AI Psychiatrist Assistant: An LLM-based Multi-Agent System for Depression Assessment from Clinical Interviews

Adam Greene[1,2*] AGREENE46@STUDENT.GSU.EDU
Neviah Blair[1,2*] NBLAIR7@STUDENT.GSU.EDU
Samin Mahdipour Aghabagher[3] SAMIN.MAHDIPOUR.AGHABAGHER@UMONTREAL.CA
Simmi Kumari[1,2] SKUMARI4@STUDENT.GSU.EDU
Micheal W. Schlund[1,2] MSCHLUND@GSU.EDU
Alex Fedorov[4] AVFEDOR@EMORY.EDU
Vince D. Calhoun[1,2,4,5] VCALHOUN@GSU.EDU
Xinhui Li[1,5†] XINHUILI@GATECH.EDU
Rogers F. Silva[1,2†] RSILVA@GSU.EDU

[1] *Tri-institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State University, Georgia Institute of Technology, Emory University, Atlanta, GA, USA*
[2] *Georgia State University, Atlanta, GA, USA*
[3] *Université de Montréal, Montréal, QC, Canada*
[4] *Emory University, Atlanta, GA, USA*
[5] *Georgia Institute of Technology, Atlanta, GA, USA*

Abstract

Depression is one of the most common mental disorders yet remains underdiagnosed. Large language models (LLMs) have shown promise in their ability to understand the semantic meaning behind medical text and automate clinical workflows through collaborative agents. Here, we propose an LLM-based multi-agent system to diagnose depression symptoms from clinical interview transcripts. Our system integrates four agents: (1) a qualitative assessment agent that identifies symptoms and risk factors, (2) a judge agent that evaluates qualitative assessment through iterative self-refinement, (3) a quantitative assessment agent that predicts clinical scores using a novel embedding-based few-shot prompting approach, and (4) a meta-review agent that integrates outputs into a comprehensive overview of a patient's mental state. The qualitative assessment agent provided coherent, specific, and reasonably accurate assessment, as evaluated by both the human expert and the judge agent. The quantitative assessment agent with few-shot prompting showed an average mean absolute error of 0.619 for symptom prediction versus 0.796 in zero-shot prompting, while the meta-review agent achieved a binary classification accuracy of 78%, comparable to that of a human expert. Our system could serve as a consultant for psychiatrists and psychologists, offering an alternative perspective on patients' mental health conditions, and thus establishing a foundation for future work on agent-aided clinical support.

Keywords: Large Language Models, Agents, Mental Health, Depression

Data and Code Availability We used the Distress Analysis Interview Corpus – Wizard of Oz (DAIC-WOZ) dataset (Gratch et al., 2014), which is available at https://dcapswoz.ict.usc.edu/. Code is available at https://github.com/trendscenter/ai-psychiatrist.

Institutional Review Board (IRB) This study used a public and de-identified dataset and did not require IRB approval.

1. Introduction

Mental disorders are among the leading contributors to the global burden of disease, posing substantial challenges to both individuals and public health systems. Depression affects more than 300 million people worldwide, and its prevalence continues to rise

[*] These authors contributed equally to this work.
[†] These authors jointly supervised this work and share correspondence.

© 2025 A. Greene, N. Blair, S.M. Aghabagher, S. Kumari, M.W. Schlund, A. Fedorov, V.D. Calhoun, X. Li & R.F. Silva.

(World Health Organization, 2025). The clinical diagnosis and treatment of mental disorders typically rely on semi-structured interviews, in which clinicians assess symptoms of patients according to their verbal responses and behavioral signals. However, this process is inherently subjective, with outcomes varying across clinicians based on their training, experience, and interpretation of the criteria (Meyer et al., 2001). Such inconsistency results in severe consequences. At the patient level, it delays accurate diagnosis, worsens clinical outcomes, and increases financial burdens due to prolonged care. At the system level, it further strains already overburdened medical services, reduces comparability across cases, and escalates overall healthcare costs. This gap, rooted in the text and speech-based nature of clinical interviews, can be addressed by leveraging artificial intelligence (AI) capabilities in natural language and speech analysis, offering a promising path toward more reliable and explainable mental health assessment.

Recent advances in large language models (LLMs) (Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020; Touvron et al., 2023) have empowered AI agents to perform complex language-based tasks (Yao et al., 2023; Schick et al., 2023; Wang et al., 2024). These LLM-based agentic systems show promising potential in clinical workflows that involve the understanding and reasoning of natural language (see Appendix A for related work). Previous related studies focused on isolated tasks, such as synthesizing interview data (Yin et al., 2025) or predicting the severity of depression (Galatzer-Levy et al., 2023; Sadeghi et al., 2024), rather than proposing a structured, explainable, and clinically grounded assessment system.

Here, we propose a multi-agent system based on LLMs to analyze clinical interviews from both quantitative and qualitative aspects to support depression assessment. The proposed system includes four specialized and collaborative agents: (1) a qualitative assessment agent that interprets interview transcripts and identifies risk factors, (2) a judge agent that evaluates qualitative assessment through iterative self-refinement, (3) a quantitative assessment agent that predicts standardized clinical scores using a novel embedding-based few-shot prompting approach, and (4) a meta-review agent that integrates outputs into diagnostic suggestions. In addition, a human expert evaluation is conducted to guarantee clinical validity. This system is designed to support psychiatrists by providing automated initial evaluations while ensuring that professional oversight remains central to the process. Our contributions are summarized as follows:

- We design an LLM-based multi-agent system that integrates four collaborative agents to analyze clinical interviews from both quantitative and qualitative perspectives.
- We develop an embedding-based few-shot prompting strategy that outperforms a naive zero-shot approach in symptom-specific prediction.
- We demonstrate that the agentic system predicts depression severity from available information with human-level accuracy.

2. Methods

2.1. Dataset

We used the Distress Analysis Interview Corpus – Wizard of Oz (DAIC-WOZ) dataset (Gratch et al., 2014), a multimodal benchmark comprising semi-structured clinical interviews in North American English. This database aims to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder (PTSD). Each interview was conducted by a virtual interviewer called Ellie, controlled by a human interviewer. Participants included people with and without depressive symptoms. Before the interview, the participants completed a set of questionnaires, including the eight-item Patient Health Questionnaire (PHQ-8) (Kroenke et al., 2009). The PHQ-8 assesses eight depression symptoms over the past two weeks, each scored from 0 (not at all) to 3 (nearly every day), resulting in a total score of 0 to 24, with a higher score indicating greater severity of depression. The PHQ-8 total score can be converted to five severity categories: no significant depressive symptoms ($0-4$), mild depressive symptoms ($5-9$), moderate ($10-14$), moderately severe ($15-19$), and severe ($20-24$). A PHQ-8 total score of 10 serves as the threshold for major depressive disorder (MDD; PHQ-8 ≥ 10). The dataset includes 189 interview sessions divided into a training set (107 participants; 63 males, 44 females), a development set (35 participants; 16 males, 19 females), and a test set (47 participants; 23 males, 24 females). We used 142 subjects from the training and development sets for our main analysis, as item-wise PHQ-8 scores were not available in the test set.

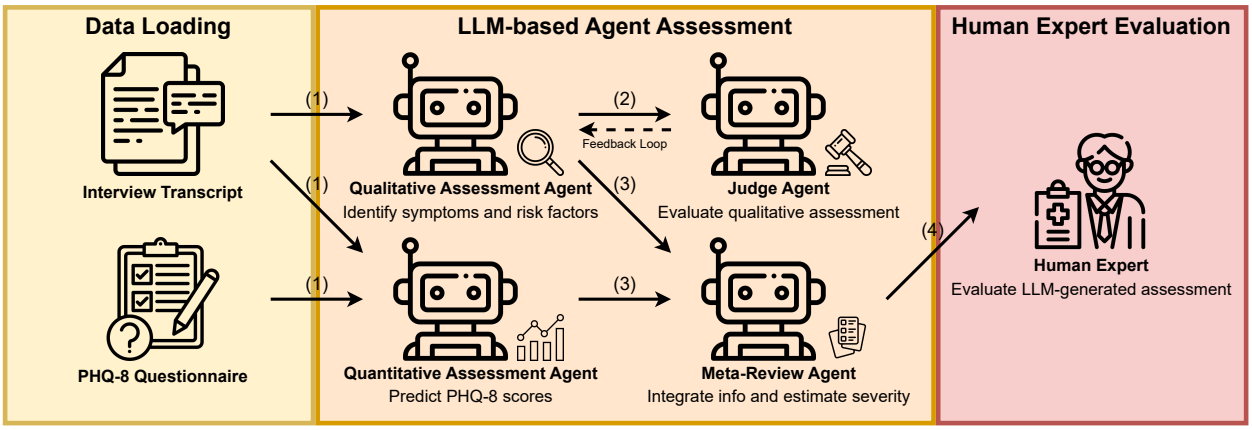

Figure 1: **Overview of the multi-agent system.** (1) The interview transcript and PHQ-8 scores are loaded into the system, and then distributed to the agents for both qualitative and quantitative assessments. (2) The qualitative assessment agent summarizes symptoms and risk factors and sends them to the judge agent for iterative evaluation. (3) The meta-review agent integrates both qualitative and quantitative information and estimates overall severity. (4) The final LLM-generated assessment is reviewed by a human expert for validation.

2.2. Models

We utilized a state-of-the-art open-weight language model, Gemma 3 with 27 billion parameters (Gemma 3 27B) (Gemma Team et al., 2025). For the embedding-based few-shot prompting approach, we used Qwen 3 8B Embedding (Zhang et al., 2025) due to its superior performance on the Massive Text Embedding Benchmark (MTEB) leaderboard (Muennighoff et al., 2022).

2.3. System Design

We propose an LLM-based multi-agent system for assessing depression symptoms from clinical interviews (Figure 1). In this system, an agent is an autonomous LLM-driven component that (1) maintains local task state or memory, (2) follows a local decision policy to determine its next step, and (3) operates in an observe-reason-act loop while communicating with other agents or humans through structured messages (Xi et al., 2023; Yao et al., 2022; Wu et al., 2023). This design follows recent LLM-agent literature that conceptualizes agents as cognitive units coupling perception, reasoning, and action, and demonstrates collaborative coordination across multiple agents (Li et al., 2023; Chen et al., 2023). Each component thus satisfies the operational definition of an LLM-based agent by maintaining internal state, executing local decision policies, and exchanging structured messages.

The system includes four collaborative agents. The qualitative assessment agent identifies PHQ-8 symptoms and risk factors based on the interview transcript. The judge agent then evaluates these assessments and, if the evaluation score falls below a threshold, triggers a feedback loop to reassess. The quantitative assessment agent predicts PHQ-8 scores using an embedding-based few-shot prompting approach. The meta-review agent integrates all information to estimate the severity of the depressive symptoms. As a whole, these agents form an interactive multi-agent pipeline wherein decisions and information propagate sequentially. Finally, a human expert evaluates the LLM-generated assessments.

2.3.1. Qualitative Assessment

We prompted Gemma 3 27B to assume the role of a psychiatrist tasked with generating an objective and concise assessment based on the participant's interview transcript, using exact quotes from the transcript as evidence. Four domains were created to categorize participant data based on the details provided during the interview: a summary of PHQ-8 symptoms (frequency or duration of depression symptoms if applicable), biological factors (for example, familial mental health history, pre-existing mental health

conditions), social factors (for example, interpersonal relationships or conflicts), and risk factors (for example, isolation, stressors) relevant to the patient's mental health. Examples of each domain were included in the prompt to help guide the model.

To verify the validity of the agent's output, we developed a judge agent, inspired by the LLM-as-a-judge approach (Zheng et al., 2023), to evaluate qualitative assessment. The judge agent scored on four metrics: specificity (avoiding overly generic statements), completeness (coverage of symptoms and their frequencies), coherence (logical consistency), and accuracy (alignment with PHQ-8). Scores were assigned on a five-point Likert scale, where a higher score is better (see Appendix B for detailed definitions of the proposed metrics). For each subject, the judge agent analyzed the qualitative assessment and explained its reasoning, using the original transcript as reference. The judge prompt included instructions to prevent hallucinations, explicitly asking the model to base its explanations on information present in the transcript and qualitative assessment, rather than making assumptions or introducing external content. The model was required to provide exact quotations from the transcript as evidence. It then assigned a numeric score ($1-5$, higher is better) based on the rubric provided. The ordering in this step (reasoning first and scoring later) was intended to provide the model with more time to think and to mitigate potential hallucinations. The judge agent took both the original transcript and the LLM-generated qualitative assessment as input for evaluation. When an original evaluation score was below four, the judge agent triggered an automatic feedback loop to further improve the qualitative assessment. The qualitative assessment agent was then given the judge agent's evaluation output and instructed to use it as a remedial tool in the re-assessment (Madaan et al., 2023). The new, modified version of the original assessment based on the evaluation feedback was re-assessed by the judge agent. The feedback loop was limited to a maximum of 10 iterations per transcript.

2.3.2. Quantitative Assessment

The quantitative assessment part of our system utilized few-shot prompting via embeddings to predict PHQ-8 scores across all eight items. We compared the few-shot prompting approach with a zero-shot approach, where we provided only the transcript to the model without any examples. Our few-shot prompting method consists of two main parts: evidence retrieval and evaluation. During evidence retrieval, Gemma 3 27B was provided with information about the PHQ-8, a given transcript, and instructions to retrieve relevant evidence associated with each individual PHQ-8 question. If no relevant evidence was found for a given PHQ-8 item, the model produced no output. Then, we took the outputted pieces of evidence and concatenated them into strings (one for each PHQ-8 question, up to eight strings total if all output was produced). Next, we used the embedding backend to retrieve relevant example transcript segments from our training set, along with their associated PHQ-8 scores by means of cosine similarity with the concatenated strings (see Section 2.4.2 for more details). During evaluation, Gemma 3 27B was subsequently prompted with information about the PHQ-8, the retrieved transcript segments, and additional instructions. The additional instructions were to search for direct transcript evidence related to each PHQ-8 question, cross-reference that evidence with the example transcript segments, and evaluate the frequency and severity of the input transcript while thinking. After that, the LLM provided reasoning for its thought process, and only then did it output an integer score ($0-3$, higher is more severe), or "N/A" (not available) if no relevant evidence was available for a given PHQ-8 question. Model performance was evaluated by computing the mean absolute error (MAE) between predicted and ground-truth PHQ-8 scores.

2.3.3. Meta Review

We designed a meta-review agent to integrate information from the interview transcript, the qualitative assessment, and the quantitative assessment. The meta-review agent integrated the qualitative and quantitative assessments, and then predicted a severity category based on depressive symptoms discussed in the interview. The model was instructed to use the available information to infer the participant's condition as accurately as possible.

2.3.4. Human Expert Evaluation

To ensure clinical validity, we included a human expert assessment as part of the evaluation process. A PhD-level researcher (X.L.) evaluated the qualitative assessment agent's outputs using the same four-metric and five-point rubrics used in the judge agent (Appendix B), with guidance from a senior clinical

expert (M.W.S.). The human evaluator also predicted a PHQ-8 total score for each participant based on the transcript. The PHQ-8 prediction performance of the human expert was compared with that of the meta-review agent.

2.3.5. Agentic System

The agentic system consists of four specialized and collaborative agents: a qualitative assessment agent, a quantitative assessment agent, a judge agent, and a meta-review agent. All agents use open-weight LLMs through the Ollama API (https://ollama.com/). Instead of using hard-coded logic, the system leverages a dynamic architecture coordinated by a central server. Each agent runs autonomously and communicates back to the server when their task is completed. The full assessment pipeline executes in approximately one minute on a MacBook Pro with an Apple M3 Pro chipset. This computation efficiency is especially noteworthy, as it enables deployment on easily accessible consumer hardware, eliminating the need for expensive cloud-based GPU infrastructure or specialized computing resources, and thereby supporting broader clinical adoption.

2.4. Embedding-based Few-shot Prompting

Our quantitative assessment first evaluated a zero-shot prompting approach and then proposed an embedding-based few-shot prompting approach to further improve performance, as we hypothesized that by providing real examples for reference, the model prediction would be more grounded in reality.

2.4.1. Data Splitting

We split 142 subjects with eight-item PHQ-8 scores from the DAIC-WOZ database into training, validation, and test sets. The training set was used to retrieve relevant evidence, the validation set to select hyperparameters, and the test set to evaluate model performance. We stratified the data according to PHQ-8 total scores and gender. We used a 41% training (58 participants), 30% validation (43), and 29% test (41) split, which provided a fairly balanced distribution (Appendix C).

2.4.2. Few-shot Prompting Workflow

In our few-shot workflow, we utilized pre-chunked and pre-embedded transcripts from our training set as a knowledge base to retrieve few-shot examples. We utilized Qwen 3 8B Embedding to embed the transcript chunks. Each transcript in the training set was segmented into chunks of N_{chunk} lines using a sliding window with step size of 2 lines. Gemma 3 27B was used to retrieve relevant evidence from the given transcript for each PHQ-8 question using direct prompting. Next, we concatenated all the retrieved evidence for a given question into a string and embedded it with Qwen 3 8B Embedding. We then compared that embedding to our pre-embedded transcripts, via cosine similarity, and pulled the N_{example} most similar example chunks. For each chunk, we identified its associated participant ID in the dataset and attached its ground truth PHQ-8 score. After this process, we obtained a string of up to $N_{\text{example}} \times 8$ reference chunks (N_{example} for each PHQ-8 question). That reference string and its corresponding score were incorporated into the main quantitative assessment prompt, which was then used to instruct Gemma 3 27B to generate predictions. Hyperparameters were systematically evaluated, including the chunk size N_{chunk}, the number of reference examples N_{example}, and the embedding dimension $N_{\text{dimension}}$. Then, optimal values were selected based on the performance of the validation set (Appendix D).

2.5. Prompt Design

Our prompt design emphasized objectivity, conciseness, and clinical interpretability (Luo et al., 2024). Because the interviews were conversational and not prestructured, we formatted the outputs using XML tags and structured JSON/CSV files to enhance readability and interpretability. For each agent, we designed both the system and user prompts. The system prompt defined the model's role as a psychiatrist and analysis principles, while the user prompts included task-related instructions. The model was instructed to provide exact quotes and explanations for its assessment or evaluation. For transcripts with incomplete information, PHQ-8 scores that could not be determined were recorded as "N/A".

2.6. Single-Prompt Experiment

To evaluate the benefits of the multi-agent system over a single-agent setup, we merged the qualitative, quantitative, and meta-review prompts into a single prompt, using only one API call to complete all three tasks. The program took a transcript as input and generated structured outputs corresponding to each component.

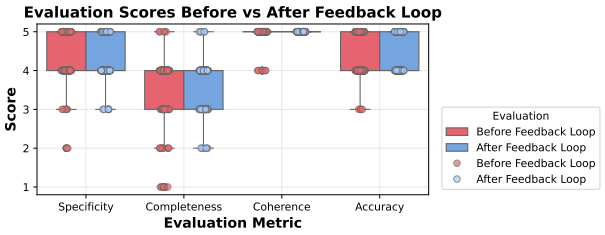

Figure 2: **Qualitative evaluation scores before and after feedback loop across four metrics on 142 participants.** The box shows the quartiles of the scores and the whiskers show the rest of the distribution.

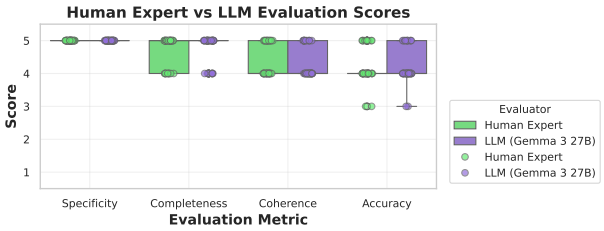

Figure 3: **Qualitative evaluation scores between human expert and LLM-based judge on the test set.** The box shows the quartiles of the scores and the whiskers show the rest of the distribution.

3. Results

3.1. Qualitative Assessment

As shown in Figure 2, the model without the feedback loop achieved the highest score on coherence (mean ± standard deviation: 4.96 ± 0.20), implying its assessments generally aligned with the expected analysis criteria. Specificity (4.37 ± 0.62) and accuracy (4.33 ± 0.53) both achieved relatively high average scores, implying the model tended to avoid vague language and aligned with PHQ-8 criteria. However, there were still several cases of vague language and moderate alignment with PHQ-8 criteria. Completeness scored the lowest (3.61 ± 0.85), suggesting the model often omitted relevant details. After the feedback loop, all metrics showed improvement, with the completeness metric showing the largest increase (3.72 ± 0.61). Coherence, specificity, and accuracy slightly improved as well (5.00 ± 0.00, 4.38 ± 0.58, and 4.36 ± 0.48, respectively). Although the feedback loop mechanism could not achieve perfect scores, possibly due to lack of mention for certain symptoms in the interviews, it demonstrates the ability to consistently improve evaluation scores across all proposed metrics.

To further examine clinical validity of the system, we compared the evaluation scores between the LLM judge and a human expert on our test set with 41 participants (Figure 3). LLM and human evaluations showed high overall agreement: specificity and coherence scores were comparable, whereas the LLM judge tended to give slightly higher ratings for completeness and accuracy, compared with the human judge.

3.2. Quantitative Assessment

We first identified optimal hyperparameters for the embedding-based few-shot prompting according to the performance of the validation set: $N_{\text{chunk}} = 8$, $N_{\text{example}} = 2$, $N_{\text{dimension}} = 4096$ (Appendix D). We then confirmed that the embedded reference transcripts formed severity-based clusters in the t-SNE visualization and that most retrieved and target chunks showed consistent PHQ-8 scores (Appendix E). Next, we conducted a quantitative assessment on 41 subjects in our test set. Subjects without sufficient evidence (determined by Gemma 3 27B) were excluded from the assessment. With optimized hyperparameters, Gemma 3 27B achieved an average MAE of 0.619 when predicting PHQ-8 scores, but in 50% of cases it was unable to provide a prediction due to insufficient evidence (Figure 4). The distribution of available scores was not even: certain symptoms, such as appetite, had few available scores, while others, such as sleep quality, had available scores for nearly all subjects. This reflects the variability in the content of the interview, with some symptoms discussed more frequently than others.

In comparing zero-shot and few-shot performance, the few-shot approach generally yielded slightly lower MAE values, indicating that the incorporation of reference examples improved prediction performance (Figure 5). Using the optimal few-shot parameters, the few-shot approach achieved an average MAE of 0.619 whereas the zero-shot approach achieved 0.796. This represents an MAE reduction of 22% compared to zero-shot. However, for symptoms such as poor appetite and moving slowly, MAE performance was highly variable due to substantially fewer subjects with available scores.

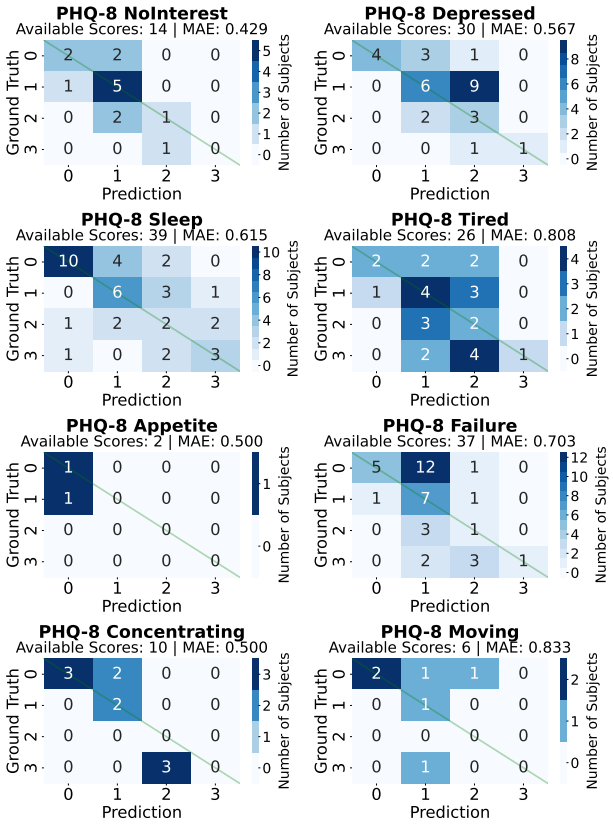

Figure 4: **Confusion matrices showing PHQ-8 prediction performance with optimal hyperparameters on the test set.**

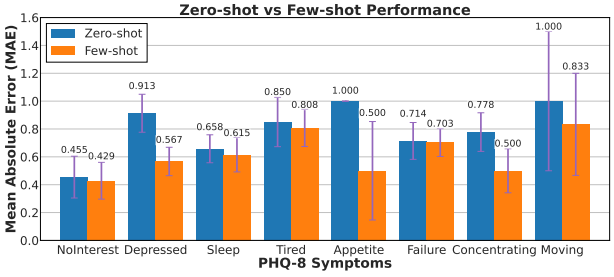

Figure 5: **Bar graph comparing zero-shot and few-shot performance on the test set.** Purple error bar shows standard error.

In addition to Gemma 3 27B, we also evaluated its variant fine-tuned on medical text, MedGemma 27B (Sellergren et al., 2025). MedGemma 27B achieved an improved average MAE of 0.505 but detected fewer chunks relevant, making fewer predictions overall (Appendix F).

3.3. Meta Review

The severity levels predicted by the LLM were largely consistent with the ground-truth labels (Figure 6a). However, 6 subjects with minimal symptoms were misclassified as having mild symptoms, which can imply the tendency of the model to overinterpret the presence of individual symptoms. We also identified 13 false negative cases (in the lower triangle), possibly due to incomplete coverage of symptoms during the interviews. Since not all eight PHQ-8 symptoms were always discussed, the model may have underestimated overall severity when provided with only partial information.

As presented in Figure 6 and Table 1, when participants were grouped by minimal and mild symptoms versus more severe symptoms, our agentic system achieved an accuracy of 78% in this binary classification task, with a precision of 0.727, recall of 0.571, and an F1 score of 0.640. In comparison, the human evaluator achieved an accuracy of 78%, with a precision of 1.000, recall of 0.357, and an F1 score of 0.526. The agentic system evaluation showed 80.5% agreement with the human evaluation, although the system tended to overestimate severity compared with the human evaluator. Our system performed comparable or slightly better than a human evaluator, despite incomplete information in the transcripts. The current evidence supports that the agentic system can detect the relevant symptom information with human-level accuracy.

3.4. Single-Prompt Experiment

To demonstrate the advantages of the multi-agent system, we conducted a single-prompt experiment by merging all assessment agents' prompts into one unified prompt. The single-prompt approach failed to generate severity scores for 18 of the 47 subjects in the DAIC-WOZ test set but achieved comparable performance for the remaining subjects. These results suggest that shorter, role-specific prompts are more robust and reliable, highlighting the necessity of the multi-agent framework (Appendix G).

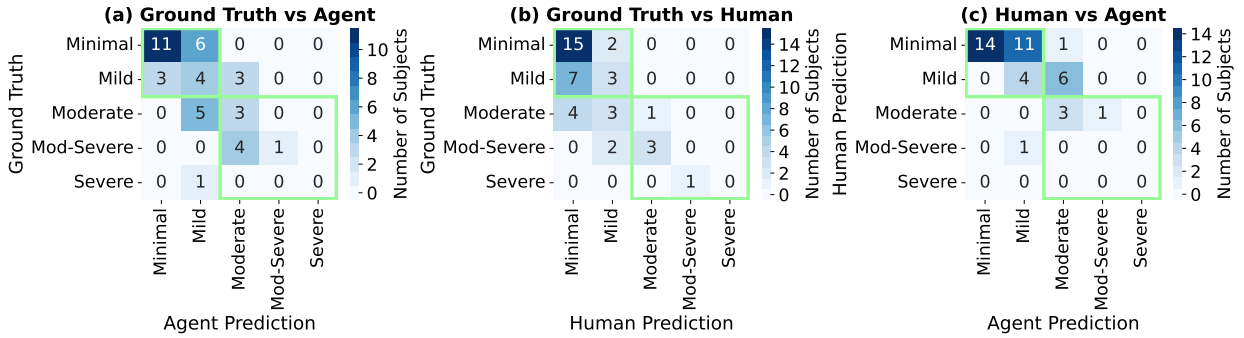

Figure 6: **Confusion matrices showing symptom severity prediction performance on the test set.** (a) Agent performance. (b) Human performance. (c) Human vs agent performance. Groupings (other depressive disorder vs MDD) are highlighted in green boxes.

Table 1: **Symptom severity prediction performance on the test set.**

Comparison	Accuracy	Balanced Accuracy	Precision	Recall	F1 Score
Ground Truth vs Agent	0.780	0.730	0.727	0.571	0.640
Ground Truth vs Human	0.780	0.679	1.000	0.357	0.526
Human vs Agent	0.805	0.803	0.364	0.800	0.500

4. Discussion

Through this work, we have analyzed how well large language models (LLMs) can diagnose depression through in-context learning and how they could be used in the psychiatry field as a whole. Our agentic workflow has provided useful insights and comments on a patient's mental health, which has shown promise for use in practical settings. Compared to similar works based on LLMs (Yang et al., 2024), we were able to achieve comparable results without the need for fine-tuning. This indicates that in-context learning can be a viable and efficient alternative for depression classification tasks.

In terms of deployment, the proposed system has modest computational overhead and short latency. Running the whole pipeline, including data loading, qualitative evaluation, quantitative scoring, and meta-review generation on a MacBook Pro with an Apple M3 Pro processor took approximately one minute to compile a report. This efficiency, achieved without the need of dedicated GPUs or specialized hardware, demonstrates the practicality and cost-effectiveness of the framework, allowing for rapid assessments on widely available consumer devices and supporting broader clinical and research use.

Our system aims to be a useful tool for psychiatrists and psychologists, providing support in their efforts to better understand their patients. To that end, we must also acknowledge the potential for ethical concerns arising from the use of LLMs in a sensitive area such as mental health. LLMs are inherently built to be probabilistic, which can be a strength in many aspects (Brown et al., 2020), but within this field, there's a precedent for misuse due to their inability to give genuine clinical judgment (Yang et al., 2024). There is a growing demand for simple, non-invasive solutions to mental health problems (Stoll et al., 2020), and even despite all the risks, some individuals try to use LLMs as a means of working through their problems. Applications such as Abby (`abby.gg`), Earkick (`earkick.com`), and Ash (`talktoash.com`) openly market themselves as AI therapists, yet the known limitations inherent to AI systems raise concerns regarding potential unintended user impact. Multiple states including Texas (Texas Attorney General Office, 2025), California (California Legislative Office, 2025), and Pennsyl-

vania (Pennsylvania General Assembly Office, 2025) have recently initiated preventive legislation. However, demand for AI therapy applications will likely continue to exist, and further investigations into ethical safeguards are key to minimize potential harm.

With that in mind, the proposed system is intended to assist, rather than replace, mental health professionals. It generates organized summaries of interview data aligned with standardized PHQ-8 definitions, flags ambiguous or high-risk cases for professional evaluation, provides explanations for its conclusions, and detects insufficient evidence without extrapolating (with potential to offer real-time suggestions to the interviewer). These outputs can support psychiatrists and psychologists by complementing manual note-taking, improving inter-rater agreement, and providing an auxiliary perspective during case consultations or primary-care screenings. Such assistive technologies are particularly valuable in resource-limited or high-throughput clinical settings. Additionally, the system's outputs may offer experts insights to inform further evaluation or decision-making.

The stochastic nature of LLMs renders a key limitation of the proposed approach. Even with fairly deterministic parameters, responses can vary across runs, making it challenging to obtain consistent performance metrics. In addition, the relatively small size of fully-labeled participants in the DAIC-WOZ database (142 subjects) limited the scope of our evaluation. While our quantitative analysis focused on predicting the eight-item PHQ-8 scores, several items were rarely discussed in interviews (see available scores in Figure 4), which limited the number of valid predictions and increased the variability in evaluation metrics. The upper bounds of the feedback loop evaluation (Figure 2) and the meta-review performance (Figure 6) likely reflect the same data limitation. Moreover, PHQ-8 scores are derived from the reported frequency of symptoms. Some participants did not explicitly describe the frequency or duration of symptoms (for example, "it depends" or "it is never easy"), making it difficult for the LLM to infer the correct score.

Beyond depression evaluation, the proposed framework is condition-agnostic and can, in principle, be extended to other mental disorders that use structured interviews and standardized diagnostic questionnaires. The modular design allows adaptation to new contexts by replacing the PHQ-8 reference framework with the relevant diagnostic constructs—such as the Generalized Anxiety Disorder-7 (GAD-7) for anxiety (Spitzer et al., 2006), the Young Mania Rating Scale (YMRS) for mania (Young et al., 1978), or the Positive and Negative Syndrome Scale (PANSS) for schizophrenia (Kay et al., 1987). Adapting the framework requires (1) assembling representative interview corpora for the target condition, (2) adjusting the assessment agents to align with the corresponding diagnostic criteria, and (3) incorporating expert-in-the-loop validation to ensure precision, interpretability, and fairness across subpopulations. Key challenges include domain-specific language variations, limited availability of high-quality labeled data for certain conditions, and heterogeneity in interview styles across sites. Addressing these issues will be essential for achieving the generalizability of the system in future clinical applications.

Future improvements could include expanding the transcript knowledge base with the extended DAIC dataset (Ringeval et al., 2019), evaluating the system on other mental disorders and assessment measures, and incorporating additional modalities, such as audio recordings or brain imaging.

5. Conclusion

This study demonstrates that LLMs can serve as useful tools for depression assessment. Our agentic system, while not an objective measure of truth, provides an interpretable view into a patient's mental state. Through iterative refinement and embedding-based few-shot prompting, our system is highly flexible, generalizable, and adaptable to various model sizes and different use cases. Practically, our system can provide additional viewpoints to practitioners and serve as an educational tool for junior psychiatrists. With fine-tuning, larger datasets, and more modalities, our system is likely to become an even more useful tool for psychiatrists to better understand their patients, thereby enhancing patient care in this age of AI.

Acknowledgments

The authors would like to thank the Math Path Program at Georgia State University funded by the Alfred P. Sloan Foundation G-2021-16960 and the NSF CREST D-MAP Program NSF2112455.

References

Falwah Alhamed, Julia Ive, and Lucia Specia. Classifying social media users before and after depression diagnosis via their language usage: A dataset and study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3250–3260, 2024.

Amna Amanat, Muhammad Rizwan, Abdul Rehman Javed, Maha Abdelhaq, Raed Alsaqour, Sharnil Pandya, and Mueen Uddin. Deep learning for depression detection from textual data. *Electronics*, 11(5):676, 2022.

Steen Andreassen, Annelise Rosenfalck, Bjørn Falck, Kristian G Olesen, and Stig Kjær Andersen. Evaluation of the diagnostic performance of the expert emg assistant munin. *Electroencephalography and Clinical Neurophysiology/Electromyography and Motor Control*, 101(2):129–144, 1996.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Sergio Burdisso, Ernesto Reyes-Ramírez, Esaú Villatoro-Tello, Fernando Sánchez-Vega, Pastor López-Monroy, and Petr Motlicek. Daic-woz: On the validity of using the therapist's prompts in automatic depression detection from clinical interviews. *arXiv preprint arXiv:2404.14463*, 2024.

California Legislative Office. Sb-579 mental health and artificial intelligence working group. Bill, 2025. URL https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB579.

Weize Chen et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in llm agents. *arXiv preprint arXiv:2308.10848*, 2023. URL https://arxiv.org/abs/2308.10848.

Zhuang Chen, Jiawen Deng, Jinfeng Zhou, Jincenzi Wu, Tieyun Qian, and Minlie Huang. Depression detection in clinical interviews with llm-empowered structural element graph. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8181–8194, 2024.

Jeffrey F. Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De la Torre. Detecting depression from facial actions and vocal prosody. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–7, 2009. doi: 10.1109/ACII.2009.5349358.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

Isaac R Galatzer-Levy, Daniel McDuff, Vivek Natarajan, Alan Karthikesalingam, and Matteo Malgaroli. The capability of large language models to measure psychiatric functioning. *arXiv preprint arXiv:2308.01834*, 2023.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, L. P. Morency, et al. The distress analysis interview corpus of human and computer interviews. In *LREC*, volume 14, pages 3123–3128, May 2014.

Lang He, Mingyue Niu, Prayag Tiwari, Pekka Marttinen, Rui Su, Jiewei Jiang, Chenguang Guo, Hongyu Wang, Songtao Ding, Zhongmin Wang, Xiaoying Pan, and Wei Dang. Deep learning for depression recognition with audiovisual cues: A review. *Information Fusion*, 80:56–86, 2022. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.

2021.10.012. URL https://www.sciencedirect.com/science/article/pii/S1566253521002207.

David Earl Heckerman, Eric J Horvitz, and Bharat N Nathwani. Toward normative expert systems: Part i the pathfinder project. *Methods of information in medicine*, 31(02):90–105, 1992.

Stanley R. Kay, Abraham Fiszbein, and Lewis A. Opler. The positive and negative syndrome scale (panss) for schizophrenia. *Schizophrenia Bulletin*, 13(2):261–276, 1987. doi: 10.1093/schbul/13.2.261. URL https://doi.org/10.1093/schbul/13.2.261.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173, 2009.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society. *arXiv preprint arXiv:2303.17760*, 2023. URL https://arxiv.org/abs/2303.17760.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, et al. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*, 2024.

Siwen Luo, Hamish Ivison, Soyeon Caren Han, and Josiah Poon. Local interpretations for explainable natural language processing: A survey. *ACM Computing Surveys*, 56(9):1–36, 2024.

Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang. Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC '16, page 35–42, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450345163. doi: 10.1145/2988257.2988267. URL https://doi.org/10.1145/2988257.2988267.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.

Gregory J Meyer, Stephen E Finn, Lorraine D Eyde, Gary G Kay, Kevin L Moreland, Robert R Dies, Elena J Eisman, Tom W Kubiszyn, and Geoffrey M Reed. Psychological testing and psychological assessment: A review of evidence and issues. *American psychologist*, 56(2):128, 2001.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022.

Pennsylvania General Assembly Office. Senate bill 631. Bill, 2025. URL https://www.palegis.us/legislation/bills/2025/sb631.

Iryna Pentina, Tyler Hancock, and Tianling Xie. Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior*, 140:107600, 2023. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2022.107600. URL https://www.sciencedirect.com/science/article/pii/S0747563222004204.

Jianing Qiu, Kyle Lam, Guohao Li, Amish Acharya, Tien Yin Wong, Ara Darzi, Wu Yuan, and Eric J Topol. Llm-based agentic systems in medicine and healthcare. *Nature Machine Intelligence*, 6(12):1418–1420, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, pages 3–12, 2019.

Nico Roos, Annette Ten Teije, André Bos, and Cees Witteveen. An analysis of multi-agent diagnosis.

In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2*, pages 986–987, 2002.

M. Sadeghi, R. Richer, B. Egger, L. Schindler-Gmelch, L. H. Rupp, F. Rahimi, B. M. Eskofier, et al. Harnessing multimodal approaches for depression detection using large language models and facial expressions. *npj Mental Health Research*, 3 (1):66, 2024.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551, 2023.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Caifeng Shan, Shaogang Gong, and Peter W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27 (6):803–816, 2009. ISSN 0262-8856. doi: https://doi.org/10.1016/j.imavis.2008.08.005. URL https://www.sciencedirect.com/science/article/pii/S0262885608001844.

Robert L Spitzer, Kurt Kroenke, Janet B W Williams, and Bernd Löwe. A brief measure for assessing generalized anxiety disorder: The gad-7. *Archives of Internal Medicine*, 166(10):1092–1097, 2006. doi: 10.1001/archinte.166.10.1092. URL https://doi.org/10.1001/archinte.166.10.1092.

Elizabeth C Stade, Shannon Wiltsey Stirman, Lyle H Ungar, Cody L Boland, H Andrew Schwartz, David B Yaden, João Sedoc, Robert J DeRubeis, Robb Willer, and Johannes C Eichstaedt. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Research*, 3(1):12, 2024.

Julia Stoll, Jonas Adrian Müller, and Manuel Trachsel. Ethical issues in online psychotherapy: A narrative review. *Frontiers in psychiatry*, 10:498439, 2020.

Waleed Bin Tahir, Shah Khalid, Sulaiman Almutairi, Mohammed Abohashrh, Sufyan Ali Memon, and Jawad Khan. Depression detection in social media: A comprehensive review of machine learning and deep learning techniques. *IEEE Access*, 2025.

Texas Attorney General Office. Attorney general ken paxton investigates meta and character.ai for misleading children with deceptive ai-generated mental health services. News Release, 2025.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Rosa M Vicari, Cecilia D Flores, André M Silvestre, Louise J Seixas, Marcelo Ladeira, and Helder Coelho. A multi-agent intelligent environment for medical knowledge. *Artificial Intelligence in Medicine*, 27(3):335–366, 2003. ISSN 0933-3657. doi: https://doi.org/10.1016/S0933-3657(03)00009-5. URL https://www.sciencedirect.com/science/article/pii/S0933365703000095. Software Agents in Health Care.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

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

Lingyun Wen, Xin Li, Guodong Guo, and Yu Zhu. Automated depression diagnosis based on facial dynamic analysis and sparse coding. *IEEE Transactions on Information Forensics and Security*, 10(7): 1432–1441, 2015. doi: 10.1109/TIFS.2015.2414392.

World Health Organization. Depression. https://www.who.int/news-room/fact-sheets/detail/depression, August 2025. Accessed: 2025-08-29.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun

Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023. URL https://arxiv.org/abs/2308.08155.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023. URL https://arxiv.org/abs/2309.07864.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024*, pages 4489–4500, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022. URL https://arxiv.org/abs/2210.03629.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Congchi Yin, Feng Li, Shu Zhang, Zike Wang, Jun Shao, Piji Li, Jianhua Chen, and Xun Jiang. Mdd-5k: a new diagnostic conversation dataset for mental disorders synthesized via neuro-symbolic llm agents. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i24.34763. URL https://doi.org/10.1609/aaai.v39i24.34763.

R. C. Young, J. T. Biggs, V. E. Ziegler, and D. A. Meyer. A rating scale for mania: reliability, validity and sensitivity. *The British Journal of Psychiatry*, 133:429–435, 1978. doi: 10.1192/bjp.133.5.429. URL https://doi.org/10.1192/bjp.133.5.429.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007. doi: 10.1109/TPAMI.2007.1110.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Appendix A. Related Work

The intersection between artificial intelligence and mental health assessment has evolved significantly over the past decade, with computational approaches to depression detection advancing from hand-crafted features to sophisticated computational architectures. This section examines four key areas of development in this field: the progression of deep learning methods, the emergence of large language models, the evolution of agentic systems, and the application of in-context learning techniques. Together, these approaches work to form the current landscape of AI-driven mental health assessment.

A.1. Deep Learning for Depression Assessment

Even before the rise of large language models, computational methods were used for mental health assessment. Cohn et al. (2009) demonstrated that by analyzing facial actions and vocal prosody in patients, depression could be detected accurately 79% of the time. Automatic Depression Detection (ADD) systems like Cohn's often relied on hand-crafted features such as Local Binary Patterns (LBP) (Shan et al., 2009), Local Phase Quantization from Three Orthogonal Planes (LPQ-TOP) (Wen et al., 2015), and Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) (Zhao and Pietikainen, 2007). Although these methods were reasonably effective, they required substantial effort, domain-specific knowledge, subjective assumptions for feature extraction, and they were unable to capture all relevant patterns from their data (He et al., 2022). More recent advances within ADD have allowed for a shift away from hand-crafted subjective assessment and towards more computational approaches. For example, the DepAudioNet model combined Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) to encode depression-related characteristics from audio data. Via these methods, DepAudioNet achieved an average F1 score of 0.61, which was a significant improvement over the baseline of 0.496 average F1 score (Ma et al., 2016). Deep learning prediction methods have continued to advance to such a point where near perfect accuracy has been achieved. More recently, one such method achieved an average F1 score of 0.98 by utilizing Long-Short Term Memory (LSTM) with a Recurrent Neural Network (RNN) (Amanat et al., 2022). Due to achieving such accuracy, the focus has shifted toward clarity and explainability (Tahir et al., 2025).

A.2. Large Language Models for Mental Health Assessment

Although deep learning models have proven to be very accurate, they lack the interpretability that would make their predictions practical. However, interpretability is where large language models (LLMs) shine. LLM based conversational agents, like Replika, are designed to act like a companion, with one of its main use-cases being aiding mental well being. During the COVID pandemic, Pentina et al. (2023) conducted a study asking active Replika users if Replika improved their mental well-being. An overwhelming majority agreed; many stated Replika acted as their support system, friend, or even romantic partner. Within the behavioral health field, the usage is not as broad, but LLMs are occasionally used by counselors to improve how they express empathy, or by therapists to identify client behaviors (Stade et al., 2024). Studies have experimented with the use of LLMs to directly predict depression, such as a study by (Alhamed et al., 2024) who found that traditional chatbots such as GPT-3 and Google Bard performed very poorly in depression classification, achieving F1 scores of 0.32 and 0.36 respectively. Some chat-based LLM's have been fine-tuned to work better for classification of depression. For example, MentaLLaMA, which used a large custom dataset with a non-agentic approach, achieved an F1 score of 0.71 across its testing categories (Yang et al., 2024).

A.3. Multi-Agent Systems for Medical Diagnosis and Assessment

Agentic systems for medical diagnosis and assessment have evolved substantially, progressing from single-agent Model-Based Diagnosis (MBD) to modern multi-agent systems powered by LLMs. Early diagnostic systems typically employed a single agent that modeled the entire system under analysis, often using MBD techniques (Roos et al., 2002). Subsequent advancements introduced probabilistic reasoning systems through Bayesian Networks, which provided a mathematically principled framework that aligned more closely with physicians' reasoning processes and yielded more clinically useful outputs (Vicari et al., 2003). Notable examples include Pathfinder for lymphatic disease diagnosis (Heckerman et al., 1992) and

MUNIN for muscle and nerve disorders (Andreassen et al., 1996).

Despite their utility, single-agent systems were limited in handling the complexity and variability of real-world medical scenarios. This led to the development of Multi-Agent Systems (MAS) and Distributed Artificial Intelligence (DAI) techniques (Roos et al., 2002). Recent LLM-based agentic systems represent a new generation of MAS, in which the LLM functions as a central reasoning module, employing strategies such as chain-of-thought reasoning and reflection to enhance performance. These systems typically comprise perception modules for multimodal input processing, memory modules for long-term information retention, and action modules for task execution (Qiu et al., 2024). A prominent example is Agent Hospital (Li et al., 2024), a comprehensive virtual environment that leverages MAS to assist clinicians in patient diagnosis. It employs Simulacrum-based Evolutionary Agent Learning (SEAL) to emulate specialized medical professionals, such as nurses and physicians. The system demonstrated high clinical accuracy, achieving 95.31% for diagnosis and 98.76% for medical examination selection.

A.4. In-Context Learning

In-context learning (ICL) has recently emerged as a significant strength of LLMs, enabling them to learn from examples provided in prompt context (Dong et al., 2022). Unlike traditional supervised learning, ICL does not require parameter updates, allowing for training-free knowledge incorporation (Brown et al., 2020). Techniques like Chain-of-Thought (CoT) prompting use ICL by having an LLM provide a step-by-step answer. This can result in significant improvements in tasks such as arithmetic and symbolic reasoning (Wei et al., 2022; Kojima et al., 2022). In the context of mental health, the incorporation of therapist questions into the LLMs prompt, instead of only patient responses, has been shown to enhance performance (Chen et al., 2024); however, these improvements may be due to an underlying bias within the therapist's questions (Burdisso et al., 2024).

Appendix B. Qualitative Evaluation Metrics

We defined four metrics to evaluate the qualitative assessment as follows:

- **Specificity**: Is the assessment specific? Mistakes include using vague or generic statements such as "the patient seems depressed".
- **Completeness**: Does the assessment cover all relevant symptoms, severities, duration/frequency? Mistakes are missed PHQ-8 symptoms, or duration/frequency details.
- **Coherence**: Is the response logically consistent? Mistakes are logically inconsistent statements or contradictions within the assessment.
- **Accuracy**: Are the signs/symptoms aligned with DSM-5 or PHQ-8? Mistakes are incorrect symptoms or incorrect duration/frequency.

Next, we mapped the number of mistakes to an integer value ranging from 1 to 5 for each metric:

- **Score of 5**: 0 mistake.
- **Score of 4**: 1 to 2 mistakes.
- **Score of 3**: 3 to 4 mistakes.
- **Score of 2**: 5 to 6 mistakes.
- **Score of 1**: 7 or more mistakes.

Appendix C. Data Splitting

We stratified 142 subjects from the DAIC-WOZ training and development sets into training, validation, and test sets based on PHQ-8 total scores and gender information. Due to the relatively small size and imbalanced distribution (only 22/142 participants in the 13+ total score range) of the dataset, we implemented automatic parsing for instances where there were two or one participants for a given PHQ-8 total score. For PHQ-8 total scores with two participants, we put one in the validation set and one in the test set. For PHQ-8 total scores with one participant, we put that one participant in the training set. This way, we prevented certain sets from missing patients within the higher PHQ-8 total score range, as we noted there were substantially less subjects with higher reported total scores.

Figure A1 shows the distributions of gender and PHQ-8 scores across the training, validation, and test sets. The distributions across the three sets were approximately balanced, ensuring representative analyses for information retrieval, hyperparameter optimization, and model evaluation.

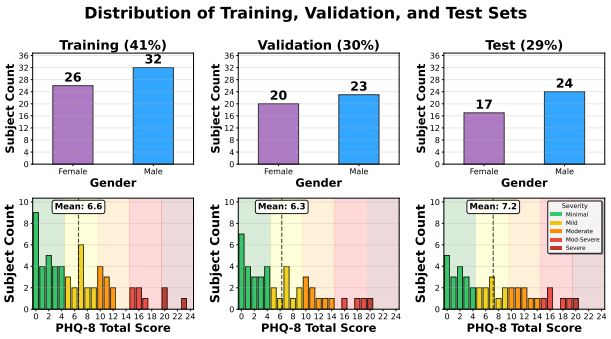

Figure A1: **Gender and PHQ-8 total score distributions across training, validation, and test sets.**

rate results. The optimal embedding dimension for Qwen 3 8B Embedding was 4096.

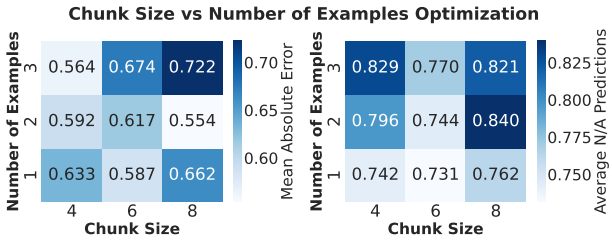

Figure A2: **Confusion matrices showing average hyperparameter performance over 3 runs on the validation set.**

Appendix D. Hyperparameter Optimization

We evaluated three hyperparameters: chunk size, number of reference examples, and embedding dimension. We built transcript chunks on a line-by-line basis, testing chunk sizes of 4, 6, and 8. We also determined how many reference examples would be pulled per question for use in the quantitative analysis prompt, including example numbers of 1, 2, and 3. Lastly, we evaluated embedding dimensions of 64, 256, 1024, and 4096.

We repeated the experiment three times for each of the nine chunk size and reference example combinations and averaged the results to mitigate variability. The two combinations with the lowest MAEs were then selected for evaluation on the test set. For each optimal combination, we performed one experiment with each of four embedding dimension options.

As shown in Figure A2, the combination of a chunk size of 8 and 2 reference examples yielded the lowest average MAE (0.554) and was therefore selected as the optimal hyperparameter setting. A chunk size of 4 with 3 reference examples also demonstrated competitive performance (MAE: 0.564). However, the lowest average number of N/A predictions was achieved with a chunk size of 6 and 1 reference example. Based on these results, we used a chunk size of 8 with 2 reference examples for all subsequent experiments.

Figure A3 shows a more straightforward pattern, with higher dimensions generally yielding more accu-

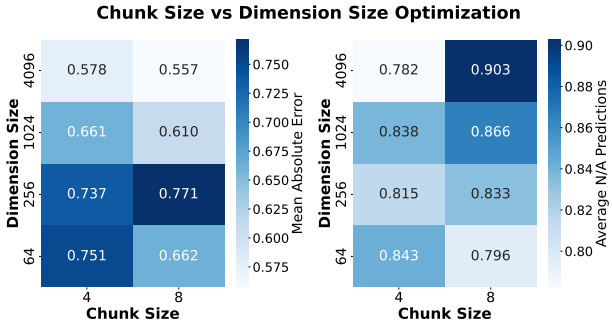

Figure A3: **Confusion matrices showing embedding dimension performance on the validation set.**

Appendix E. Retrieval Statistics with Gemma 3 27B

In the t-SNE graph, shown in Figure A4, the medians of the group without significant depressive symptoms and the group with mild symptoms remained close, whereas the medians of the groups with more severe symptoms were more dispersed. This may reflect the inherent heterogeneity of depression and variations in how individuals describe their experiences. For instance, one participant may experience depression following the loss of a family member, while another may experience it after losing a job. Both cases were labeled as severe depression, yet their textual embeddings may be far apart in the latent space.

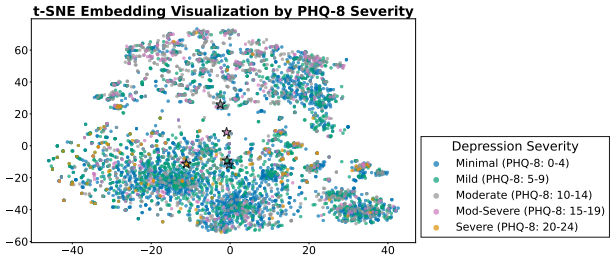

Figure A4: **t-SNE projection of training transcript embeddings, colored by severity.** The star shows the median of each severity group. $N_{\text{chunk}} = 8$.

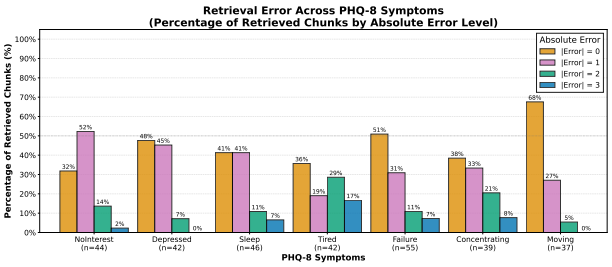

Figure A5: **Histogram of training transcript embeddings, colored by absolute error.** The symptom of poor appetite is missing because it was not discussed in the interviews.

Meanwhile, each bar on the histogram shown in Figure A5 represents the proportion of retrieved embeddings yielding a certain PHQ-8 symptom error level relative to the true score of the subject being evaluated. These absolute error values compared a ground truth PHQ-8 symptom score from the transcript being analyzed to the one from the transcript a given chunk came from (according to the PHQ-8 symptom). For example, if a chunk retrieved for PHQ-8-Sleep comes from a participant with a PHQ-8-Sleep score of 2 and the transcript being analyzed has a ground truth PHQ-8-Sleep score of 1, then the absolute error would be $|\text{Error}| = |2 - 1| = 1$.

From these histograms, we observed that for 6 out of the 7 symptoms with relevant chunks retrieved, most chunks had an absolute error of 0 (depressed: 48%; sleep: 41%; tired: 36%; failure: 51%; concentrating: 38%; moving: 68%)—indicating strong PHQ-8 score agreement between the retrieved and target chunks in subjects for which predictions were made.

Additionally, we noted that PHQ-8–Appetite had no successfully retrieved reference chunks during inference. Upon closer inspection, Gemma 3 27B did not identify any evidence related to appetite issues in the available transcripts, resulting in no reference retrieval for that symptom.

Appendix F. MedGemma Results

We evaluated MedGemma 27B with the same optimal hyperparameters determined for Gemma 3 27B and continued to use Qwen 3 8B Embedding for few-shot embeddings. As shown in Figure A6, MedGemma 27B had an edge over Gemma 3 27B in most categories overall, achieving an average MAE of 0.505, 18% less than Gemma 3 27B, although the number of subjects detected as having available evidence from the transcripts was smaller with MedGemma. Figure A7 further shows MedGemma 27B performance on subjects with detected evidence exceeds Gemma 3 27B performance in most individual categories. Note that for the same transcript, different models may or may not detect evidence, leading to different number of available scores, i.e., number of subjects for which the model was able to make a prediction.

Appendix G. Single-Prompt Experiment

We merged the prompts of the three agents (qualitative assessment, quantitative assessment, and meta-review agents) into a single prompt, and compared the single-prompt approach with our multi-agent system. With the single-prompt approach, the LLM struggled to follow instructions in the longer context—failing to output severity scores for 18 of the 47 subjects in the DAIC-WOZ test set. Among the remaining subjects with valid predictions, the single-prompt approach achieved comparable results to the multi-agent system (accuracy: 0.793, precision: 0.625, recall: 0.625, F1 score: 0.625). However, 38.3% of subjects lacked valid predictions, limiting its applicability in real-world settings. These findings suggest that shorter, specialized prompts assigned to specific agents are more robust and reliable, underscoring the necessity of the multi-agent framework.

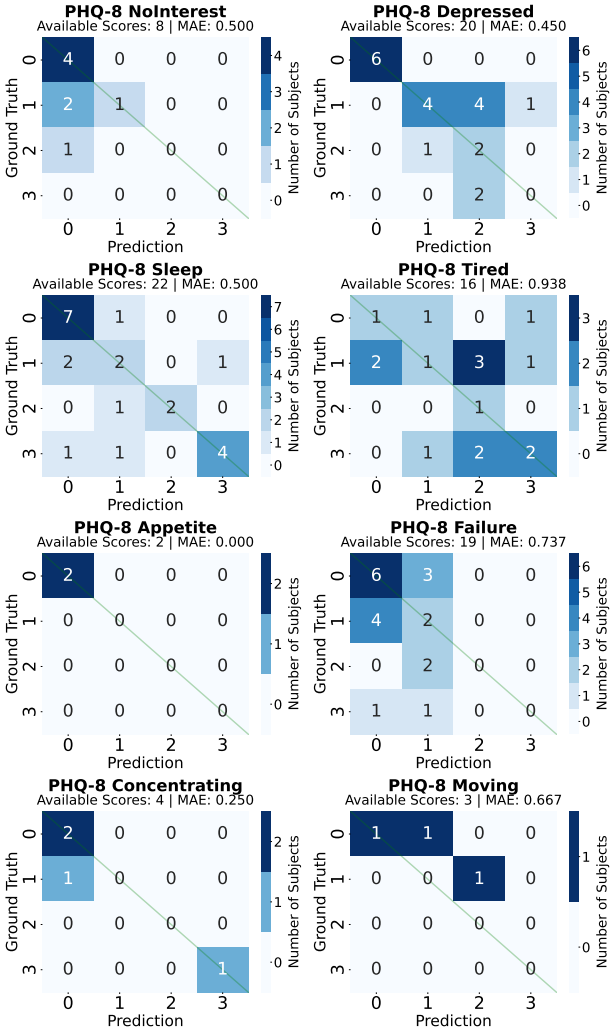

Figure A6: **MedGemma 27B confusion matrices showing PHQ-8 prediction performance with optimal hyperparameters on the test set.**

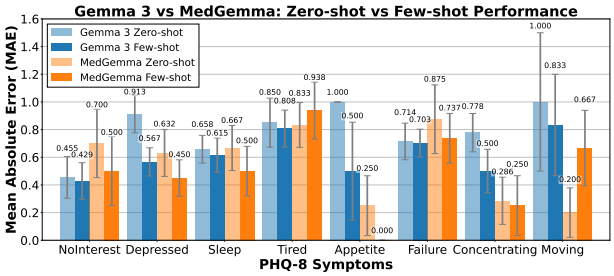

Figure A7: **MedGemma 27B bar graph comparing zero-shot and few-shot performance on the test set.** Purple error bar shows standard error.