deviation of ratings for each model.

giving the following values: H-statistic: 21.2020, p-value: 0.0003, indicating statistically significant differences among the ratings for the five types of responses.

To further examine these differences, a post-hoc Dunn's test with Bonferroni correction was conducted (see Appendix F). The results indicate significant differences between Zephyr-7B-alpha and several other models, including Mistral-7B-Instruct-v0.3 ($p = 0.0012$) and Orca-2-13b ($p = 0.0054$). Qwen2-7B-Instruct also showed a statistically significant difference from Mistral-7B-Instruct-v0.3 ($p = 0.0387$). No significant differences were found among most other model pairs, suggesting broadly similar performance in the context of this study.

Table 9 indicates that Orca-2-13b was the most preferred (42 times), while Zephyr-7b-alpha was the least preferred (17 times). The Chi-square test (Chi-square statistic: 19.3750, p-value: 0.0007) reveals significant preference differences among the models.

Due to the limited sample size, no differences were observed between the ratings provided by student raters and experienced raters.

Model	W-stat	p-value
Therapist Response	0.9063	7.04719e-08
Mistral-7B-Instruct-v0.3	0.9020	4.07591e-08
Orca-2-13b	0.8913	1.08807e-08
Zephyr-7b-alpha	0.8914	1.10034e-08
Qwen2-7B-Instruct	0.9016	3.86513e-08

Table 8: Normality Test (Shapiro-Wilk) results for model ratings.

6 Discussion

This study investigates the use of Large Language models (LLMs)(specifically those with up to 13 billion parameters) in generating therapeutic responses for multi-turn, chat-based interventions

Model	Preferred Responses
Therapist Response	26
Mistral-7B-Instruct-v0.3	35
Orca-2-13b	**42**
Qwen2-7B-Instruct	20
Zephyr-7b-alpha	17

Table 9: Preferred responses count for each model.

aimed at preventing the use of child sexual abuse material (CSAM). Unlike prior research that relies on larger LLMs like GPT-4 (Inaba et al., 2024), our approach prioritizes data privacy and safety, aligning with ethical concerns and ensuring that sensitive content is handled with care. We employed a Retrieval-Augmented Generation (RAG) approach, integrating a post-trained, fine-tuned BERT-based retriever model alongside an LLM to generate therapist-like responses. To evaluate the effectiveness of this setup, we conducted an ablation study, which was evaluated automatically using BERTScore. Despite observing variations in performance, statistical tests indicated no significant differences, suggesting that the differences in scores may not be meaningful. This lack of significant difference may partly be due to limitations of BERTScore in capturing the nuanced benefits of retrieval-guided responses. Based on these findings, we selected the baseline setup for further evaluation through a manual review by 14 psychotherapists. The manual evaluation revealed that Mistral-7B-Instruct-v0.3 and Orca-2-13b were preferred over the therapist response in most cases. Despite the small sample size, our human evaluation involved domain experts in a highly specialized therapeutic context, providing clinically meaningful insights.

While the study focused on evaluating model performance and did not delve deeply into the impact of different fine-tuning techniques, the findings emphasize the importance of choosing the right model architecture and setup. Furthermore, the observed discrepancies between automated and expert evaluations highlight the limitations of automatic metrics like BERTScore in assessing the therapeutic value of responses. As noted in previous studies (Filienko et al., 2024), BERTScore may not fully capture the nuances required to evaluate the effectiveness of therapeutic interventions. The lower BERTScore values for the CACTUS dataset (see Appendix D) may result from its fully LLM-generated nature, in contrast to the authentic therapist–patient interactions in our primary data. As the prompt was specifically designed for the CSAM context to capture the linguistic characteristics of therapist messages, it may be less effective when applied to synthetic datasets like CACTUS. This highlights the distinct nature of real therapeutic dialogues.

Our model's creativity and coherence were also influenced by hyperparameters, such as temperature settings. We chose a temperature of 0.7 to strike a balance between creativity and coherence, minimizing excessive randomness while ensuring some degree of variability in the responses. Research suggests that temperature settings subtly affect creativity (Zhao et al., 2024), with higher temperatures fostering more novel outputs (Peeperkorn et al., 2024). This aspect of model behavior requires further exploration in future work.

7 Conclusion and Future Work

This study highlights the feasibility of using LLMs for chat-based therapy in the context of CSAM prevention, emphasizing the need to balance model performance with privacy and safety considerations. The results suggest that models such as Mistral-7B-Instruct-v0.3 and Orca-2-13B are capable of generating contextually appropriate and therapeutically relevant responses. These findings offer valuable insights into LLM applications in sensitive domains, where patient messages may contain explicit content that many models struggle to interpret. Although not yet suitable for direct patient use, the approach shows promise as a supportive tool for therapists, offering AI-generated suggestions to reduce their workload.

Our work emphasizes the importance of combining both retrieval and generation components to optimize model performance in therapy settings. The ablation study showed that the retrieval mechanism plays a crucial role in enhancing the effectiveness of response generation, while the generation component is essential for producing contextually appropriate responses that mimic therapists. These findings highlight the potential of LLMs to replicate or mimic therapist responses, emphasizing the importance of balancing both retrieval and generation in therapeutic applications, rather than relying exclusively on therapy principles. This was largely possible due to the availability of domain-specific therapy dialogues, which provided concrete examples that guided the LLMs in producing contextually appropriate responses.

Looking ahead, future work could improve retrieval via better data labeling or intent detection, and further explore prompt engineering, given its strong impact on model outputs. Temperature settings for generating responses with LLMs also need investigation (especially in this domain), as they affect response creativity and coherence. Expanding the study with larger, more diverse samples can offer deeper insights into model behavior across therapeutic contexts. Using LLM-as-a-judge approach could be valuable where expert therapists are scarce, but due to the sensitive nature of the dialogues, this would also require smaller locally run LLMs rather than GPT, which has strict content moderation. While LLMs show promise in areas like CSAM prevention, continued research is required to address challenges and ensure their safe and effective use.

8 Limitations

Using AI and LLMs in therapy for sensitive topics like Child Sexual Abuse Material (CSAM) involves strict privacy and ethical considerations, which prevented the use of third-party APIs for handling patient data. To address this, we employed locally hosted models, which required working with smaller-parameter LLMs and a fine-tuned BERT-based embedding model. While these models may be considered limited compared to the latest large-scale architectures, this choice was necessary to ensure data privacy, faster inference, and real-time responsiveness, all of which are key requirements for potential deployment in therapeutic settings. Additionally, many state-of-the-art LLMs, such as GPT-4 or newer LLaMA variants impose content moderation policies that prevent them from processing explicit language. These constraints made them unsuitable for our application, where the ability to interpret and respond to sensitive content is essential. Although this limited our access to the very latest model capabilities, it enabled the development of a responsible, privacy-preserving system aligned with the considerations of working in high-risk clinical domains.

9 Ethical Considerations

Ethical approval for using patient data, including dialogue data, was granted for research purposes and the training of AI models, with strict adherence to ethical guidelines and proper consent protocols. Approval was obtained from the relevant institutional review board. To ensure the protection of patient data, all AI models were run locally, keeping the data confined to the study environment. Prior to any publication, the data will undergo thorough anonymization to safeguard patient privacy, which is why it cannot be published at this time. Although if required a sample of the dataset can be shared upon request. Additionally, further ethical considerations will be necessary for the deployment of an automated chatbot based on these findings.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Aaron T Beck. 1963. Thinking and depression: I. idiosyncratic content and cognitive distortions. *Archives of general psychiatry*, 9(4):324–333.

Samuel Bell, Clara Wood, and Advait Sarkar. 2019. Perceptions of chatbots in therapy. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6.

Agata Berdowska and Katarzyna Zdanowicz-Cyganiak. 2024. 11 digital therapists-the potential and limitations of artificial intelligence. *Trust and Artificial Intelligence: Development and Application of AI Technology*, page 120.

Desirée Bill and Theodor Eriksson. 2023. Fine-tuning a llm using reinforcement learning from human feedback for a therapy chatbot application.

Christine Bundy. 2004. Changing behaviour: using motivational interviewing techniques. *Journal of the royal society of medicine*, 97(Suppl 44):43.

Siyuan Chen, Mengyue Wu, Kenny Q Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. Llm-empowered chatbots for psychiatrist and patient simulation: application and evaluation. *arXiv preprint arXiv:2305.13614*.

Yujin Cho, Mingeon Kim, Seojin Kim, Oyun Kwon, Ryan Donghan Kwon, Yoonha Lee, and Dohyun Lim. 2023. Evaluating the efficacy of interactive language therapy based on llm for high-functioning autistic adolescent psychological counseling. *arXiv preprint arXiv:2311.09243*.

Olive Jean Dunn. 1964. Multiple comparisons using rank sums. *Technometrics*, 6(3):241–252.

Daniil Filienko, Yinzhou Wang, Caroline El Jazmi, Serena Xie, Trevor Cohen, Martine De Cock, and Weichao Yuwen. 2024. Toward large language models as a therapeutic tool: Comparing prompting techniques to improve gpt-delivered problem-solving therapy. *arXiv preprint arXiv:2409.00112*.

Colm Gannon, Arjan AJ Blokland, Salla Huikuri, Kelly M Babchishin, and Robert JB Lehmann. 2023. Child sexual abuse material on the darknet. *Forensische Psychiatrie, Psychologie, Kriminologie*, 17(4):353–365.

Zainab Iftikhar, Sean Ransom, Amy Xiao, and Jeff Huang. 2024. Therapy as an nlp task: Psychologists' comparison of llms and human peers in cbt. *arXiv preprint arXiv:2409.02244*.

Michimasa Inaba, Mariko Ukiyo, and Keiko Takamizo. 2024. Can large language models be used to provide psychological counselling? an analysis of gpt-4-generated responses using role-play dialogues. *arXiv preprint arXiv:2402.12738*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Luoma Ke, Song Tong, Peng Chen, and Kaiping Peng. 2024. Exploring the frontiers of llms in psychological applications: A comprehensive review. *arXiv preprint arXiv:2401.01519*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Zoha Khawaja and Jean-Christophe Bélisle-Pipon. 2023. Your robot therapist is not your therapist: understanding the role of ai-powered mental health chatbots. *Frontiers in Digital Health*, 5:1278186.

William H Kruskal and W Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621.

Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991*.

Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, et al. 2024. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. *arXiv preprint arXiv:2407.03103*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles. *arXiv preprint arXiv:2407.00870*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

Ben Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 1.

Ggaliwango Marvin, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. 2023. Prompt engineering in large language models. In *International conference on data intelligence and cognitive informatics*, pages 387–402. Springer.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, et al. 2023. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*.

Christian Muise, Tathagata Chakraborti, Shubham Agarwal, Ondrej Bajgar, Arunima Chaudhary, Luis A Lastras-Montano, Josef Ondrej, Miroslav Vodolan, and Charlie Wiecha. 2019. Planning for goal-oriented dialogue systems. *arXiv preprint arXiv:1910.08137*.

Hongbin Na. 2024. Cbt-llm: A chinese large language model for cognitive behavioral therapy-based mental health question answering. *arXiv preprint arXiv:2403.16008*.

Jingping Nie, Hanya Shao, Yuang Fan, Qijia Shao, Haoxuan You, Matthias Preindl, and Xiaofan Jiang. 2024. Llm-based conversational ai therapist for daily functioning screening and psychotherapeutic intervention via everyday smart devices. *arXiv preprint arXiv:2403.10779*.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? *arXiv preprint arXiv:2405.00492*.

Faisal Rahutomo, Teruaki Kitasuka, Masayoshi Aritsugi, et al. 2012. Semantic cosine similarity. In *The 7th international student conference on advanced science and technology ICAST*, volume 4, page 1. University of Seoul South Korea.

Leonardo Ranaldi and Andre Freitas. 2024. Aligning large and small language models via chain-of-thought reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1812–1827.

Samuel Sanford Shapiro and Martin B Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611.

Hao Shen, Zihan Li, Minqiang Yang, Minghui Ni, Yongfeng Tao, Zhengyang Yu, Weihao Zheng, Chen Xu, and Bin Hu. 2024. Are large language models possible to conduct cognitive behavioral therapy? *arXiv preprint arXiv:2407.17730*.

Jiho Shin, Clark Tang, Tahmineh Mohati, Maleknaz Nayebi, Song Wang, and Hadi Hemmati. 2023. Prompt engineering or fine tuning: An empirical assessment of large language models in automated software engineering tasks. *arXiv preprint arXiv:2310.10508*.

Chiyu Song, Hongliang He, Haofei Yu, Pengfei Fang, Leyang Cui, and Zhenzhong Lan. 2021. Uni-encoder: A fast and accurate response selection paradigm for generation-based dialogue systems. *arXiv preprint arXiv:2106.01263*.

Zhenpeng Su, Xing Wu, Wei Zhou, Guangyuan Ma, and Songlin Hu. 2023. Dial-mae: Contextual masked auto-encoder for retrieval-based dialogue systems. *arXiv preprint arXiv:2306.04357*.

Lipeipei Sun, Tianzi Qin, Anran Hu, Jiale Zhang, Shuojia Lin, Jianyan Chen, Mona Ali, and Mirjana Prpa. 2024. Persona-l has entered the chat: Leveraging llm and ability-based framework for personas of people with complex needs. *arXiv preprint arXiv:2409.15604*.

Sara Syed, Zainab Iftikhar, Amy Wei Xiao, and Jeff Huang. 2024. Machine and human understanding of empathy in online peer support: A cognitive behavioral approach. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, et al. 2023. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yang Wu, Masayuki Mukunoki, Takuya Funatomi, Michihiko Minoh, and Shihong Lao. 2011. Optimizing mean reciprocal rank for person re-identification. In *2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 408–413. IEEE.

Haojie Xie, Yirong Chen, Xiaofen Xing, Jingkai Lin, and Xiangmin Xu. 2024. Psydt: Using llms to construct the digital twin of psychological counselor with personalized counseling style for psychological counseling. *arXiv preprint arXiv:2412.13660*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Jeyoon Yeom, Hakyung Lee, Hoyoon Byun, Yewon Kim, Jeongeun Byun, Yunjeong Choi, Sungjin Kim, and Kyungwoo Song. 2024. Tc-llama 2: fine-tuning llm for technology and commercialization applications. *Journal of Big Data*, 11(1):100.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yunpu Zhao, Rui Zhang, Wenyi Li, Di Huang, Jiaming Guo, Shaohui Peng, Yifan Hao, Yuanbo Wen, Xing Hu, Zidong Du, et al. 2024. Assessing and understanding creativity in large language models. *arXiv preprint arXiv:2401.12491*.

A Prompt for LLMs

This section describes the uniform prompt applied across all LLMs for generating responses for the baseline setup and the Ablation 1.

```
Conversation history:
{dialogue_history}

Instructions:
From the list of example responses provided by a fine-tuned BERT model, select and modify the most appropriate
response to align with the client's recent message. If none are suitable, generate a new response.

Key points:
- The response must be concise and suited for live chat.
- The response should be short consisting 1 sentence only.
- Maintain an open-ended, and directive tone.
- Avoid overwhelming the client with too much information.
- Do not suggest specific actions (e.g., "delving deeper" or "discussing further" or "Let's talk about").
- Encourage reflection or sharing, consistent with the therapist's conversational style.
- Remember, each therapy session is unique.

Example responses, with their dot product scores computed by the fine-tuned BERT model in descending order:
{example_responses_text}

Final response: <therapist_response>
```

Figure 2: Prompt used for LLMs to generate responses with retrieved examples (Baseline setup)

```
Instructions:
Generate the most appropriate next response as a "Therapist" to align with the client's recent message given
the conversation history:
{dialogue_history}


Key points:
- The response must be concise and suited for live chat.
- The response should be short consisting 1 sentence only.
- Maintain an open-ended, and directive tone.
- Avoid overwhelming the client with too much information.
- Do not suggest specific actions (e.g., "delving deeper" or "discussing further" or "Let's talk about").
- Encourage reflection or sharing, consistent with the therapist's conversational style.
- Remember, each therapy session is unique.

Final response: <therapist_response>
```

Figure 3: Prompt used for LLMs to generate responses without retrieved examples (Ablation 1)

B Example Responses from an LLM and a Therapist in a Therapy Dialogue

Dialogue history/context:

Therapist: Did you use any material for sexual arousal since the last session?
Client: No illegal content.
Therapist: That's quite an accomplishment!
Therapist: Are you feeling proud about it?
Client: No, not really.
Therapist: Why not?
Client: I just didnt really have the urge to I dont think.
Client: So it's not like i had to stop myself.

LLM-generated response:

Are there any other activities that help you feel relaxed or fulfilled in a similar way as before?

Therapist's response:

Perhaps simply discussing it helped to decrease your urge?

Figure 4: Example dialogue excerpt from the study, showing the subsequent responses from both a therapist and an LLM (Mistral-7b-Instruct-v0.3). These responses were obtained using the baseline setup explained in section 4.7.1. Here, "Client" refers to "Patient" or consumers of Child Sexual Abuse Material (CSAM).

C Example Item from User Study

This figure illustrates a sample dialogue item presented to therapists during the user study. The interface displays multiple response options for a given conversation history. Next to each response is a radio button allowing the participant to select the most preferred response. Additionally, each response can be rated on a 5-point Likert scale based on its appropriateness within the conversation context.

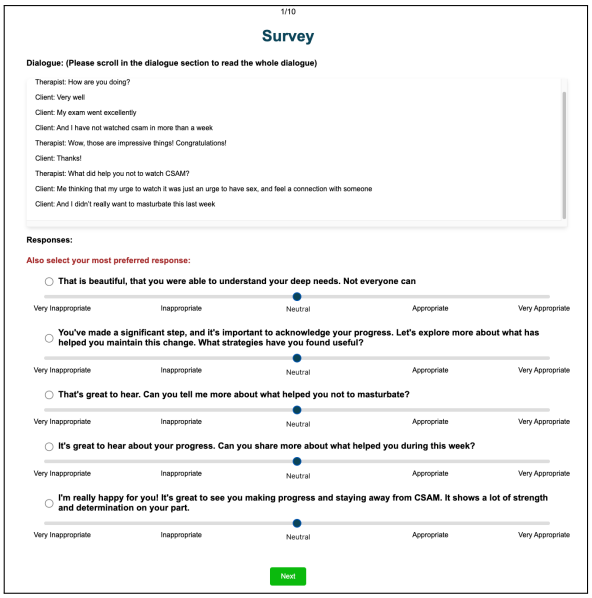

Figure 5: An example dialogue item used in the user study with therapists. Here, "Client" refers to "Patient" or consumers of Child Sexual Abuse Material (CSAM).

D Result on Another Multi-turn Counseling Dataset

This section presents evaluation results on the CACTUS dataset (Lee et al., 2024), using BERTScore to compare model responses to real counselors, reporting F1, Precision, and Recall scores.

Model	F1	Precision	Recall
Mistral-7B-Instruct-v0.3	**0.6075**	0.6058	**0.6093**
Orca-2-13b	0.5999	0.5909	**0.6093**
Qwen2-7B-Instruct	0.6068	**0.6083**	0.6053
Zephyr-7b-alpha	0.5885	0.5809	0.5965

Table 10: BERTScore evaluation on the CACTUS dataset: precision, recall, and F1 scores comparing model responses to real counselor replies.

E Stacked bar chart comparing four LLMs and real therapist response across manual ratings

This section presents a stacked bar chart comparing the ratings of four LLMs and a real therapist response, visualizing how each was evaluated across various appropriateness levels, based on participant ratings (average results in Table 7).

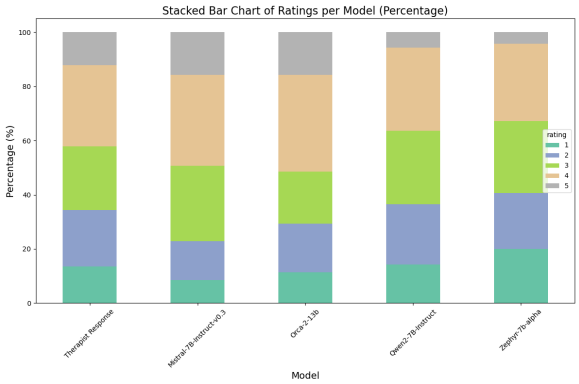

Figure 6: Stacked bar chart depicting a comparison of the models and therapist response across the ratings.

F Post-hoc Dunn's Test Results with Bonferroni Correction

This section reports p-values from pairwise comparisons between model and therapist responses using Dunn's test, with Bonferroni correction for multiple comparisons, assessing statistical significance in therapist ratings.

Model	Mistral-7B-Instruct-v0.3	Orca-2-13b	Qwen2-7B-Instruct	Therapist Response	Zephyr-7b-alpha
Mistral-7B-Instruct-v0.3	1	1	0.0387	0.6825	0.0012
Orca-2-13b	1	1	0.1215	1	0.0054
Qwen2-7B-Instruct	0.0387	0.1215	1	1	1
Therapist Response	0.6826	1	1	1	0.4370
Zephyr-7b-alpha	0.0012	0.0054	1	0.4370	1

Table 11: Post-hoc Dunn's Test Results with Bonferroni Correction: p-values for pairwise model comparisons.