Constructing a Knowledge-Guided Mental Health Chatbot with LLMs

Xi Fan SA22011218@MAIL.USTC.EDU.CN
Lishan Yang YLISHAN@MAIL.USTC.EDU.CN
Xiangyu Wang* SA312@USTC.EDU.CN
Derui Lyu DRLV@MAIL.USTC.EDU.CN
Huanhuan Chen* HCHEN@USTC.EDU.CN
University of Science and Technology of China Hefei, China

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

The global shortage of mental health resources has severely impacted the ability to address psychological distress, affecting approximately 658 million people. Despite the effectiveness of psychotherapy and counseling, less than 35% of those in need receive help. Traditional conversational agents often lack emotional support, leading to mechanical interactions that detract from user experience. This paper introduces the "Mental Health Chatbot", a conversational agent based on a pre-trained large language model. This chatbot innovatively uses retrieval-augmentation techniques to extract relevant knowledge from psychological diagnostics and treatment manuals, providing tailored psychotherapeutic interventions. It effectively identifies mental disorders and their severity, suggesting appropriate interventions. Evaluated through pre-trained model similarity comparisons, large language model scoring, and expert assessments, results show that the Mental Health Chatbot enhances the accuracy of smaller models and accelerates the inference speed of larger models through retrieval-augmentation. The optimized training process enables more human-like interactions, improving user experience and demonstrating the chatbot's potential and practical application in addressing mental health challenges.

Keywords: Large Language Model, Conversational Agent, Knowledge-Guided, Retrieval-Augmentation, Psychotherapy

1. Introduction

The global shortage of mental health resources poses a significant challenge, with an estimated 658 million people worldwide suffering from mental distress, a number that has risen by 50% over the past three decades Collaborators et al. (2022). Although psychotherapy and counseling are effective for treating conditions like anxiety, depression, Post-Traumatic Stress Disorder, and eating disorders, only about 35% of those affected receive treatment, and fewer than 25% consult professionals Chen and Cardinal (2021). In response, conversational agents or chatbots are increasingly vital in mental health, recognized for their effectiveness in alleviating negative emotions and promoting healthy lifestyles Narynov et al. (2021). Additionally, symptom checker chatbots represent a significant advancement, assessing symptoms through simulated human interactions and highlighting the growing potential of conversational technology in mental health care You et al. (2023).

Building on traditional methods, the emergence of large language models has opened new avenues for conversational agents. The LLaMA series, exemplifying models based on

© 2024 X. Fan, L. Yang, X. Wang*, D. Lyu & H. Chen*.

the Transformer architecture and generative pre-training, can predict subsequent words by analyzing vast datasets. These models are fine-tuned using niche-specific data and human feedback Christiano et al. (2017) , sparking widespread discussions about their potential applications in mental health interventions van Heerden et al. (2023). While preliminary studies have demonstrated the promise of patient-centric chatbots, these robots still face challenges in replicating the nuances of human emotional communication Zhu et al. (2024); Lyu et al. (2024). Moreover, the robust capabilities of these models offer prospects for supporting systematic therapeutic interventions by professional counselors, an area that remains to be further explored in the academic realm.

However, the complexity and diversity of mental health issues present significant challenges for developing and evaluating chatbots. Difficulties include obtaining accurate data due to privacy concerns and the subjective nature of mental illness symptoms. Additionally, reliance on screening tools like the PHQ-9, which lack diagnostic precision and overlook individual personality traits, complicates effective patient feedback even for experienced professionals Salaheddin and Mason (2016).

In response, this paper introduces a novel conversational agent, the "Mental Health Chatbot", based on a pre-trained large language model. This chatbot interacts with users to identify their mental disorder and severity, and provides tailored psychological treatments using corresponding treatment manuals.

The main contributions of this study are as follows:

- Utilizing psychological diagnostic and treatment manuals to guide the Mental Health Chatbot in accurately identifying users' disorders and suggesting treatment plans.
- Developing a comprehensive evaluation metric for assessing the performance of the chatbot in disease classification, session management, and their integration.
- The chatbot enables real-time interactions, generating personalized treatment strategies, thereby assisting in mental health care and reducing the workload of psychotherapists.

2. Related Works

LLMs represent a pivotal advancement in the development of general artificial intelligence. Despite their unprecedented performance across a broad range of tasks, these models still face numerous challenges, such as hallucinations Bang et al. (2023), command adherence Bai et al. (2022), and handling long texts An et al. (2023). To address these issues, researchers have proposed a method known as "retrieval-enhanced generation", which strengthens the model's capabilities by integrating external assistance Borgeaud et al. (2022). Retrievers play a crucial role in connecting LLMs with necessary external components, enabling them to perform a variety of downstream tasks. There are several types of retrievers, each designed to specialize in different functions: 1) Knowledge retrievers provide support for knowledge-intensive tasks by supplying external information Wang et al. (2023, 2024). 2) Tool retrievers select appropriate tools, allowing LLMs to interact effectively with the physical world Chen et al. (2023a, 2024). 3) Example retrievers locate and use pre-cached

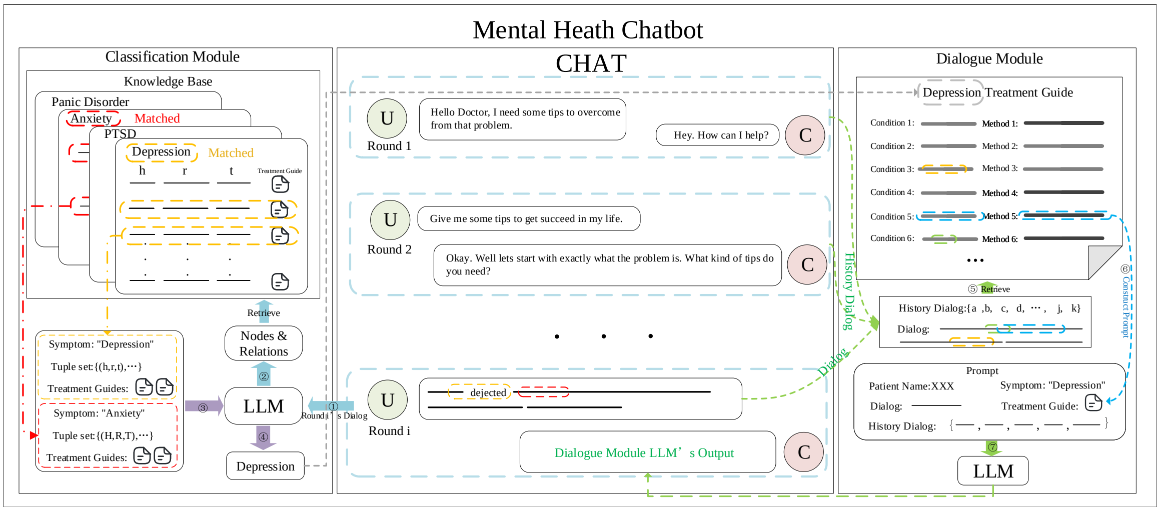

Figure 1: Workflow of the Mental Health Chatbot system.

examples to automatically generate LLM prompts, facilitating learning in specific contexts Ban et al. (2023a,b). 4) Memory retrievers help models collect and utilize information beyond the immediate conversational context to support long-sequence generation Bertsch et al. (2024).

Many self-screening methods for mental health are widely used today Brown et al. (2016). While online help is preferred by many , existing tools mainly rely on closed-ended questions for limited evaluations, potentially overlooking crucial information compared to face-to-face or open-ended questions. Innovative research Liu et al. (2022) has introduced an AI-based online mental state examination (MSE) via web browsers. Experience Sampling Method (ESM) is extensively utilized in Human-Computer Interaction (HCI) research for various physical and mental health screenings, including self-reporting Parkinson's Disease symptoms, chronic pain, and designing health technologies for Bipolar Disorder Adams et al. (2018). Meanwhile, the CaiTI project employs open-ended questions, allowing users to freely discuss any topic, a rarity in academic research. Building on this, modern research continues to explore the use of LLMs for predicting and assessing mental health states through online text data. Researchers have evaluated several LLMs in mental health prediction tasks Radwan et al. (2024) and developed systems using RoBERTa Liu et al. (2019) and Llama-65b Touvron et al. (2023) to classify mental disorders, severe depression, self-rated depression, and anxiety based on time-series multimodal features.

Traditional conversational therapy methods, such as rule-based Medeiros and Bosse (2018) and data-driven approaches Yao et al. (2022), often result in mechanical interactions that lack necessary emotional support, adversely affecting user experience. However, the advent of LLMs offers a pivotal solution to these shortcomings. LLMs, trained on extensive datasets, accumulate vast prior knowledge and enhanced reasoning capabilities. Advanced models based on the Transformer architecture, such as GPT-4, PaLM 2, and LLaMA2, demonstrate superior performance in medical-related natural language processing tasks. For instance, GPT-4 has facilitated medical innovations including managing patient discharge records, summarizing clinical trials, and providing ethical guidelines Waisberg et al. (2023).

Google's specialized medical LLMs, Med-PaLM 2 Singhal et al. (2023), deliver precise medical consultation responses. Moreover, researchers are exploring LLMs for psychological counseling services. A study Nie et al. (2022) utilized GPT-3 to develop a home-based AI therapist that detects abnormalities in psychological states and daily functions, providing comforting responses to users. Additionally, an AI-assisted tool Lai et al. (2023) based on the Chinese pre-trained model WenZhong has been developed for effective Q&A in psychological counseling scenarios. The potential of ChatGPT is also being explored for simulating dialogues between psychiatrists and patients with mental disorders, highlighting its capability in mimicking complex interpersonal interactions Chen et al. (2023b).

3. Methodology

Our constructed framework for psychological counseling dialogue models primarily consists of two modules: a mental disorder classification model based on psychological counseling dialogues and a conversational robot model based on psychological counseling dialogues. The specific framework and flowchart are as shown in Figure 1.

3.1. Classification Model Using Counseling Dialogues

The mental disorder classification module aims to categorize patient speech to determine the presence of mental health issues and further refine the diagnosis to specific types of mental disorders. This provides crucial evidence for diagnostic work.

3.1.1. Knowledge graph retrieval augmentation in mental health diagnosis

This study enhances NLP model performance by introducing retrieval-augmentation technology, utilizing additional knowledge resources , particularly knowledge bases, preferred for their retrieval speed and computational efficiency, are defined as $\mathcal{G} = (V, E)$ where V represents nodes (concepts, objects, events, etc.), and E represents the semantic relationships between these entities. Following DSM-5 guidelines, A mental disorder knowledge base optimized for retrieval was constructed:

$$\mathcal{G} = \bigcup_{i=1}^{n} \{(v_i, \mathcal{A}(v_i))\} \tag{1}$$

where $\mathcal{G}$ denotes the knowledge graph, n the number of knowledge points, v_i a node, and $\mathcal{A}(v_i)$ the attributes of node v_i.

The process for retrieval augmentation using knowledge graphs is described in the following contents. The input dataset is initially defined as $\mathcal{D}_1$.

Retrieval Preprocessing: For each data entry $d_i \in \mathcal{D}_1$, a large language model L is used to extract relevant nodes and relations:

$$q_i = L(P_q(d_i)), \quad r_i = L(P_r(d_i)) \tag{2}$$

where P_q and P_r are prompts designed to extract node sets Q and relation sets R, respectively, enhancing the model's comprehension and response quality.

Prompt $P_1'(d_i)$

You are a mental health expert. Extract key nodes/relations from the psychological counseling dialogue: "d_i".Guidance: List all important entities/relations in the sentence.

Knowledge Matching: After retrieval preprocessing, The sets Q and R within the previously constructed knowledge graph $\mathcal{G}$ are utilized to search for matching relationships and node sets. The search operation is conducted as follows:

$$A_q = \bigcup_{i=1}^{|Q|} \text{search}(V, q_i), \quad V \in \mathcal{G}, \quad A_r = \bigcup_{i=1}^{|R|} \text{search}(E, r_i), \quad E \in \mathcal{G}, \tag{3}$$

where the search function is defined as:

$$\text{search}(G, x_i) = \{y \in G | \text{sim}(y, x_i) > \theta\}, \tag{4}$$

where $\text{sim}(y, x_i)$ denotes the similarity between element y and x_i, with θ being the preset similarity threshold.

After obtaining the appropriate sets A_q and A_r, their neighborhoods are determined as follows:

$$N(A_q) = \bigcup_{q \in A_q} N(q), \quad N(A_r) = \bigcup_{(v,w) \in A_r} (\{v, w\}) \cup N(v) \cup N(w), \tag{5}$$

where $N(v) = \{u \in V : (u, v) \in E \text{ or } (v, u) \in E, \quad E \in \mathcal{G}\}$.

Having obtained their first-order neighborhoods, $N(A_q)$ and $N(A_r)$ together form a subgraph $G_A = \{N(A_q) \cup N(A_r)\} \subset G$. Then traverse each edge $(v, w) \in G$ to assemble a set $S = \mathcal{A}(w) \cup \mathcal{A}(v) \cup \mathcal{A}(e_{(v,w)})$. This set is used to construct a prompt to inquire with the large model L_{mat} whether this set matches the data d_i as follows:

Prompt $P_2'(S, d_i)$

You are a mental health expert. Output a number from [0,1.0] indicating the match degree between data d_i and the attribute set of relationship tuple S, where closer to 1 means S better represents d_i.

A hyperparameter θ is chosen so that when the output value from the model L_{mat} exceeds θ, the relationship (v, w) can be included into $\mathcal{V}_A$. Ultimately, $\mathcal{V}_A$ represents the matched knowledge.

Knowledge Extraction: After identifying the relevant node set $\mathcal{V}_A$, the next step involves extracting information by accessing each node's attributes, represented as $K_A = \{\mathcal{A}(v) | v \in \mathcal{V}_A\}$, which includes all relevant node attributes.

Knowledge Integration: The extracted knowledge is then integrated to enhance the model's response using an integration function $I(K_A, \mathcal{M}) = f(\mathcal{M}, \text{agg}(K_A))$, where $\mathcal{M}$ is the model's current state and $\text{agg}(K_A)$ compiles the knowledge into an integrable form. The function f performs the integration, updating the model state to $\mathcal{M}'$.

Dynamic Decision Support: The updated model state $\mathcal{M}'$ is then used for decision-making, described by the decision function $g(\mathcal{M}', q)$, which generates responses or recommendations based on the updated state and original query.

3.1.2. Mental Disorder Classification Model

In everyday psychological dialogue processes, initial conversations often contain extraneous information such as greetings that are not relevant to the task at hand. To enhance data efficiency and model performance, retrieval-enhanced training is initiated from the second round of dialogue. Specifically, the psychological dialogue dataset, represented by $\mathcal{D}$, includes themes $\mathcal{D}^{'}_{theme}$ and dialogues $\mathcal{D}^{'}_{dialog}$. Data from the second round onward in each session is utilized, and the focus for retrieval-enhanced training is on optimizing the model's learning process with this selected data. The setup is formally defined as follows:

$$\mathcal{D}' = \bigcup_{i=1}^{m} \{s_i^t | t \geq x\}, \quad L_{tr} = \mathrm{train}(L, P), \quad P(\mathcal{P}_{sys}, \mathcal{G}, \mathcal{D}') \tag{6}$$

where m is the number of sessions, x indicates the dialogue round, L_{tr} is the trained model, L is the base language model, and P is our training prompt that integrates external knowledge from $\mathcal{G}$ into the training process.

Prompt $P_3'(\mathcal{P}_{sys}, \mathcal{G}, \mathcal{D}')$

As a psychological counseling expert with access to the mental disorder diagnosis knowledge base $\mathcal{G}$, analyze the consultation themed $\mathcal{D}'_{theme}$ and content $\mathcal{D}'_{dialog}$. Use the knowledge base and your expertise to assess if the consultant exhibits any psychological disorder symptoms and provide your professional judgment.

Training Process: The training function train(L, P) involves the base model L adjusting its parameters through exposure to P, enhancing its task-specific performance. The steps include: 1) Initialization: Starting with the base model L and initializing all necessary parameters. 2) Data Integration and Processing: Utilizing $\mathcal{D}'$ for forward and backward propagation, integrating data from $\mathcal{G}$ to improve data comprehension. 3) Optimization Iteration: Conducting multiple training iterations as specified by P to optimize parameters. 4) Model Evaluation and Adjustment: Periodically assessing and adjusting model performance and training strategies during the training process.

3.2. Conversational Robot Model for Psychological Counseling Dialogues

The conversational module is central to the system, integrating patient dialogue with the treatment plan. It takes a conversational prompt as input and delivers the final dialogue outcome. This involves understanding and analyzing the patient's speech, applying the treatment plan, and generating responses and recommendations. The module's output, presented as session results, offers personalized psychological counseling services.

3.2.1. Text-Based Retrieval-Augmentation Method

The overall algorithm is outlined in Algorithm 1. A treatment manual $\mathcal{M}$, consisting of n levels, is expressed as $\mathcal{M} = \{M_1, M_2, M_3, \ldots, M_n\}$, where each level provides a text description. The most detailed information is contained in the final level M_n. The objective is to locate the most relevant treatment instructions for a specific query by sequentially searching through each level.

Algorithm 1 Text Retrieval-Augmentation Method for Hierarchical Treatment Manuals

```
Require: $M_1, M_2, \ldots, M_n$ - Levels of the treatment manual, $n$ - Total number of levels, $a$
         - Similarity threshold, $k$ - Number of top matching entries to retain per level
Ensure: Final selected treatment information
Initialize the similarity threshold $a$  for $i = 1$ to $n$ do
    Initialize the current level entry set $S = \emptyset$  foreach entry $j \in M_i$ do
        Calculate similarity $\mathrm{sim}(q, j)$  if $sim(q, j) > a$ then
            Add $j$ to the set $S$
        end
    end
    if $|S| < k$ then
        repeat
            $a \leftarrow a - \delta$  Update the set $S$
        until $|S| \geq k$
    end
    $S_i \leftarrow \mathrm{top}_k(S)$  if $i < n$ then
        $M_{i+1} \leftarrow S_i$
    else
        return $S_i$
    end
end
```

3.2.2. Agent Module for Determining Psychological Disorders

In this section, a conversational robot was trained with prior knowledge of patients' psychological disorders, enabling quantitative comparisons with other pre-trained models. As observed in previous sections, initial dialogue rounds typically contain excess irrelevant information, such as greetings. The same method was employed to enhance training efficiency using the dataset designated as $\mathcal{W}'$. For dialogue data excluded from $\mathcal{W}'$, represented as $\overline{\mathcal{W}'} = \mathcal{W} - \mathcal{W}'$, a distinct strategy was incorporated into the training process. For each $\overline{w'i}$ in $\overline{\mathcal{W}'}$, the prompt $P'_{\overline{\mathcal{W}'}}(w)$ was used for training:

Prompt $P'_{\overline{\mathcal{W}'}}(w)$

You are an expert in psychological counseling. A patient with a psychological disorder is consulting you, and their dialogue content is: $\overline{w'_i}$. Based on prior patient information and your expertise, provide a professional response.

For each element w'_i in $\mathcal{W}'$, the following prompt $P_{\mathcal{W}'}(w, \delta, \mathcal{T})$ is used for training:

Prompt $P'_{\mathcal{W}'}(w, \delta, \mathcal{T})$

You are an expert in psychological counseling. A patient with δ is consulting you, and their dialogue content is: w'_i. The treatment plan for δ is: $\mathcal{T}$. Using prior patient information and your expertise, provide a response.

where δ represents the predefined type of the patient's psychological disorder, and $\mathcal{T}$ is the treatment plan related to δ retrieved from the treatment manual $\mathcal{M}$ using the text retrieval-Augmentation method.

3.3. Knowledge-Guided Mental Health Chatbot

In practical settings, diagnosing psychological disorders is dynamic, relying on real-time patient interactions rather than preset information. Initially, the system uses generic prompts $P_{\overline{\mathcal{W}'}}(w)$ to maintain conversation continuity. From the x^{th} round, it classifies diseases based on session content using a classification model. If a patient's description is ambiguous or unrelated to psychological disorders, it's marked as "Unrelated", and text retrieval-augmentation is used to find relevant treatment plans. Conversely, if a specific disease δ is diagnosed, targeted searches are conducted in the treatment manual related to that disease.

The classification module continually monitors patient expressions, adjusting classifications with new data to ensure accuracy and timeliness. The system uses prompts $P_{\mathcal{W}'}(w, \delta, \mathcal{T})$ to guide response generation based on the treatment plan $\mathcal{T}$ and confirmed disease δ, improving interactiveness and diagnostic precision. The module updates the patient's disorder classification throughout the interaction, which continues until the dialogue ends. This ongoing interaction enhances adaptability and user experience, ensuring timely and personalized responses.

4. Experiments

4.1. Experimental Design

4.1.1. Data Description

This research utilizes two datasets. Dataset A, derived from the cleaned efaqa dataset Chen et al. (2022), includes 1667 entries labeled with psychological disorders such as Depression, Anxiety, Bipolar Disorder, PTSD, Panic Disorder, Eating Disorder, and Unrelated, as shown in Figure 2. It consists of dialogues from online mental health forums involving multiple counselors, suitable for classification research due to varied dialogue styles and unverified counselor professionalism. Dataset B, from the ESConV dataset Liu et al. (2021), contains 1300 dialogues between psychological consultants and counselors, focusing on mental counseling services, as depicted in Figure 2. It is ideal for training conversational agents with its more standardized dialogue data. Both datasets, from existing research, have distinct purposes: Dataset A for classification and Dataset B for conversational agents.

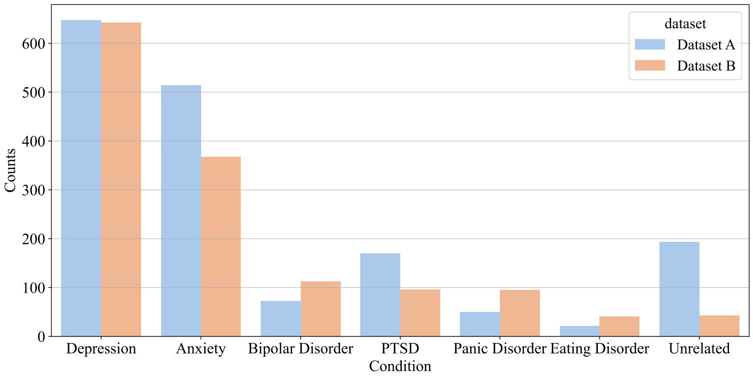

Figure 2: Distribution of psychological conditions across datasets.

4.1.2. Metrics

We employ three evaluation metrics to assess the quality of outputs from our framework.

Pre-trained Model Quantification: Performance is measured by cosine similarity between model outputs and actual dataset values, using Word2Vec Church (2017), RoBERTa Liu et al. (2019), and SBERT Wang and Kuo (2020). A penalty mechanism corrects overestimated scores for outputs longer than the ground truth.

Large Language Model Quantification: We use GPT-4 to score discrepancies from 0 to 100 based on closeness to ground truth. The prompt $P(d_t, d_s)$ evaluates similarity, expertise, fluidity, and empathy, each up to 25 points.

Prompt $P'_5(d_t, d_s)$

As a language expert, your task is to evaluate sentence d_s against reference d_t and assign a score up to 100 points based on: 1. Similarity—semantic closeness to d_t, scoring up to 25 points. 2. Expertise—accuracy of knowledge in d_s, with expert presentation scoring up to 25 points. 3. Fluidity—language fluency, with smoother expressions scoring up to 25 points. 4. Empathy—compassion or understanding shown by d_s, scoring up to 25 points.

Human Evaluation Metric: A subjective scoring method by independent reviewers evaluates outputs on professionalism, fluency, and empathy, each rated up to 100 points for a comprehensive assessment.

4.1.3. Experimental Setup and Knowledge Graph Preparation

In this study's experimental setup, we utilized a server with dual-slot Intel(R) Xeon(R) Platinum 8352Y processors (2.20GHz) and six NVIDIA(R) Tesla(R) A100 GPUs. Low-Rank Adaptation (LoRA) was the training method, with a batch size of 8 per GPU over 5 epochs and a learning rate of 0.0003, incorporating a warm-up covering 1% of the epochs. The benchmark model was Qwen2-7B.

For the classification module, we developed a knowledge graph:

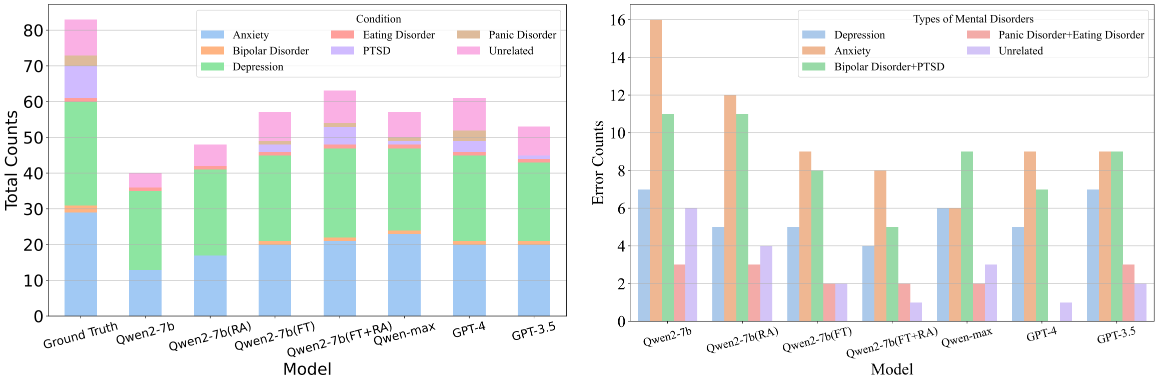

Figure 3: **Left**: Accuracy distribution of mental health diagnoses across computational models. **Right**: Error distribution across models for different mental disorders.

* "FT" denotes a model that has undergone fine-tuning. The same applies below.

Table 1: Overall accuracies of classification modules.

	Qwen2-7b	Qwen2-7b (FT)	Qwen2-7b(RA)	Qwen2-7b (FT+RA)	Qwen-max (RA)	GPT-4 (RA)	GPT-3.5 (RA)
Accuracy (%)	50.60	68.67	57.83	74.70	68.67	73.49	66.26
Precision (%)	25.78	65.03	47.82	78.15	74.92	79.02	54.11
Recall (%)	37.24	62.46	43.05	69.64	60.43	75.01	55.13

* "RA" denotes a model that has utilized a retrieval-augmentation strategy.. The same applies below.

$$D = (w, (r, (s_1, d_1), (s_2, d_2), \ldots)), \tag{7}$$

where D represents the mental condition, w the condition name, r the related symptoms, and (s_i, d_i) specific symptoms and explanations.

We streamlined the knowledge base with 248 items and about 1,000 nodes from Dataset A to enhance retrieval efficiency and ensure professionalism and practicality.

4.2. Experimental Results

4.2.1. Classification Module

In this section, Dataset A was used to create training, testing, and evaluation sets with an 18:1:1 ratio for fine-tuning large language models. After training, accuracy on the test set increased from 50.60% to 74.70%. We further evaluated performance using popular large models (GPT-4, GPT-3.5, Qwen-max), summarized in Table 1. The distribution of correct and incorrect classifications is shown in Figure 3.

In experiments with Dataset B, it was found unsuitable for classification tasks. We divided 200 samples into equal training and test sets. Post-training, the classification accuracy reached 53%, as depicted in Figure 4. This underscores the impact of dataset selection on model performance.

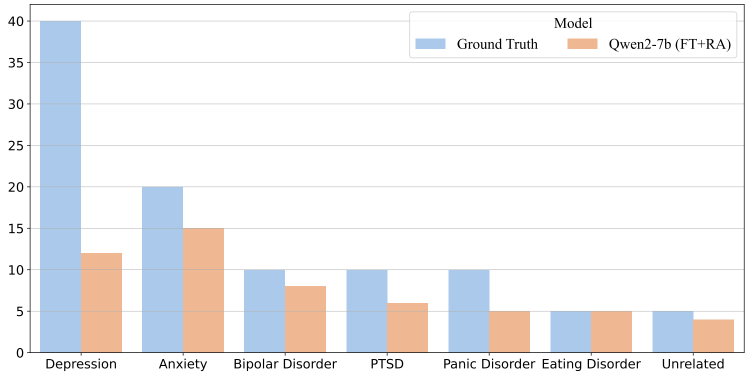

Figure 4: Classification results on dataset B using trained model compared to ground truth.

Table 2: Comparison of GPT-4 and human evaluation scores for different models

	Qwen2-7b	Qwen2-7b (FT)	Qwen2-7b (RA)	Qwen2-7b (FT+RA)	Qwen-max (RA)	GPT-4 (RA)	GPT-3.5 (RA)
GPT-4 Score	55.55	62.74	56.29	63.58	59.95	65.47	62.27
Human Evaluation Score	64.75	71.25	63.25	77.50	70.25	76.30	72.50

4.2.2. Dialogue Module

In this section, we developed a conversational agent for patients with psychological disorders and compared it quantitatively with pre-trained models. We used data from the third round of conversations in dataset $\mathcal{W}$, forming $\mathcal{W}'$ for efficient training by excluding greetings and small talk. We also processed excluded data $\overline{\mathcal{W}'}$ to maximize data utilization.

The strategy improved the agent's ability to generate relevant responses, enhancing its usefulness in psychological counseling. We evaluated the model using cosine similarity with pre-trained models (Word2Vec, RoBERTa, SBERT), compared to untrained and non-augmented models, as well as large models like GPT-4, GPT-3.5, and Qwen-max (Figure 5). Penalties for longer sentences significantly improved performance. Additionally, GPT-4 scored outputs from 0 to 100 for alignment with ground truth, summarized in Table 2.

We used a scoring-based subjective evaluation method to understand the effects of model outputs, with independent evaluators scoring professionalism, fluency, and empathy for 30, 30, and 40 points respectively. The average scores are documented in Table 2. Results show our model slightly outperforms GPT-4, mainly because it produces optimally lengthed outputs, unlike others that generate lengthy and superfluous content. However, the semantic differences between our model and others are minimal, aligning with findings from pre-trained model evaluations using Word2Vec, RoBERTa, and SBERT.

4.2.3. Knowledge-Guided Mental Health Chatbot

Based on prior sections, we constructed a prediction framework that starts categorizing from the third dialogue round in each session, split into two parts based on the use of retrieval-augmentation techniques and model types. We established various experimental setups considering whether modules are fine-tuned or use large models like GPT-4, GPT-3.5, and Qwen-max, and the application of retrieval-augmentation. We used three evaluation

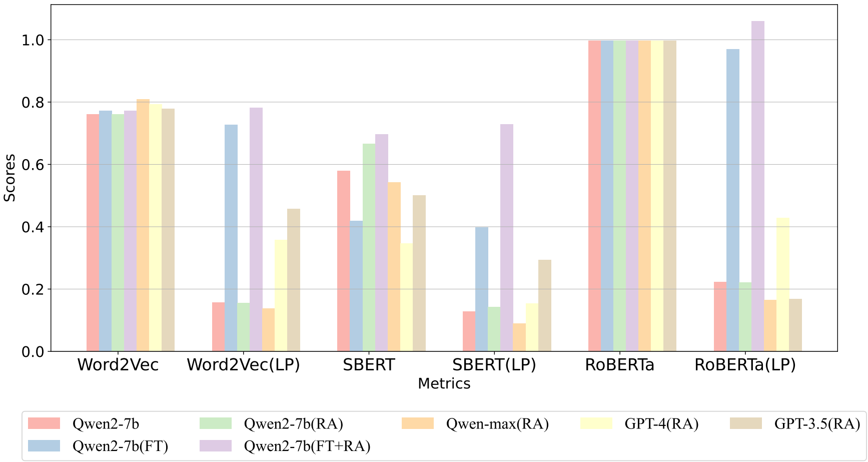

Figure 5: Comparison of model outputs using various pre-trained models and strategies.

* "LP" denote the application of a penalty mechanism for longer sentences .

Table 3: Experimental results of the Knowledge-Guided Mental Health Chatbot.

Classification Module	Conversation Module	Word2Vec/RoBERTa/SBERT Similarity	GPT-4 Score	Expert Average Score
Qwen2-7b (FT+RA)	Qwen2-7b (FT+RA)	0.6557 / 0.6762 / 0.8826	61.28	72.50
Qwen2-7b (FT+RA)	Qwen2-7b (RA)	0.1324 / 0.2124 / 0.1417	51.68	56.75
Qwen2-7b (FT+RA)	Qwen2-7b (FT)	0.3423 / 0.3625 / 0.4528	57.65	64.28
Qwen2-7b (FT+RA)	Qwen2-7b	0.1120 / 0.1223 / 0.0936	51.48	54.23
Qwen2-7b (RA)	Qwen2-7b (FT+RA)	0.4566 / 0.4671 / 0.5537	56.43	63.27
Qwen2-7b (FT)	Qwen2-7b (FT+RA)	0.6035 / 0.6437 / 0.7312	60.58	69.68
Qwen2-7b	Qwen2-7b (FT+RA)	0.4336 / 0.4512 / 0.5238	53.64	60.17
GPT-4 (RA)	GPT-4	0.3375 / 0.1175 / 0.3865	59.43	71.33
GPT-4 (RA)	GPT-4 (RA)	0.3463 / 0.1336 / 0.4023	60.36	72.75
GPT-3.5 (RA)	GPT-3.5	0.3323 / 0.2765 / 0.1023	58.67	62.17
GPT-3.5 (RA)	GPT-3.5 (RA)	0.3425 / 0.2654 / 0.1136	59.72	66.31
Qwen-max (RA)	Qwen-max	0.1215 / 0.0687 / 0.1433	57.63	66.35
Qwen-max (RA)	Qwen-max (RA)	0.1312 / 0.0753 / 0.1541	58.37	67.23

metrics, including penalties for long sentences in similarity assessments. Results are in Table 3, offering clear experiment configurations for straightforward comparison and analysis.

The results show our framework's performance lies between GPT-3.5 and GPT-4. Retrieval-augmentation significantly helps smaller models but has less effect on larger models. Additionally, the two modules in our framework influence each other, with the dialogue module having a more significant impact than the classification module. This is likely due to overlapping treatment methods for psychological disorders. Fine-tuning combined with retrieval augmentation leads to the most significant improvement in performance.

Still, larger models with retrieval-augmentation infer faster than those without. During inference, GPT-4 exhibited an average time per query of 2.12 seconds with retrieval augmentation, compared to 3.45 seconds without it. A similar pattern was seen with GPT-3.5, where the average time was 2.47 seconds per query with retrieval augmentation versus 3.09 seconds without it. Our framework met expected outcomes.

5. Conclusion

This study presents a mental health chatbot framework designed to efficiently provide psychological counseling and treatment to patients with mental disorders. The framework leverages advanced large language models, fine-tuning, and retrieval-augmentation techniques to accurately diagnose psychological illnesses during interactions. The retrieval-augmentation technique enables the chatbot to extract relevant knowledge from psychological diagnostics and treatment manuals, offering tailored interventions. Diagnostic results and treatment plans are integrated as prefixes in dialogues to facilitate targeted interactions.

In practice, the framework starts with standard interactions for the first three rounds, then adjusts dialogue based on classification outputs. Evaluated using pre-trained model similarity comparisons, large language model scoring, and expert assessments, results show the framework slightly outperforms GPT-4, especially when using the fine-tuned Qwen2-7b model. Retrieval-augmentation enhances the accuracy of smaller models and speeds up inference for larger models. Optimized training reduces redundant outputs, making interactions more human-like and improving user experience. These advancements demonstrate the chatbot's potential and practical utility in addressing mental health challenges.

6. Ethical Considerations

The data used in this study are sourced from the efaqa and ESConV datasets, both of which have accompanying ethical statements in the relevant literature. These datasets have undergone de-identification processes to ensure compliance with ethical standards and to prevent any privacy concerns. This information was omitted in the initial draft, and we acknowledge this oversight. The final version of the paper addresses this by clearly stating that all data used adhere to ethical guidelines. Future work will explore the model's ethical implications, maintaining a focus on responsible AI development.

7. Acknowledgements

This work was supported in part by the National Key R&D Program of China (No. 2021ZD0111700), and in part by the National Nature Science Foundation of China (No. 62137002, 62176245).

References

Alexander T Adams, Elizabeth L Murnane, Phil Adams, Michael Elfenbein, Pamara F Chang, Shruti Sannon, Geri Gay, and Tanzeem Choudhury. Keppi: A tangible user interface for self-reporting pain. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

Chenxin An, Shansan Gong, Ming Zhong, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. L-Eval: Instituting standardized evaluation for long context language models. *arXiv preprint arXiv:2307.11088*, 2023.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, and Huanhuan Chen. Causal structure learning supervised by large language model. *arXiv preprint arXiv:2311.11689*, 2023a.

Taiyu Ban, Xiangyu Wang, Xin Wang, Jiarun Zhu, Lvzhou Chen, and Yizhan Fan. Knowledge extraction from national standards for natural resources: A method for multi-domain texts. *Journal of Database Management*, 34(1):1–23, 2023b.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew Gormley. Unlimiformer: Long-range transformers with unlimited length input. *Advances in Neural Information Processing Systems*, 36, 2024.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *Proceedings of the International Conference on Machine Learning*, pages 2206–2240, 2022.

Menna Brown, Noelle O'Neill, Hugo van Woerden, Parisa Eslambolchilar, Matt Jones, Ann John, et al. Gamification and adherence to web-based mental health interventions: a systematic review. *JMIR Mental Health*, 3(3):e5710, 2016.

Lyuzhou Chen, Taiyu Ban, Xiangyu Wang, Derui Lyu, and Huanhuan Chen. Mitigating prior errors in causal structure learning: Towards LLM driven prior knowledge. *arXiv preprint arXiv:2306.07032*, 2023a.

Lyuzhou Chen, Xiangyu Wang, Taiyu Ban, Muhammad Usman, Shikang Liu, Derui Lyu, and Huanhuan Chen. Research ideas discovery via hierarchical negative correlation. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2):1639–1650, 2024.

Shanquan Chen and Rudolf N Cardinal. Accessibility and efficiency of mental health services, united kingdom of great britain and northern ireland. *Bulletin of the World Health Organization*, 99(9):674, 2021.

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

Youren Chen, Yong Li, and Ming Wen. Chinese psychological qa database and its research problems. In *Proceedings of the 9th International Conference on Dependable Systems and Their Applications*, pages 786–792. IEEE, 2022.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017.

Kenneth Ward Church. Word2vec. *Natural Language Engineering*, 23(1):155–162, 2017.

GBD 2019 Mental Disorders Collaborators et al. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. *The Lancet Psychiatry*, 9(2):137–150, 2022.

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

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*, 2021.

Yanchen Liu, Stephen Xia, Jingping Nie, Peter Wei, Zhan Shu, Jeffrey Andrew Chang, and Xiaofan Jiang. AIMSE: Toward an AI-based online mental status examination. *IEEE Pervasive Computing*, 21(4):46–54, 2022.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Derui Lyu, Lyuzhou Chen, Taiyu Ban, Xiangyu Wang, Qinrui Zhu, Xiren Zhou, and Huanhuan Chen. Leveraging multisource label learning for underground object recognition. *IEEE Transsctions on Geoscience and Remote Sensing*, 62:1–14, 2024.

Lenin Medeiros and Tibor Bosse. Using crowdsourcing for the development of online emotional support agents. In *Highlights of Practical Applications of Agents, Multi-Agent Systems, and Complexity: The PAAMS Collection: International Workshops of PAAMS*, pages 196–209, 2018.

Sergazy Narynov, Zhandos Zhumanov, Aidana Gumar, Mariyam Khassanova, and Batyrkhan Omarov. Chatbots and conversational agents in mental health: a literature review. In *2021 21st International Conference on Control, Automation and Systems*, pages 353–358, 2021.

Jingping Nie, Hanya Shao, Minghui Zhao, Stephen Xia, Matthias Preindl, and Xiaofan Jiang. Conversational AI therapist for daily function screening in home environments. In *Proceedings of the 1st ACM International Workshop on Intelligent Acoustic Systems and Applications*, pages 31–36, 2022.

Ahmad Radwan, Mohannad Amarneh, Hussam Alawneh, Huthaifa I Ashqar, Anas AlSobeh, and Aws Abed Al Raheem Magableh. Predictive analytics in mental health leveraging LLM embeddings and machine learning models for social media analysis. *International Journal of Web Services Research*, 21(1):1–22, 2024.

Keziban Salaheddin and Barbara Mason. Identifying barriers to mental health help-seeking among young adults in the UK: a cross-sectional survey. *British Journal of General Practice*, 66(651):e686–e692, 2016.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Alastair C van Heerden, Julia R Pozuelo, and Brandon A Kohrt. Global mental health services and the impact of artificial intelligence–powered large language models. *JAMA Psychiatry*, 80(7):662–664, 2023.

Ethan Waisberg, Joshua Ong, Mouayad Masalkhi, Sharif Amit Kamran, Nasif Zaman, Prithul Sarker, Andrew G Lee, and Alireza Tavakkoli. GPT-4: a new era of artificial intelligence in medicine. *Irish Journal of Medical Science (1971-)*, 192(6):3197–3200, 2023.

Bin Wang and C-C Jay Kuo. SBERT-WK: A sentence embedding method by dissecting BERT-based word models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2146–2157, 2020.

Xiangyu Wang, Taiyu Ban, Lyuzhou Chen, Muhammad Usman, Yifeng Guan, Derui Lyu, Jian Cheng, Huanhuan Chen, Cyril Leung, and Chunyan Miao. Decentralised knowledge graph evolution via blockchain. *IEEE Transactions on Services Computing*, 17(1):169–182, 2023.

Xiangyu Wang, Taiyu Ban, Lyuzhou Chen, Xingyu Wu, Derui Lyu, and Huanhuan Chen. Knowledge verification from data. *IEEE Transactions on Neural Networks and Learning Systems*, 35(3):4324–4338, 2024.

Binwei Yao, Chao Shi, Likai Zou, Lingfeng Dai, Mengyue Wu, Lu Chen, Zhen Wang, and Kai Yu. D4: a Chinese dialogue dataset for depression-diagnosis-oriented chat. *arXiv preprint arXiv:2205.11764*, 2022.

Yue You, Chun-Hua Tsai, Yao Li, Fenglong Ma, Christopher Heron, and Xinning Gui. Beyond self-diagnosis: how a chatbot-based symptom checker should respond. *ACM Transactions on Computer-Human Interaction*, 30(4):1–44, 2023.

Qinrui Zhu, Derui Lyu, Xi Fan, Xiangyu Wang, Qiang Tu, Yibin Zhan, and Huanhuan Chen. Multi-model consistency for LLMs' evaluation. In *Proceedings of the 2024 International Joint Conference on Neural Networks*, pages 1–8, 2024.