MentalRAG: Developing an Agentic Framework for Therapeutic Support Systems

Francisco R. E Silva[1,2][a], Pedro A. Santos[1,2][b] and João Dias[2,3,4][c]
[1]*Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, Lisboa, Portugal*
[2]*INESC-ID, Rua Alves Redol, 9, Lisboa, Portugal*
[3]*Faculty of Science and Technology, University of Algarve, Campus de Gambelas, Faro, Portugal*
[4]*CISCA, Campus de Gambelas, Faro, Portugal*

Keywords: Artificial Intelligence, Large Language Models, Mental Health, Retrieval Augmented Generation, Agentic Workflow.

Abstract: This paper introduces MentalRAG, a multi-agent system built upon an agentic framework designed to support mental health professionals through the automation of patient data collection and analysis. The system effectively gathers and processes high-sensitivity mental health data from users. It employs locally run open-source models for most tasks, while leveraging advanced state-of-the-art models for more complex analyses, ensuring the maintenance of data anonymity. The system's models have showed improvements in delivering empathetic and contextually adaptive responses, particularly in sensitive contexts such as emotional distress and crisis management. Notably, an integrated agent for detecting levels of suicidal ideation allows the system to assess and respond sensitively to diverse levels of risk, promptly alerting mental health professionals as needed. This innovation represents a stride towards creating a more reliable, efficient, and ethically responsible mental health support tool, capable of addressing both patient and doctor needs effectively while minimizing associated risks.

1 INTRODUCTION

Mental health is a global issue, with millions affected around the world. In the United States, one in five adults experience mental illness annually, and serious conditions such as schizophrenia or bipolar disorder affect 1 in 25. The economic impact of mental health disorders, such as depression and anxiety, leads to global productivity losses estimated at $1 trillion annually (National Alliance on Mental Illness, 2023). As patient numbers increase, mental health professionals face overwhelming caseloads, highlighting the urgent need for innovative solutions that can improve care delivery while easing the workload on healthcare providers.

Mental health, as defined by the World Health Organization (WHO), extends beyond the absence of mental disorders to encompass positive attributes such as emotional resilience, psychological functioning, and the ability to manage everyday stresses. It influences how individuals think, feel, and behave, affecting their capacity to navigate relationships, make decisions, and contribute productively to society. Given the rising rates of mental health issues worldwide, understanding the complex interaction between biological, psychological, and environmental factors is essential. Research in psychology seeks to identify these underlying processes, enabling professionals to diagnose mental illnesses accurately and develop effective interventions aimed at improving individuals' quality of life.

In response to these challenges, we present a system using large language models (LLMs) to assist mental health professionals by automating patient data collection and analysis. The system incorporates LLM-powered agents that handle tasks such as chatting with patients, anonymizing data, detecting suicide ideation risk, and generating detailed reports. By leveraging patient history, previous session data, and psychologist feedback, the system personalizes responses and clinical recommendations. It focuses on key areas such as suicide risk assessment, diag-

[a] https://orcid.org/0009-0005-9834-8615
[b] https://orcid.org/0000-0002-1369-0085
[c] https://orcid.org/0000-0002-1653-1821

Silva, F. R. E., Santos, P. A. and Dias, J.
MentalRAG: Developing an Agentic Framework for Therapeutic Support Systems.
DOI: 10.5220/0013267400003938
In *Proceedings of the 11th International Conference on Information and Communication Technologies for Ageing Well and e-Health (ICT4AWE 2025)*, pages 46-57
ISBN: 978-989-758-743-6; ISSN: 2184-4984

nostics, and generating actionable insights, while ensuring that final decision-making remains in the hands of human experts. Local execution of critical system components guarantees data security, addressing both privacy and performance concerns.

Understanding the complexity of mental health conversations requires advanced approaches, as standard models fail to grasp the emotional depth and subtle nuances involved. Our solution uses natural language processing techniques to handle the intricate nature of these discussions. It combines a RAG process, which retrieves relevant information to enrich context, with specialized mental health datasets to improve analysis. This ensures a more accurate interpretation of patient emotions and intentions, providing nuanced insights for psychological care.

The lack of tools that integrate seamlessly into psychologists' workflows is a challenge in mental health. Our system overcomes this by fine-tuning LLMs with mental health data and incorporating a user-friendly interface that adapts to different stages of care, from diagnosis to progress monitoring. The integration of RAG technology enriches the model's understanding, providing contextually relevant information. This allows for more effective clinical decisions while keeping psychologists at the center of the process, ensuring both efficiency and the ethical management of patient care.

Figure 1 shows an overview of our proposed system.

Our contributions are:

- We propose, implement, and evaluate an agentic framework to collect and analyze high-sensitive mental health data from users, where most of the agents run "small" open-source models locally, while more demanding tasks are performed by state-of-the-art powerful models over anonymized data.
- We tested several open source models and, in particular, found that LLaMA-3-8B-Instruct (Dubey et al., 2024) model and Mental-LLaMA-3-8B-Instruct, our fine-tuned model, outperforms ChatCounselor model (Liu et al., 2023) and existing open-source models in the benchmark Counselling Bench.

This paper is organized as follows: Section 2 contains an overview of the State of the Art; Section 3 describes the datasets used to fine tune and validate our model; in Section 4 the approach is presented with detailed description of the agents, techniques and solution; Section 5 shows the evaluation of the presented approach; in Section 6 we present the conclusions.

2 RELATED WORK

In this section we present an overview of related work, focusing on describing recent advances on large language models for the mental health domain, and overview of agent centric approaches.

2.1 Mental-LLM

Recent studies have initiated evaluations of LLMs on mental health-related tasks, with a focus on zero-shot settings and simple prompt engineering. Preliminary findings suggest that LLMs can predict mental health disorders using natural language, although with performance limitations compared to state-of-the-art domain-specific NLP models. This performance gap is attributed to the general-purpose nature of existing LLMs, which are not specifically trained for mental health tasks.

2.1.1 Mental-RoBERTa

Mental-RoBERTa and its counterpart MentalBERT (Xu et al., 2024) are masked language models specifically tailored for the domain of mental healthcare. These models represent a significant contribution to the field, as they are the first attempt to pre-train language models for mental healthcare applications.

MentalBERT and MentalRoBERTa are pre-trained on a corpus collected from Reddit, which includes discussions from various mental health-related subreddits. This approach aims to capture the nuances and linguistic characteristics of language used in the context of mental health discussions, making these models particularly suited for analyzing and understanding text data related to mental health conditions.

The performance of MentalRoBERTa was evaluated across several mental health detection tasks, demonstrating superior results in comparison to both general-domain models like BERT (Devlin et al., 2019) and RoBERTa (Liu, 2019), and other domain-specific models such as BioBERT (Lee et al., 2019) and ClinicalBERT (Huang et al., 2019). This indicates the significant advantage of domain-specific pretraining for mental health applications.

2.1.2 Mental-FLAN-T5

The FLAN-T5 model (Chung et al., 2024) represents a significant advancement in the field of natural language processing through instruction finetuning. The FLAN-T5 models have been fine-tuned on a diverse set of instructions to enhance their zero-shot, few-shot, and chain-of-thought reasoning capabilities, outperforming their T5 (Raffel et al., 2020) coun-

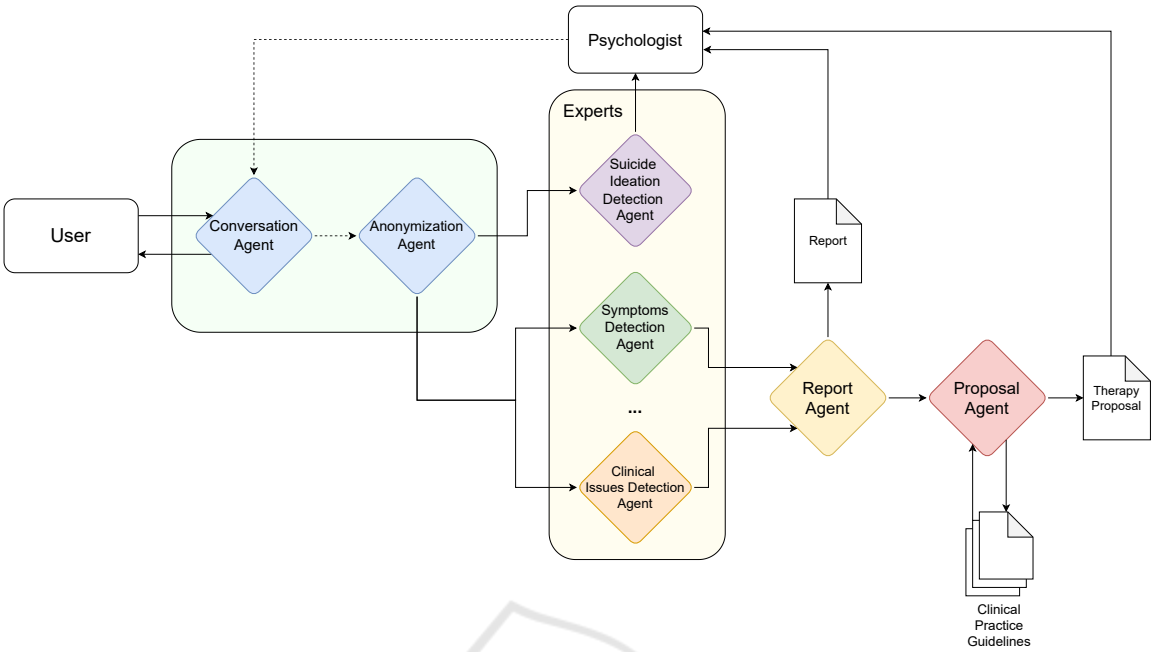

Figure 1: Architecture of MentalRAG.

terparts and even surpassing the performance of larger models like PaLM 62B (Chowdhery et al., 2023) on certain benchmarks.

Mental-FLAN-T5 was developed by instruction finetuning FLAN-T5 on the six tasks across the four datasets simultaneously. This approach aimed to enhance the model's capability in the mental health domain. The instruction finetuning involved updating model parameters over 3 epochs using Adam optimizer, with a learning rate of $2e^{-5}$ and a cosine scheduler warmup ratio of 0.03, utilizing eight Nvidia A100 GPUs.

Compared to larger models like GPT-3.5 (Brown et al., 2020) and GPT-4 (OpenAI et al., 2023), Mental-FLAN-T5 achieved significantly better results when fine-tuned for mental health-related tasks, exhibiting a 10.9% improvement in balanced accuracy over GPT-3.5, despite being 15 times smaller. This outcome highlights the capabilities of Mental-FLAN-T5 in managing complex, multitask scenarios, reinforcing its suitability for nuanced tasks such as suicide risk prediction on mental health datasets containing user-level annotations based on the Columbia Suicide Severity Rating Scale (C-SSRS) guidelines. (Xu et al., 2024)

2.1.3 Mental-Alpaca

Alpaca (Taori et al., 2023) is a fine-tuned language model developed from Meta's LLaMA 7B model (Touvron et al., 2023a). Created by researchers at the Center for Research on Foundation Models (CRFM) at Stanford University, Alpaca stands out for its cost-effective and efficient training process, aimed at fostering academic research in instruction-following models.

The creation of Mental-Alpaca involved instruction-based fine-tuning on mental health-specific tasks. This model was further refined using a combination of mental health datasets to handle tasks such as depression, stress, and suicidal ideation detection. Fine-tuning with instruction learning techniques enables Mental-Alpaca to excel in complex, multi-task environments, particularly in binary suicide ideation prediction (Xu et al., 2024).

2.2 Agentic Workflow

The Agentic Workflow is based on the concept of multi-agent systems (MAS), where autonomous agents collaborate to solve complex tasks by leveraging their specialized capabilities. Each agent focuses on a specific task, working in parallel or independently to contribute to the overall objective, similar to human teamwork. This modular approach breaks down intricate problems into smaller tasks, allowing agents to efficiently handle subtasks. Agents are designed to dynamically adjust their actions based on ongoing feedback, ensuring a flexible and robust system (Singh et al., 2024). For example, agents in a travel planning system can specialize in areas like

data collection or itinerary optimization, streamlining the process and improving efficiency.

A key aspect of the Agentic Workflow is the integration of external tools, allowing agents to make more informed decisions and mitigate issues like language model hallucinations. By connecting to APIs or specialized software, agents can retrieve accurate data, enhancing their problem-solving capabilities. Studies using tool-enhanced language models, such as "Gorilla" (Patil et al., 2024) and "MM-REACT" (Yang et al., 2023), demonstrate the effectiveness of this approach in expanding the range of tasks agents can handle, from visual intelligence to API-based queries. Additionally, agent interactions are optimized through techniques like parallelization and feedback loops, where agents iteratively improve their performance, leading to more adaptive and precise systems (Capital, 2024). This collaborative and decentralized decision-making model has applications across various domains, from smart grids to urban traffic management, and represents a step toward more autonomous and sophisticated AI systems.

3 DATASETS

As with the primary distinctions between the LLaMA-2 (Touvron et al., 2023b) and LLaMA-3 models, which were attributable to the quality and scale of the data, we adopted a comparable approach by concentrating on identifying a suitable dataset to enhance counselling-style responses. The Psych8k dataset, developed by (Liu et al., 2023), was used. It comprises 260 English-language counseling conversations, extracted from recorded clinical sessions between patients and certified professionals. The aforementioned conversations encompass five principal subject areas: The categories of mental health concerns, stress, family and relationships, minority groups, and others.

Furthermore, four additional datasets were gathered for the purpose of evaluating the models employed in a variety of tasks, see Table 1. The UMR Suicidality Dataset (Shing et al., 2018), an effort from the University of Maryland, represents a significant contribution to the field of suicide risk assessment, as it is the first to rely solely on social media content. This dataset was developed using the *r/SuicideWatch* subreddit. It enables researchers to identify anonymous users who may have signs of suicidality. The data was annotated at four levels. The framework offers a comprehensive evaluation of suicide risk, categorising posts as either "No Risk", "Low Risk", "Moderate Risk", or "Severe Risk". Furthermore, crowdsourced annotations were compared with expert evaluations and baseline predictive modeling experiments were conducted to demonstrate the potential of this dataset to advance suicide prevention through social media analysis.

The Crisis Response QA dataset (Jin et al., 2023) comprises a series of pivotal inquiries pertaining to the administration of mental health crises, meticulously assembled from esteemed sources such as the "Responding to Mental Health Crisis" and "Navigating a Mental Health Crisis" manuals. These guides provide an invaluable framework for the effective management of mental health crises, offering best practice guidance. The dataset comprises question-answer pairs that have been reviewed by medical students to ensure quality assurance.

The USMLE-Mental dataset (Jin et al., 2023) represents a specialized subset of the MedQA dataset, comprising multiple-choice questions sourced from professional medical board examinations, including the United States Medical Licensing Examination (USMLE). This dataset is centred on mental health-related questions from the USMLE and the Step 2 Clinical Knowledge (CK) sections. A manual review was conducted to confirm the relevance of the extracted questions to mental health. This process resulted in the annotation of 727 questions, which provide valuable insights into mental health knowledge within the medical field.

The final dataset is the Counselling Bench, which is a comprehensive framework designed to evaluate the performance of chatbots in psychological counselling contexts. This dataset extends beyond the assessment of basic question-answering capabilities to evaluate the capacity of chatbots to replicate the subtleties of counselling strategies employed in professional mental health support. The dataset comprises 229 open-ended queries, which simulate real-world counselling conversations. This provides a robust testing ground for chatbots to demonstrate their capability in delivering therapeutic support.

4 APPROACH

Our approach was based on the agentic workflow, which entails equipping each agent with specialized knowledge and distinct tasks. This permits greater flexibility in the selection of the large language models (LLMs) for each agent and the fine-tuning of the models for specific tasks. In our case, we performed fine-tuning using auto-regressive training exclusively on the Conversation Agent, thereby enhancing its conversational and counselling capabilities. Further-

Table 1: Datasets used for models evaluation.

Task	Dataset	Format	Data Size	Language
Mental Health QA	USMLE-mental	Question-Answering	727	English
Mental Health QA	Crisis Response QA	Question-Answering	153	English
Suicide Risk Classification via Online Text	UMR Suicidality Dataset	Classification	1000	English
Psychological Counseling	Counseling Bench	Generation	229	English

more, a RAG process in the Proposal Agent was implemented, enabling it to generate therapy proposals based on mental health clinical guidelines. This allows the agent to update its context in real time. We also implemented an agent that detects suicide ideation levels, enabling the system to assess and respond with sensitivity to varying levels of ideation risk and warn the mental health professionals. Finally, the system architecture solution was implemented, creating each agent along with the interaction flow between them.

4.1 Solution

The Conversation Agent is the primary point of interaction between the user and our system, as shown in Figure 1. It communicates with the user via a chatbot interface or potentially a general companion agent, allowing for real-time exchanges where the user sends messages and receives responses. The agent's primary function is to engage with the user on a subject of their choice, supporting them through a conversational process that builds trust and facilitates the detection of mental health issues.

Once the user has finished the session, the Anonymization Agent takes over, responsible for anonymizing the messages exchanged by the user and the Conversation Agent. This involves replacing identifiable information such as names, emails, phone numbers, and addresses with tokens, ensuring that personal data remain securely stored on the server and is not shared over the Internet. By anonymizing the data, this approach significantly improves the security and confidentiality of user data.

The Expert Agents play a crucial role in the diagnosis of mental health disorders. The Symptoms Detection Agent focuses on identifying specific symptoms of mental health disorders by analyzing anonymized messages from the conversation, producing a symptom-based diagnosis along with a detailed rationale for each symptom detected. In contrast, the Clinical Issues Detection Agent takes a broader approach, assessing the conversation to identify overarching clinical mental health issues. Although both agents process the same conversation, the Symptoms Detection Agent focuses on individual symptoms, whereas the Clinical Issues Detection Agent provides a comprehensive and evidence-based evaluation of potential clinical issues as a whole. The Suicide Detection Agent detects whether suicide ideation is present or not, by processing anonymized messages. If it detects that suicide ideation is present, it assigns a risk level between 1 and 5. The Report Agent acts the consensus agent as its primary function is to receive diagnostics from the expert agents, compare these diagnostics, and extract only the overlapping or common information to produce the report. The Proposal Agent is responsible for generating comprehensive final proposals consisting of both a report and suggested therapeutic interventions for the detected symptoms generated by the RAG process described above.

This multi-agent system enables our system to detect mental health issues with greater accuracy, provide personalized support, and facilitate evidence-based treatment plans while maintaining user privacy and security.

Although this system is operated by independent agents, the sensitive nature of the domain (mental health) requires the integration of a human validator into the workflow, a technique known as "human-in-the-loop." In this particular case, the psychologist serves the function of validating the content that is to be provided to the Conversation Agent as context, see Figure 2. This specifically pertains to the therapy proposal and, in addition, the psychologist offers feedback based on the session conducted. This approach guarantees that the agent engaged in communication with the patient does not receive information that has been hallucinated or otherwise erroneous, generated by previous agents. This ensures the accuracy and quality of the content delivered. By incorporating the psychologist into this process, we ensure the dependability of the system in a crucial and sensitive domain such as mental health.

4.2 Conversational Agent Fine-Tuning

To optimize the LLaMA-3-8B-Instruct model for psychological counselling, the Quantized Low-Ranking Adaptation (QLoRA) (Dettmers et al., 2023) method was employed to facilitate efficient fine-tuning on a single NVIDIA GeForce RTX 4090 GPU with 24GB of memory. The model was quantized using 4-

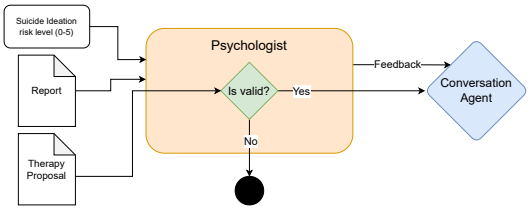

Figure 2: Human-in-the-Loop solution for MentalRAG system.

bit precision with the NormalFloat (NF4) data type, which enabled retention of near full-precision accuracy while significantly reducing memory usage. Additionally, bfloat16 was used as the compute data type, optimizing the model's efficiency during training by leveraging the GPU's tensor core.

The fine-tuning process involved training the model for 5 epochs using a batch size of 4 per device with gradient accumulation steps set to 1 and a learning rate of $2e^{-5}$ to optimize convergence. The model was trained on the Psych8k dataset, comprising authentic counselling conversations conducted by licensed professionals, which was tokenized using a maximum-length truncation strategy to ensure consistency. To manage memory usage more effectively, gradient checkpointing was implemented, allowing the model to train on larger sequences without exceeding the memory limitations of the GPU. The autoregressive instruction tuning objective was used to enhance the model's ability to generate responses that are professional and empathetic, with real-time monitoring using *wandb* tracking providing insights into the model's performance during training.

4.3 Retrieval Augmented Generation (RAG)

The RAG process began with the collection of clinical practice guidelines for prevalent mental health disorders, including bipolar disorder, eating disorders, psychosis and schizophrenia, major depressive disorder, post-traumatic stress disorder, substance use disorder, and suicide risk. Following this, an ingestion phase was conducted during which each document was splitted into smaller chunks. These segments were subsequently transformed into high-dimensional vectors using the LLaMA-3 embedding model and stored in the ChromaDB vector database for further analysis. To generate a comprehensive final proposal with suggested therapeutic interventions, the system performs similarity searches to retrieve the most relevant clinical information. This information is used to contextually enrich the input report and detected symptoms, enabling the system to generate therapy

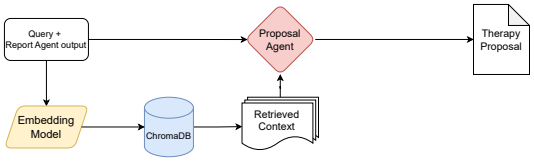

Figure 3: Overview of the RAG process implemented.

recommendations that are well informed and aligned with established clinical practices.

4.4 Agents Implementation

The implementation of a variety of agents, each designed to perform specific tasks within the therapeutic support framework, is described in Table 2. Each agent utilizes a large language model, such as Mental-LLaMA-3-8B-Instruct and GPT-4o, to address a variety of functions. These include conversation handling, anonymization, suicide ideation detection and clinical issue assessment. It should be noted that small open-source models are employed in expert agents primarily for the purpose of evaluating their performance in specific tasks. However, these models are fully interchangeable with state-of-the-art, off-the-shelf models, such as Claude 3.5 [1] Sonnet or GPT-4o [2], thereby providing flexibility for enhanced functionality as needed. This structured approach enables each agent to leverage model strengths that are relevant to its functional requirements.

Table 2: Overview of agents and corresponding models.

Agent	Model
Conversation Agent	Mental-LLaMA-3-8B-Instruct
Anonymization Agent	LLaMA-3-8B-Instruct
Suicide Ideation Detection Agent	GPT-4o
Clinical Issues Detection Agent	LLaMA-3-8B-Instruct
Symptoms Detection Agent	Mistral-7B
Report Agent	LLaMA-3-8B-Instruct
Proposal Agent	LLaMA-3-8B-Instruct

5 EVALUATION

The evaluation process was divided into two components: a models evaluation and an expert evaluation conducted by a PhD-qualified psychiatrist. The models evaluation was structured around three distinct tasks, namely a mental health knowledge task, a suicide risk classification task and an emotional support task. In order to complete this assessment, four datasets were used. Subsequently, the psychiatrist

[1] https://www.anthropic.com/news/claude-3-5-sonnet
[2] https://platform.openai.com/docs/models/gpt-4o

conducted a series of manual tests followed by an in-depth interview to gain insights into the system's performance.

5.1 Models Results

In order to evaluate the system, which follows an agentic workflow, all agents running different (or not) LLMs were tested in accordance with their predefined tasks. To illustrate, the Conversation Agent utilizes the instruction-tuned model, Mental-LLaMA-3-8B-Instruct, which is evaluated in accordance with psychological counselling metrics.

Models such as GPT-4o and GPT-4o-mini (OpenAI et al., 2024), LLaMA-3-8B-Instruct, Mistral-7B (Jiang et al., 2023) and our fine-tuned model, Mental-LLaMA-3-8B-Instruct, were evaluated on most of the tasks. A summarization of model details is shown in Table 3.

5.1.1 Mental Health Knowledge Task

In evaluating this task, the prompt is comprised of the question itself, as well as the available answer options, presented in a multiple-choice question-answering format. Given the objective nature of the questions, accuracy was used as the evaluation metric.

It is crucial to consider both the model size and the fine-tuning techniques applied when analysing the results from the Crisis Response QA and USMLE-mental datasets (see Table 4). The largest model, GPT-4o demonstrated superior performance on both tasks, achieving a score of 93.46% on Crisis Response QA and 81.16% on USMLE-mental. The model's substantial size enables it to process intricate tasks and generalise effectively across a range of domains, including medical reasoning and crisis intervention.

In comparison, GPT-4o-mini (6-13B parameters) performs almost as well on Crisis Response QA with 91.50%, but exhibits a more pronounced decline on the USMLE-mental task (66.30%). This demonstrates that although smaller models can still demonstrate proficiency in general tasks, their reduced capacity constrains their ability to process the complexity required for specialised medical tasks.

LLaMA-3-8B-Instruct (8B parameters) demonstrates a moderate performance, achieving 70.59% on Crisis Response QA and 52.27% on USMLE-mental. This indicates that mid-sized models are capable of handling general tasks but face challenges in more specialized domains, such as medical question-answering. Similarly, Mistral-7B (7B parameters) achieves 81.70% on Crisis Response QA but performs poorly on USMLE-mental (34.94%), indicating that models of this size are less well-suited to domain-specific tasks.

A noteworthy aspect for examination is Mental-LLaMA-3-8B-Instruct, which was tailored to meet the requirements of psychological counselling tasks through the utilization of a QLoRA adapter-based methodology (Dettmers et al., 2023). Although the model was optimized for mental health counselling, it exhibits suboptimal performance on the USMLE-mental dataset, achieving a score of 23.11%. This suboptimal performance can be attributed to the QLoRA quantisation method employed for fine-tuning. Although QLoRA is highly memory-efficient, the process of quantizing the model's weights to 4-bit precision introduces quantization errors, particularly in the presence of outliers or complex scenarios, such as medical question-answering. Such errors impact the model's capacity to generalize across domains beyond the one for which it was fine-tuned. This is particularly relevant in the context of the Mental-LLaMA-3-8B-Instruct model, which was fine-tuned on a psychological counselling instruction dataset. Consequently, it is more specialized and less capable of handling medical reasoning tasks, such as those in USMLE-mental.

In contrast, models such as GPT-4o-mini, which were not subjected to QLoRA, demonstrate considerably superior performance across a range of diverse tasks. This comparison highlights the trade-off between the efficiency gained through techniques such as QLoRA and the loss in generalization.

5.1.2 Suicide Risk Classification Task

The UMR Suicidality Dataset (Shing et al., 2018) was used to evaluate this task and the evaluation prompt is composed of the text to classify and the corresponding dataset labels. The metrics used were accuracy, recall, precision and F1-score.

This evaluation reveal significant performance variations across models in terms of their overall accuracy, recall, precision and F1-score, and label-specific performance. The models evaluated, GPT-4o, GPT-4o-mini, LLaMA-3-8B-Instruct, and Mistral-7B, demonstrated strengths in different areas but also showed notable shortcomings that limit their readiness for real-world application.

From the general metrics (Accuracy, Recall, F1-Score), Mistral-7B outperformed the other models with a higher score in accuracy (37.51%) and F1-score (35.19%), indicating its relatively better ability to correctly classify cases. GPT-4o and GPT-4o-mini showed lower overall performance across all metrics, with GPT-4o scoring the lowest at 19.61% in F1-score and GPT-4o-mini only slightly better. The low results

Table 3: Details of the models used in the evaluation phase.

Model	Modality	Parameters (Billions)	Context Length (Tokens)	Language	Access
GPT-4o	Pretrained	≈ 175	32000	English	API
GPT-4o-mini	Pretrained	$\approx 6-13$	$\approx 8000-16000$	English	API
LLaMA-3-8B-Instruct	Pretrained	8	8192	English	Weights
Mental-LLaMA-3-8B-Instruct	Fine-tuned	8	8192	English	Weights
Mistral-7B	Pretrained	8	8192	English	Weights

Table 4: PsyEval benchmark evaluation results.

	Crisis Response QA	USMLE-mental
Mental-LLaMA-3-8B-Instruct	71.24	23.11
LLaMA-3-8B-Instruct	70.59	52.27
GPT-4o	**93.46**	**81.16**
GPT-4o-mini	91.50	66.30
Mistral-7B	81.70	34.94

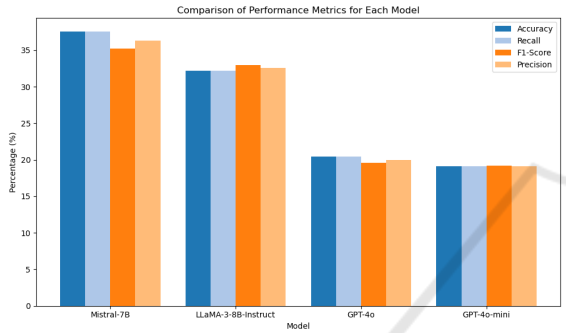

Figure 4: Models performance.

Table 5: Model Performance Metrics for Suicide Risk Classification.

Model	Accuracy (%)	Recall (%)	F1-Score (%)	Precision (%)
Mistral-7B	**37.51**	**37.51**	**35.19**	**36.31**
LLaMA-3-8B-Instruct	32.20	32.20	32.94	32.57
GPT-4o	20.40	20.40	19.61	20.00
GPT-4o-mini	19.10	19.10	19.21	19.15

may be attributed to the presence of OpenAI safeguards that prohibit the evaluation of text that may evoke suicidal ideation. LLaMA-3-8B-Instruct performed reasonably well in precision but showed room for improvement in overall classification effectiveness compared to Mistral-7B.

A review of the accuracy of the individual risk labels (see Figure 5) revealed the following:

LLaMA-3-8B-Instruct model demonstrated a noteworthy degree of accuracy in identifying cases classified as "Severe Risk" (43.71%) and "Moderate Risk" (38.29%). However, its performance in the "No Risk" (15.70%) and "Low Risk" (11.76%) categories was significantly lower, indicating a difficulty in detecting non-risk cases.

Mistral-7B also demonstrated proficiency in identifying "Moderate Risk" and "Severe Risk" cases, with accuracy rates of 51.59% and 25.53%, respectively. However, it exhibited notable challenges in accurately distinguishing "Low Risk" and "No Risk" scenarios, with accuracy rates of 2.12% and 15.24%, respectively.

GPT-4o demonstrated a high degree of accuracy in identifying individuals not at risk, with a score of 74.11% in the No Risk category. However, its performance in other risk categories was markedly poor, with low risk at 4.93% and severe risk at 9.74%. This suggests that it is unable to cope with more subtle forms of classification.

GPT-4o-mini exhibited a superior overall equilibrium compared to the GPT-4o, displaying a comparable potency in the "No Risk" category (63.49%) while also demonstrating marginal superiority in the "Low Risk" and "Moderate Risk" categories.

In general, Mistral-7B and LLaMA-3-8B-Instruct demonstrated superior performance in identifying individuals belonging to higher-risk categories, whereas GPT-4o exhibited a greater capacity for detecting those with no risk.

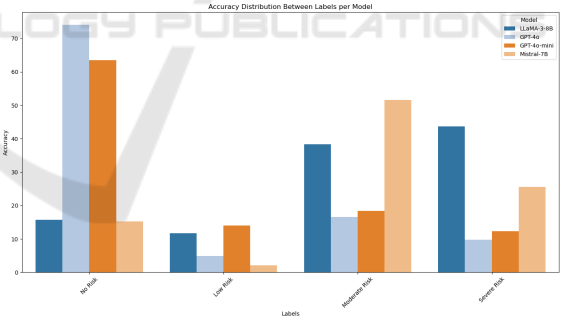

Figure 5: Model's performance on Classification Task (Metrics: Accuracy 100%).

Despite these insights, the overall accuracy scores across all models remain relatively low, with Mistral-7B reaching around 37.51% total accuracy, the highest among the models. LLaMA-3-8B-Instruct follows closely behind, while GPT models lag with an overall accuracy below 21%. These results suggest that the overall performance is not yet robust enough for deployment in real-world mental health scenarios, particularly where the stakes are high, such as suicide prevention.

Given the sensitivity of mental health appli-

cations, where misclassification could lead to either missed interventions or unnecessary escalations, these models need further refinement. Additionally, the relatively low performance of even the best models emphasizes the need for more training data and more comprehensive fine-tuning to enhance their ability to distinguish between subtle differences in risk levels, especially when applied to real-world suicide ideation detection tasks.

In summary, while the models evaluated in these classification tasks show potential, particularly in their specialized domains, their current accuracy levels are insufficient for real-world use in critical mental health settings.

5.1.3 Emotional Support Task

The Counselling Bench dataset (Liu et al., 2023) and its prompt design were used to evaluate this task, structured around seven core counseling skills that challenge large language models (LLMs) to respond empathetically and effectively in various scenarios such as anxiety, depression, and addiction (Jin et al., 2023). With 229 open-ended queries, it serves as a robust measure of LLM performance in delivering therapeutic support. The performance of Psych8k-7B, Mental-LLaMA-3-8B-Instruct, and LLaMA-3-8B-Instruct was compared across key counselling metrics, including information, direct guidance, approval and reassurance, and restatement, reflection, and listening (see Table 6). Despite Psych8k-7B being identified as the state-of-the-art model in the original paper (Liu et al., 2023), both LLaMA-3-based models have since surpassed it, with notable differences in performance.

1. **Information:** Mental-LLaMA-3-8B-Instruct outperforms with a score of 8.00, compared to LLaMA-3-8B-Instruct (7.50) and Psych8k-7B (6.25), indicating its superior accuracy in providing relevant psychological data.
2. **Direct Guidance:** Both Mental-LLaMA-3-8B-Instruct and LLaMA-3-8B-Instruct score 8.67, showing equal effectiveness in offering actionable advice, while Psych8k-7B lags with 6.67.
3. **Approval and Reassurance:** Mental-LLaMA-3-8B-Instruct (7.80) and LLaMA-3-8B-Instruct (7.60) perform similarly, with Psych8k-7B scoring 5.80, indicating a weaker capacity to provide emotional support.
4. **Restatement, Reflection, and Listening:** LLaMA-3-8B-Instruct (7.84) slightly surpasses Mental-LLaMA-3-8B-Instruct (7.68), with Psych8k-7B showing deficiencies at 6.12.
5. **Interpretation:** LLaMA-3-8B-Instruct leads with 8.00, followed by Mental-LLaMA-3-8B-Instruct (7.27) and Psych8k-7B (6.27), highlighting the latter's lower interpretive abilities.
6. **Self-Disclosure:** LLaMA-3-8B-Instruct scores 7.26, slightly higher than Mental-LLaMA-3-8B-Instruct (7.04), while Psych8k-7B trails at 5.61, indicating less effectiveness in personal insight.
7. **Obtain Relevant Information:** LLaMA-3-8B-Instruct leads with 8.00, followed by Mental-LLaMA-3-8B-Instruct (7.68), and Psych8k-7B underperforms with 6.39.

In conclusion, despite previously being the state-of-the-art model, Psych8k-7B has been surpassed by both Mental-LLaMA-3-8B-Instruct and LLaMA-3-8B-Instruct across almost all categories. This shift in performance demonstrates the impact of more effective fine-tuning and model development. In particular, the LLaMA-3-based models, notably Mental-LLaMA-3-8B-Instruct, exhibit enhanced capabilities in critical areas such as providing accurate information, offering actionable guidance, and providing emotional support. In consideration of the Conversation Agent's role in interacting with patients, which entails the utilization of personal details, therapy proposals and feedback, it is evident that Mental-LLaMA-3-8B-Instruct is the optimal selection. It is particularly adept at providing information and direct guidance, which are essential for advising patients based on their specific therapy context. The model's efficacy in providing approval and reassurance renders it an effective conduit for emotional support, which is a crucial element in sensitive counseling conversations.

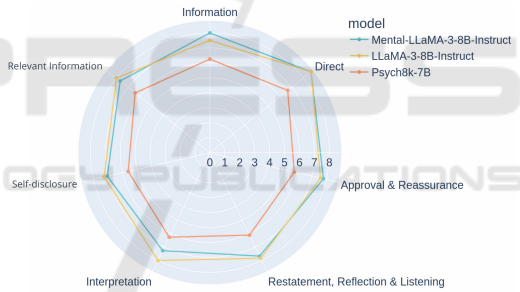

Figure 6: Model's performance on Emotional Support Task (Metrics: LLM-as-Judge 0-10 scale).

The model's proficiency in restatement, reflection, listening, and interpretation ensures its capacity to respond empathetically, thereby fostering rapport with patients through the reflection of their concerns. In comparison to Psych8k-7B, which it outperforms in nearly all metrics, our model is better equipped to ad-

Table 6: Counselling-bench evaluation (Scale: 0 to 10).

	Mental-LLaMA-3-8B-Instruct	LLaMA-3-8B-Instruct	Psych8k-7B
Information	**8.00**	7.50	6.25
Direct Guidance	**8.67**	**8.67**	6.67
Approval & Reassurance	**7.80**	7.60	5.80
Restatement, Reflection & Listening	7.68	**7.84**	6.12
Interpretation	7.27	**8.00**	6.27
Self-disclosure	7.04	**7.26**	5.61
Obtain Relevant Information	7.68	**8.00**	6.39

dress the subtleties of psychological counselling in a personalized and empathetic manner.

5.2 Expert Evaluation

The expert evaluation was conducted by a doctoral-level psychiatrist who began by simulating patient interactions with the system, addressing a range of general mental health issues. In this process, the psychiatrist engaged with the system's responses and analyzed its diagnostic assessments, reporting, and therapeutic recommendations. Following this simulation, an in-depth interview was conducted with the expert, focusing on their overall experience and insights into the system's efficacy and reliability. The key findings from this interview form the basis of the conclusions presented in this subsection, covering aspects such as empathy in responses, sensitivity to patient needs, human-AI collaboration, and clinical usability.

Empathy and Response Guidelines: The psychiatrist noted that while the system demonstrated empathy, responses should be more concise to avoid overwhelming distressed users. Emphasis should be placed on brief, active listening interactions, incorporating reflexive listening and validation techniques to balance information collection with reassurance.

Sensitivity and Quality of Responses: Sensitivity in addressing patients, especially those with suicidal ideation, was highlighted as critical. The agent should help patients express emotions and avoid providing unrealistic hope, instead offering grounded guidance and promoting calming strategies.

Human-AI Collaboration: A collaborative approach to suicide risk assessment was recommended, with both AI and clinicians evaluating the risk independently. The integration of established risk scales and attention to GDPR compliance were emphasized, particularly regarding when alerts are triggered.

Suicide Risk Detection Actions: The system should involve human intervention in high-risk cases, sending alerts to clinicians and directing patients to appropriate support services, while providing immediate emotional support and empathetic follow-up.

Clinical Usability: The interface was deemed intuitive, but improvements were suggested, such as better organization of patient records and quicker access to contact information for urgent cases.

Legal and Regulatory Issues: The psychiatrist raised concerns about the system's classification under EU medical device regulations, especially in high-risk areas like suicide prevention. Legal safeguards were recommended to clarify responsibility in cases of misclassification.

Future Improvements: Suggested enhancements included adding voice interaction and a clear introduction to the system's capabilities and limitations. The system could also be useful in academic settings to meet the growing demand for mental health services.

5.3 Discussion

The fine-tuned Mental-LLaMA-3-8B-Instruct model outperformed the original LLaMA-3-8B-Instruct in emotional support tasks, highlighting the effectiveness of task-specific fine-tuning for therapeutic communication. However, this specialization resulted in decreased performance on general medical knowledge tasks, illustrating the trade-off between specialization and generalization.

The GPT-4o model proved to be the best performer in both general knowledge tasks and the specialized domain of suicide risk detection, making it ideal for high-stakes tasks. Meanwhile, GPT-4o-mini offered a cost-effective alternative for routine tasks, maintaining competitive performance in less complex areas.

Improving the system's efficiency by optimizing model architectures and exploring model compression techniques will help reduce resource consumption and increase response times. Future efforts should also address regulatory concerns in clinical applications, ensuring compliance with emerging AI frameworks and providing transparency in AI decision-making.

Enhancing the RAG process and expanding the system's knowledge base in ChromaDB will further improve the accuracy and relevance of proposed therapies, making the system more robust in real-world

mental health scenarios.

Finally, our framework has a plug-and-play design that facilitates the use of both small open-source models and larger, state-of-the-art models. While the current implementation uses lightweight open source models to keep the system efficient and accessible, the framework is compatible with more advanced, pretrained models that can be integrated off-the-shelf as needed. This flexibility allows for scalable model upgrades, providing a pathway to incorporate more powerful agents as needed without changing the system architecture. Importantly, the expert agents processes only anonymized data, ensuring that any model replacement or upgrade respects user privacy and data security.

6 CONCLUSION

This paper presents the development and evaluation of the proof of concept MentalRAG system, an agentic framework for therapeutic support systems, with a particular focus on suicide risk assessment and psychological counselling. The system's refined models exhibited notable enhancements in the provision of empathetic and contextually tailored responses, particularly in delicate scenarios such as emotional distress and crisis management.

The MentalRAG system employs a multi-agent structure that leverages distinct specialized agents to handle various aspects of mental health support, from symptom identification to therapeutic proposal generation. The architecture of the system begins with the Conversation Agent, which interacts with users in real time, followed by the Anonymization Agent to protect user privacy. Expert agents, including the Symptoms Detection, Clinical Issues Detection, and Suicide Detection Agents, provide diagnostic insights by analyzing conversation data. The Report Agent synthesizes these diagnostics into a cohesive report, while the Proposal Agent generates personalized therapeutic recommendations. This multi-agent approach, combined with a human-in-the-loop validation process, ensures accurate, secure, and ethically guided support in mental health applications.

Despite its potential, several limitations remain. The evaluation was done by a single psychiatrist, in future works it needs more specialists to reduce the possibility of a bias. Also, the system's "black box" nature poses a risk in mental health applications, necessitating further safeguards to ensure safe and empathetic interactions, mitigating biases and errors due to its nature. Future work should focus on incorporating multimodal inputs (e.g., voice, video) to enhance interaction depth and integrating additional datasets for domain-specific fine-tuning.

By addressing these areas, the MentalRAG system can evolve into a more reliable, efficient, and ethically sound mental health support tool, better equipped to meet user needs while minimizing risks.

ACKNOWLEDGEMENTS

The authors are deeply grateful to Dr. Daniel Neto, whose contributions to the expert evaluation were invaluable. His detailed feedback and specialized knowledge as a PhD psychiatrist provided critical insights that significantly enhanced the quality and relevance of this project. This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UIDB/50021/2020 and FCT Project WSMART ROUTE+ reference 2022.04180.PTDC.

REFERENCES

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Capital, S. (2024). What's next for AI Agentic Workflows ft. Andrew Ng of AI Fund.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. (2023). Palm: scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24(1).

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma,

S., et al. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Dubey, A., Jauhri, A., and et al., A. P. (2024). The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Huang, K., Altosaar, J., and Ranganath, R. (2019). Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jin, H., Chen, S., Wu, M., and Zhu, K. Q. (2023). Psyeval: A comprehensive large language model evaluation benchmark for mental health. *arXiv preprint arXiv:2311.09189*.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Liu, J. M., Li, D., Cao, H., Ren, T., Liao, Z., and Wu, J. (2023). Chatcounselor: A large language models for mental health support. *ArXiv*, abs/2309.15461.

Liu, Y. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

National Alliance on Mental Illness (2023). Mental health by the numbers. https://nami.org/mhstats. Last checked on 2023-12-12.

OpenAI et al. (2024). Hello gpt-4o. https://openai.com/index/hello-gpt-4o/. Last checked on 2024-10-12.

OpenAI, R. et al. (2023). Gpt-4 technical report. *ArXiv*, 2303:08774.

Patil, S. G., Zhang, T., Wang, X., and Gonzalez, J. E. (2024). Gorilla: Large language model connected with massive APIs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Shing, H.-C., Nair, S., Zirikly, A., Friedenberg, M., Daumé III, H., and Resnik, P. (2018). Expert, crowdsourced, and machine assessment of suicide risk via online postings. In Loveys, K., Niederhoffer, K., Prud'hommeaux, E., Resnik, R., and Resnik, P., editors, *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, LA. Association for Computational Linguistics.

Singh, A., Ehtesham, A., Kumar, S., and Khoei, T. T. (2024). Enhancing ai systems with agentic workflows patterns in large language model. In *2024 IEEE World AI IoT Congress (AIIoT)*, pages 527–532.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. (2023). Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023a). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H., Martin, L., Stone, K. R., et al. (2023b). Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Xu, X., Yao, B., Dong, Y., Gabriel, S., Yu, H., Hendler, J., Ghassemi, M., Dey, A. K., and Wang, D. (2024). Mental-LLM: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32.

Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., and Wang, L. (2023). Mm-react: Prompting chatgpt for multimodal reasoning and action. *ArXiv*.