MEMEMINDS: RAG-Enhanced System for Entertainment-Based Psychological Diagnosis

Minzhou Pan
Student, MSECE

Jianqing Ma
Student, MSCS

Jiarong Zhu
Student, MSCS

Renlei Huang
Student, MSCS

Yujia Zhou
Student, MSECE

Qihui Fan
Student, MSDS

Abstract

Large language models (LLMs) have shown remarkable success across various natural language processing tasks, yet applying them to complex domains like psychological diagnosis remains a significant challenge. In this paper, we present MEMEMINDS, a novel framework that leverages retrieval-augmented generation (RAG) and an ensemble voting mechanism to enable LLMs to deliver more reliable, contextually rich psychological assessments. MEMEMINDS first integrates external knowledge sources to enhance domain specificity, then employs multiple state-of-the-art LLMs whose individual judgments are consolidated through a voting process. This collaborative approach refines diagnostic reasoning, reduces model biases, and fosters a higher level of interpretability.

We evaluate MEMEMINDS on both synthetic and real-world patient case descriptions, demonstrating an 82.4% accuracy in identifying psychological patterns—approaching near-expert performance. These results underscore the potential of retrieval-enhanced, ensemble-driven LLMs to improve the quality, accessibility, and efficiency of text-based mental health support. By blending scalable automation with informed, evidence-based reasoning, MEMEMINDS marks a significant step toward more nuanced and trustworthy computational psychology tools.

1 Introduction

The emergence of large language models (LLMs) has propelled artificial intelligence (AI) into new frontiers, producing remarkable success across a range of natural language processing tasks (Brown et al., 2020; Chowdhery et al., 2022). Yet, applying LLMs to specialized domains—particularly psychological diagnosis—remains non-trivial. The complexity of human cognition, the diversity of symptom presentations, and the subtlety required for clinically meaningful insights pose significant hurdles to direct, off-the-shelf LLM deployment (Abbe and Brandon, 2014; Luo et al., 2019). In short, effective text-based psychological assessment demands a depth of domain expertise and contextual understanding that current models, trained primarily on general-purpose data, struggle to achieve.

Although recent work has explored the use of LLMs for mental health applications, such as screening for depressive symptoms or identifying risk factors in written narratives, these approaches often rely on shallow keyword matching or heuristic rules, failing to fully integrate external clinical knowledge (Chekroud and Koutsouleris, 2018; Abbe and Brandon, 2014). Without domain-specific context, LLMs risk producing overly generic, unreliable, or even misleading assessments. To address these limitations, we introduce MEMEMINDS, a novel system that augments LLM reasoning with retrieval-augmented generation (RAG) (Lewis et al., 2020) and an ensemble voting mechanism inspired by electoral fusion methods (Zhao et al., 2024). By drawing on authoritative knowledge sources—such as curated symptom summaries and diagnostic guidelines—MEMEMINDS mitigates hallucinations and enriches each model's internal reasoning. In addition, the ensemble voting framework leverages multiple state-of-the-art LLMs to reach a consensus, ensuring that diagnostic conclusions are supported by a diverse set of expert-like perspectives (Nguyen et al., 2022).

The motivation for MEMEMINDS is rooted in the urgent need for accessible, high-quality mental health support. While mental health disorders affect a significant portion of the global population, traditional care models are often limited by cost, stigma, and a scarcity of qualified professionals (Organization et al., 2019). Conventional

assessment methods—such as clinical interviews and self-report inventories—are time-consuming, resource-intensive, and susceptible to response biases (Tarescavage and Ben-Porath, 2014; Paulhus and Vazire, 2007). These challenges hinder timely access to diagnostic feedback and subsequent interventions, leaving many individuals undiagnosed or untreated. By combining the scalability and real-time adaptability of LLMs with retrieval-augmented knowledge and ensemble-driven decision-making, MEMEMINDS aims to streamline the diagnostic process. This, in turn, promises more prompt, reliable, and confidential mental health support to a broader population, especially those previously underserved by traditional systems.

In this paper, we present the design, implementation, and thorough evaluation of MEMEMINDS. Our results demonstrate that it can accurately identify psychological patterns across diverse datasets, achieving near-expert performance in text-based diagnoses. These findings illuminate the transformative potential of retrieval-enhanced, ensemble-driven LLM frameworks for computational psychology, reducing barriers to mental health care, and setting the stage for more human-centered and evidence-backed AI-driven assessments.

2 Background & Related Work

LLMs in Mental Health Diagnosis. The integration of large language models (LLMs) into mental health care offers a promising pathway toward more accessible and timely support, yet significant barriers remain. Traditional clinical services often suffer from high costs, limited availability of trained practitioners, and logistical challenges, making it difficult for many individuals to obtain professional help (Coombs et al., 2021). Recent natural language processing (NLP) breakthroughs have improved our capacity to analyze human affective states through multiple modalities—text, speech, and facial expressions—thereby paving the way for more nuanced understandings of emotional and psychological well-being (Dalvi et al., 2021; De Boer et al., 2018). Building on these advances, LLMs have demonstrated potential to facilitate natural, context-sensitive interactions that can support a wide range of mental health assessments and interventions (Lawrence et al., 2024). Nonetheless, current systems often struggle to deliver responses that are both empathetic and psychologically insightful, underscoring the need for models that transcend superficial affect detection and instead capture the subtle complexities of human cognition and emotion (Khare et al., 2024).

Advances in Psychological Diagnosis with AI. A growing body of research highlights the value of AI-driven tools in improving psychological evaluation and intervention. For example, Ji et al. (2022) introduced MentalBERT, a domain-adapted model fine-tuned on mental health-related corpora that demonstrated promising capabilities in identifying psychological conditions from social media content. Similarly, Zhang et al. (2023a) developed a multi-modal framework that integrates text, speech, and facial cues to offer more holistic mental health assessments. These pioneering efforts underscore the potential of AI systems to augment clinical practice, reduce resource bottlenecks, and expand access to care. However, the path forward is not without challenges. As these models move beyond proof-of-concept demonstrations, issues of explainability, interpretability, and ethical considerations related to privacy and bias become paramount (Saeidnia et al., 2024). Close collaboration between AI researchers, clinicians, and ethicists is vital to ensure that these tools align with clinical validity, cultural sensitivity, and equitable standards of care.

Retrieval Augmented Generation. While LLMs have achieved remarkable success on general tasks, their application in specialized domains often exposes knowledge gaps and a tendency to produce factually inaccurate outputs or "hallucinations," especially when dealing with context beyond their training data (Kandpal et al., 2023; Zhang et al., 2023b). Retrieval-Augmented Generation (RAG) offers a promising remedy by incorporating external knowledge bases to supply relevant, up-to-date information (Lewis et al., 2020; Gao et al., 2023). RAG frameworks leverage semantic similarity to locate and integrate the most pertinent document fragments into the generation process, thereby enhancing both accuracy and domain specificity. Still, RAG's effectiveness hinges on the quality of its retrieval component. Poorly chosen documents can degrade model reliability, reinforcing the importance of precise and efficient retrieval methods (Chen et al., 2024). To address these issues, tools like **FAISS** (Facebook AI Similarity Search) provide a scalable solution for fast, accurate similarity searches over large vector databases (Meta,

2024). By leveraging techniques from vector indexing and clustering, FAISS refines the retrieval pipeline, improving the relevance of fetched knowledge chunks and minimizing the risk of incorporating non-pertinent information (Ghadekar et al., 2023). In turn, these advancements in retrieval technology hold the potential to further strengthen LLM-based diagnostic systems, ensuring that complex clinical assessments are grounded in solid, context-aware evidence.

3 Methodology

Our proposed system, MEMEMINDS, integrates Retrieval-Augmented Generation (RAG) and multi-agent decision-making to deliver contextually grounded, clinically relevant psychological assessments. We first retrieve a corpus of potentially relevant documents using a hybrid retrieval approach, then generate follow-up queries that refine the diagnostic context. Subsequently, multiple LLM agents collaborate through a two-stage voting process to identify the most likely diagnosis. Figure 2 provides an overview of the entire pipeline.

3.1 Problem Setup and Notation

Let q represent the user's initial query, which may include symptoms, feelings, or concerns related to their mental health. Our objective is to identify a set of candidate diagnoses $\mathcal{D} = d_1, d_2, \ldots, d_m$ that best reflect the user's condition. We leverage a local knowledge repository $\mathcal{R}$, composed of textual summaries from reputable sources like Wikipedia, to inform our reasoning. Each document $r \in \mathcal{R}$ corresponds to a description of a mental or neurological condition.

To overcome knowledge gaps and subtle distinctions between closely related disorders, we employ a hybrid retrieval framework. This framework balances keyword-driven precision with semantic depth, ensuring that the retrieved information is both contextually rich and lexically relevant.

3.2 Retrieval-Augmented Generation

3.2.1 Knowledge Representation

We store the repository $\mathcal{R}$ in a NoSQL database (MongoDB) for flexible and efficient querying. Each document $r \in \mathcal{R}$ is mapped into a d-dimensional embedding vector $\mathbf{e}r \in \mathbb{R}^d$ using a high-quality encoding function $f\text{enc}(\cdot)$. For a given user query q, its embedding is:

$$\mathbf{e}q = f\text{enc}(q). \tag{1}$$

These embeddings capture semantic structure, enabling retrieval of documents that are not only topically relevant but also contextually aligned with the user's underlying concerns.

3.2.2 Hybrid Retrieval

To ensure robust retrieval, we fuse two complementary methods: keyword-based BM25 (Robertson et al., 2009) and embedding-based FAISS (Ghadekar et al., 2023).

BM25 Retrieval. BM25 computes relevance scores based on term frequency and inverse document frequency. For query q and document r, the BM25 score is defined as:

$$\text{score}_{\text{BM25}}(q, r) = \sum_{w \in q} \text{IDF}(w) \cdot \frac{(k+1)\text{TF}(w, r)}{\text{TF}(w, r) + k\left(1 - b + b \cdot \frac{|r|}{|\bar{r}|}\right)}, \tag{2}$$

where k and b are tuning parameters, $|r|$ is the length of the document r, and $|\bar{r}|$ is the average document length.

FAISS Retrieval. FAISS leverages dense vector embeddings to capture semantic similarities. We compute cosine similarity:

$$\text{score}_{\text{FAISS}}(q, r) = \frac{\mathbf{e}_q \cdot \mathbf{e}_r}{|\mathbf{e}_q||\mathbf{e}_r|}. \tag{3}$$

We retrieve the top-K_{BM25} documents using BM25 and the top-K_{FAISS} documents via FAISS. The combined candidate set is:

$$\mathcal{R}_q = \text{Top}_{K_{\text{BM25}}}(\text{BM25}(q, \mathcal{R})) \cup \text{Top}_{K_{\text{FAISS}}}(\text{FAISS}(q, \mathcal{R})). \tag{4}$$

This hybrid approach ensures that the final retrieved set $\mathcal{R}_q$ balances lexical precision with deeper conceptual relevance, enhancing the knowledge available to subsequent modules.

3.3 Follow-up Query Generation

The user's initial query q may lack detail about symptom duration, severity, or environmental context, hindering differential diagnosis. To refine the diagnostic clarity, we introduce a follow-up query q' generated by a Large Language Model (LLM) specifically tuned for clinical inquiry:

$$q' = F_{\text{QG}}(q, \mathcal{R}q), \tag{5}$$

where $F\text{QG}(\cdot)$ is a function implemented by a state-of-the-art LLM (e.g., GPT-4). Guided by standardized diagnostic criteria (e.g., DSM-5), q' focuses on eliciting key clinical indicators, enabling more accurate and context-specific reasoning in downstream stages.

3.4 Multi-Agent Diagnostic Inference

3.4.1 Scoring Diagnoses

After gathering the user's initial query q, the retrieved documents $\mathcal{R}_q$, the follow-up query q', and the user's response to q', we present this enriched information to three specialized LLM agents: A_1, A_2, and A_3. Each agent is responsible for evaluating a set of candidate diagnoses $\mathcal{D}$, assigning a likelihood score:

$$s_i(d_j) \in 1, 2, 3, 4, 5, \quad i \in 1, 2, 3, ; d_j \in \mathcal{D}. \quad (6)$$

A score of 1 indicates minimal likelihood, while a score of 5 suggests a strong likelihood of the condition. Agents may also propose new diagnoses d_{new} not originally included in $\mathcal{D}$ if the retrieved context or user responses justify exploring additional conditions.

We aggregate the agent scores to measure overall consensus:

$$S(d_j) = \sum_{i=1}^{3} s_i(d_j). \quad (7)$$

We then filter these diagnoses, retaining those that surpass a threshold $\theta = 5$:

$$\mathcal{D}' = d_j \mid S(d_j) > \theta. \quad (8)$$

This filtering step narrows attention to a smaller, more promising subset of candidate diagnoses that resonate strongly among multiple experts.

3.4.2 Final Voting

In the second voting round, the same three agents reassess only the reduced subset $\mathcal{D}'$, taking into account the full conversational and retrieved context. Following their reassessment, we select the final diagnosis $d^\star$:

$$d^\star = \arg\max_{d_j \in \mathcal{D}'} S(d_j). \quad (9)$$

We also identify the agent $A_\star$ that most strongly supported $d^\star$ in this final evaluation.

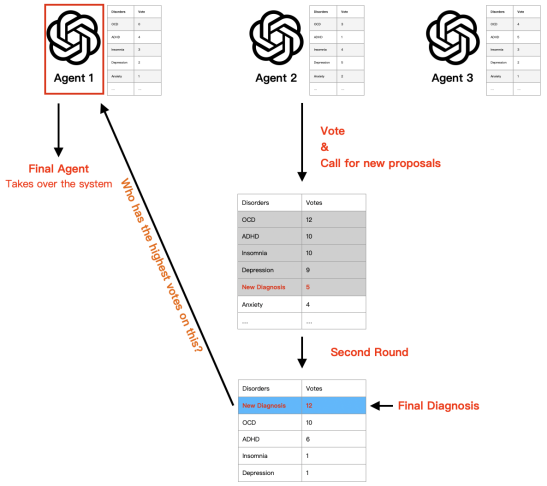

Figure 1: Schematic of the proposed two-stage voting mechanism. In the first stage, three agents independently score each candidate diagnosis based on retrieved context and user responses, filtering out low-scoring options. In the second stage, agents reevaluate the refined candidate set, culminating in the selection of a final, consensus-backed diagnosis. The agent most invested in the chosen diagnosis then becomes the primary conversational partner, ensuring consistent, informed dialogue moving forward.

3.5 Final Response Generation

Once the final diagnosis $d^\star$ is determined, the agent $A_\star$—which demonstrated the strongest support for $d^\star$—is tasked with generating the final user-facing response:

$$r^\star = F_{\text{Resp}}(q, q', \mathcal{R}q, d^\star), \quad (10)$$

where $F\text{Resp}(\cdot)$ is the response generation function implemented by the selected agent.

The final response provides a concise, clinically grounded rationale for the chosen diagnosis, along with actionable suggestions or referrals. The user may continue interacting with $A_\star$ for further clarification, support, or detailed reasoning. This continuity ensures that the user experiences a coherent and context-aware conversation, ultimately enhancing trust and engagement.

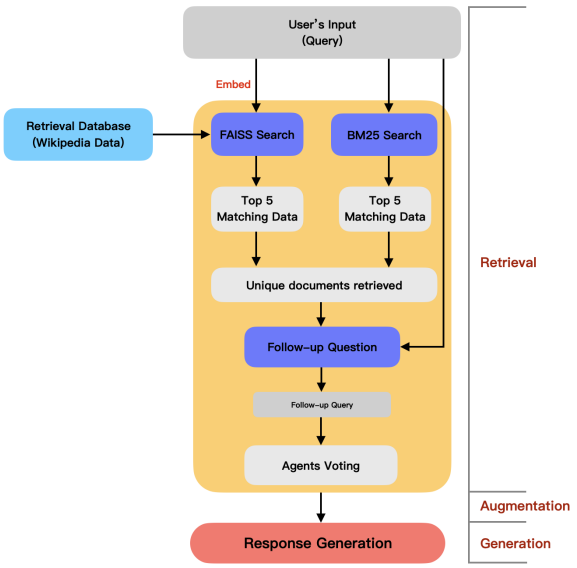

Figure 2: Overview of the MEMEMINDS pipeline. The system begins with the user's initial query, performing hybrid retrieval (BM25 + FAISS) to build a context-rich reference set. A follow-up query q' is generated to clarify critical clinical aspects. Multiple LLM agents then collaboratively score candidate diagnoses in two voting rounds, narrowing down to a final, consensus-backed diagnosis. The chosen agent provides the final response, maintaining continuity and clinical relevance in the ongoing user interaction.

4 Evaluation

In this section, we present a comprehensive evaluation of MEMEMINDS, assessing its diagnostic accuracy, agent collaboration, and robustness across both real-world and AI-generated test cases. We first outline the evaluation settings, followed by an analysis of system-level performance, agent collaboration, and differences between real-world and synthetic datasets.

4.1 Evaluation Settings

MEMEMINDS was evaluated on a benchmark of 500 psychological diagnostic cases. This benchmark comprised two components: (1) real-world inspired cases derived from summaries of 74 mental disorders listed in the DSM-5 and related datasets (e.g., MDD-5k), and (2) 300 AI-generated patient self-descriptions created using large language models to simulate realistic yet synthetic psychological profiles. These combined datasets offered a diverse range of symptoms, severity levels, and contextual details, providing a challenging testbed for our system.

Throughout the evaluation, three LLM agents (`gpt-4o`, `gpt-3.5`, `gpt-4`) operated within the RAG framework. We compared MEMEMINDS's performance against baseline LLM configurations, including those without external knowledge bases or voting mechanisms. We also evaluated a simplified `gpt-4o` model with no external knowledge and no fine-tuning to highlight the benefits of RAG and multi-agent voting.

4.2 Overall Accuracy

Model / Setting	Hit Rate
GPT4o (no KB/voting)	63.6%
GPT4o* (with KB)	71.3%
MEMEMINDS (full system)	82.4%
DSM-5 (human-level)	98%

* denotes GPT4o with access to a knowledge base

Table 1: Hit Rate of Different Systems. MEMEMINDS significantly outperforms baseline models and approaches near-expert performance.

MEMEMINDS achieved an overall accuracy of 82.4% (412 correct out of 500 cases) as shown in Table 1. This result outperformed both the `gpt-4o` baseline without external resources (63.6%) and the same model enhanced with a knowledge base but no voting mechanism (71.3%). Although there remains a gap compared to expert-level DSM-5 evaluation (98%), MEMEMINDS's significant improvement underscores the effectiveness of combining retrieval augmentation with multi-agent consensus.

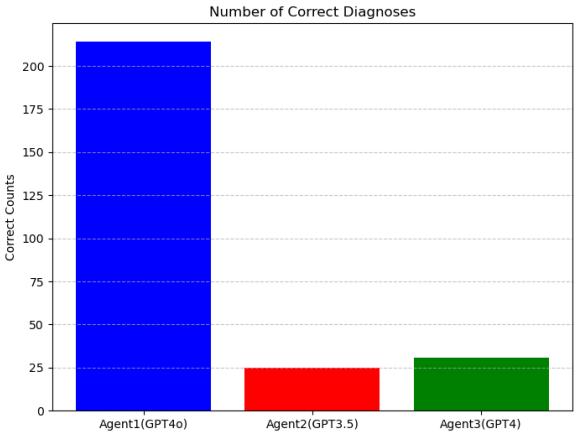

Figure 3: Number of Correct Cases per Agent. Although `gpt-4o` delivered the majority of correct outputs, the collaborative voting mechanism allowed even weaker agents (`gpt-3.5` and `gpt-4`) to contribute meaningfully, improving the overall diagnostic accuracy.

4.3 Agent Collaboration and Performance

Within the multi-agent framework, each agent ranked candidate diagnoses and contributed their unique perspective. Figure 3 shows that `gpt-4o` provided 366 correct diagnoses, substantially outperforming the other two agents (`gpt-3.5` and `gpt-4`, which contributed 39 and 45 correct diagnoses, respectively). Despite these performance differences, the collaborative voting mechanism ensured that all agents could influence the final outcome. Even less accurate agents played a valuable role by occasionally introducing new insights or confirming less common diagnoses.

Importantly, approximately 6.8% of cases underwent a form of "self-correction" during the final voting stage. In these instances, the initially favored diagnosis (A) was ultimately discarded in favor of an alternative diagnosis (B) after reconsideration. When B aligned with the ground truth, it exemplified the agents' ability to refine their reasoning collectively, reaching a more accurate conclusion than any single agent could have produced alone.

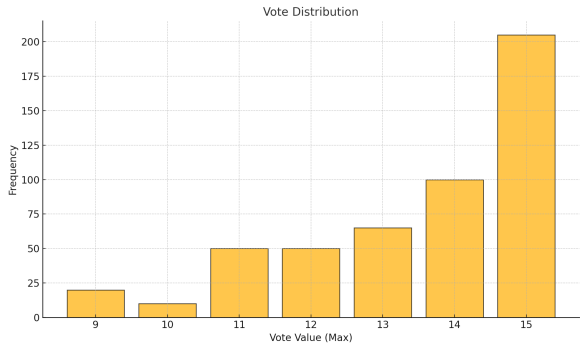

Figure 4: Distribution of Final Votes. The majority of final votes are concentrated near the high end of the scoring scale, indicating strong consensus among the agents. The average final vote is 13.5 out of 15, with a minimum of 9 and a maximum of 15, suggesting minimal divergence across agents.

Figure 4 illustrates the distribution of final vote totals across diagnoses. With an average of 13.5 and a narrow range centered around the upper end of the scoring scale (9–15), the agents consistently converged on strong recommendations. This tight clustering indicates that the voting mechanism effectively harnessed the agents' strengths, reducing uncertainty and increasing diagnostic confidence.

4.4 Performance on AI-Generated Cases

The system demonstrated particularly strong performance on the 300 AI-generated cases, correctly diagnosing 280 (93.3%). In these scenarios, the patterns were often clearer and more aligned with training data distributions, allowing the agents to reach unanimous decisions (15 out of 15 possible votes) with high confidence. Among the 20 incorrect diagnoses, 13 included the correct option as a candidate but failed during the final selection stage, suggesting that while clarity is generally higher in synthetic cases, subtle decision-making errors can still occur.

In comparison, performance on real-world-inspired data was lower, reflecting the inherent complexity, ambiguity, and variability of genuine patient descriptions. This discrepancy underscores the importance of continual refinement, domain adaptation, and real-world validation to ensure robust generalization beyond synthetic benchmarks.

4.5 Insights and Future Directions

The results highlight the value of integrating retrieval-augmented knowledge and multi-agent voting for complex diagnostic tasks. While MEMEMINDS has yet to reach human-expert performance, its substantial accuracy gains compared to baseline methods underscore the potential of such an ensemble approach. Future work may involve refining the prompt engineering strategies, incorporating more specialized clinical models, and expanding the knowledge bases to improve the fidelity and interpretability of diagnoses—especially for complex, real-world presentations.

5 Limitations and Future Works

MEMEMINDS faces several limitations that constrain its clinical applicability and generalizability. First, the scarcity of real-world diagnostic data—due to privacy restrictions—forces reliance on publicly available or AI-generated datasets (e.g., MDD-5k), which may lack the complexity, diversity, and cultural nuance found in genuine patient cases. This limitation not only reduces the system's realism but may also inflate performance on synthetic scenarios. Second, the absence of substantial input from licensed mental health professionals limits the clinical validity and interpretability of the system's reasoning. While relying on DSM-5 and related resources provides a structured foundation, direct expert collaboration could yield more accu-

rate and context-sensitive insights. In addition, the system's multi-agent architecture, although effective, reveals imbalances in agent performance (e.g., `GPT-4o` dominating the decision-making), potentially affecting diagnostic consistency in edge cases. Diagnostic granularity also remains a challenge: the system frequently defaults to high-level classifications rather than more precise subtypes. Moreover, its psychological assessments are intended for self-awareness or entertainment rather than clinical use, raising potential ethical and privacy concerns if users misconstrue outputs as professional advice. Looking ahead, future efforts should enhance data coverage and complexity by integrating more diverse and culturally sensitive datasets, as well as collaborating on the creation of dedicated psychological corpora for text-based diagnostics. Strengthening ties with clinical professionals and domain experts can guide refinement of the system's reasoning processes, improve interpretability, and ensure alignment with established best practices. Further experimentation with specialized models and model ensembles trained specifically for mental health contexts may also boost accuracy and adaptability. Fine-tuning on real-world counseling transcripts and employing machine learning techniques (e.g., supervised or reinforcement learning) to learn optimal vote weights can further improve system performance. Ultimately, deeper cross-disciplinary engagement—from psychology to linguistics and beyond—will be essential for refining MEMEMINDS into a more clinically robust and contextually grounded tool.

6 Conclusion

This project demonstrates the potential of implementing a RAG-enhanced GPT model for informal, entertainment-focused psychological insights. By integrating similarity search methods like FAISS and BM25 with state-of-the-art LLMs, the system provides an engaging platform for exploring psychological patterns and fostering self-awareness. The two-round voting mechanism and multi-agent collaboration significantly enhance diagnostic accuracy and reasoning, while the integration of Wikipedia content and DSM-5 guidelines ensures contextual relevance.

Although not intended for clinical use, the system showcases promising benchmarks, with an accuracy of 82.4% in simulated scenarios. This highlights its potential as a foundation for future advancements in text-based psychological diagnostics. Future efforts should focus on addressing limitations, such as dataset diversity and expert collaboration, to refine the system further and explore its applicability in real-world psychological assessment contexts.

References

Allison Abbe and Susan E Brandon. 2014. Building and maintaining rapport in investigative interviews. *Police practice and research*, 15(3):207–220.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Adam Mourad Chekroud and Nikolaos Koutsouleris. 2018. The perilous path from publication to practice. *Molecular psychiatry*, 23(1):24–25.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Nicholas C Coombs, Wyatt E Meriwether, James Caringi, and Sophia R Newcomer. 2021. Barriers to healthcare access among us adults with mental health challenges: A population-based study. *SSM-population health*, 15:100847.

Chirag Dalvi, Manish Rathod, Shruti Patil, Shilpa Gite, and Ketan Kotecha. 2021. A survey of ai-based facial emotion recognition: Features, ml & dl techniques, age-wise datasets and future directions. *Ieee Access*, 9:165806–165840.

JN De Boer, AE Voppel, MJH Begemann, HG Schnack, F Wijnen, and IEC Sommer. 2018. Clinical use of semantic space models in psychiatry and neurology: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, 93:85–92.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Preman Ghadekar, Sahil Mohite, Omkar More, Praiwal Patil, Shubham Mangrule, et al. 2023. Sentence meaning similarity detector using faiss. In *2023 7th*

International Conference On Computing, Communication, Control And Automation (ICCUBEA), pages 1–6. IEEE.

Qiu Huachuan. 2024. Smile: Sentiment and emotion analysis datasets [github repository]. https://github.com/qiuhuachuan/smile.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Smith K Khare, Victoria Blanes-Vidal, Esmaeil S Nadimi, and U Rajendra Acharya. 2024. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information Fusion*, 102:102019.

Hannah R Lawrence, Renee A Schneider, Susan B Rubin, Maja J Matarić, Daniel J McDuff, and Megan Jones Bell. 2024. The opportunities and risks of large language models in mental health. *JMIR Mental Health*, 11(1):e59479.

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

Minxia Luo, Gerold Schneider, Mike Martin, and Burcu Demiray. 2019. Cognitive aging effects on language use in real-life contexts: A naturalistic observation study. *PloS one*, 14(10):e0224386.

Meta. 2024. Faiss. https://ai.meta.com/tools/faiss/. Accessed: 2024-10-27.

Anh Nguyen, Jason Yosinski, and Jeff Clune. 2022. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436.

Nobody-ML. 2024. Soulstar datasets [github repository]. https://github.com/Nobody-ML/SoulStar/blob/main/datasets/data.json.

World Health Organization et al. 2019. Mental disorders. *https://www.who.int/news-room/fact-sheets/detail/mental-disorders*.

Delroy L Paulhus and Simine Vazire. 2007. The self-report method. In *Handbook of research methods in personality psychology*, volume 1, pages 224–239. The Guilford Press.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Hamid Reza Saeidnia, Seyed Ghasem Hashemi Fotami, Brady Lund, and Nasrin Ghiasi. 2024. Ethical considerations in artificial intelligence interventions for mental health and well-being: Ensuring responsible implementation and impact. *Social Sciences*, 13(7):381.

SmartFlowAI. 2024. Emollm datasets [github repository]. https://github.com/SmartFlowAI/EmoLLM/tree/main/datasets.

Randy Stinchfield, John McCready, Nigel E Turner, Susana Jimenez-Murcia, Nancy M Petry, Jon Grant, John Welte, Heather Chapman, and Ken C Winters. 2016. Reliability, validity, and classification accuracy of the dsm-5 diagnostic criteria for gambling disorder and comparison to dsm-iv. *Journal of Gambling Studies*, 32:905–922.

Anthony M Tarescavage and Yossef S Ben-Porath. 2014. Psychotherapeutic outcomes measures: A critical review for practitioners. *Journal of clinical psychology*, 70(9):808–830.

Congchi Yin, Feng Li, Shu Zhang, Zike Wang, Jun Shao, Piji Li, Jianhua Chen, and Xun Jiang. 2024. Mdd-5k: A new diagnostic conversation dataset for mental disorders synthesized via neuro-symbolic llm agents. *arXiv preprint arXiv:2408.12142*.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023b. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Xiutian Zhao, Ke Wang, and Wei Peng. 2024. An electoral approach to diversify LLM-based multi-agent collective decision-making. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2712–2727, Miami, Florida, USA. Association for Computational Linguistics.

Appendix

A Contribution

Qihui Fan: RAG System coding, retrieval database building, survey and implementation on details of proposed RAG systems, and report writing.

Jianqing Ma: Front-end coding and UI design, survey on related works, survey on future optimization, and report writing.

Jiarong Zhu: Preliminary testing on proposed RAG systems, comprehensive benchmarks on proposed system in comparison with general LLMs, and presentation slides design.

Yujia Zhou: Prompt engineering, survey on DSM 5 diagnostic manual, survey on different large language models, benchmarks on proposed system, testing and comprehensive results analysis.

Renlei Huang: Prompt engineering, survey on DSM 5 diagnostic manual, comprehensive benchmarks on proposed system, testing and results analysis.

Minzhou Pan: Proposal writing and polishing, discuss on system optimization.

B Acknowledgment

Trista Y. Lu, M.S. in Clinical Mental Health Counseling, Associate Director, Office of Global Affairs, Stony Brook University, as **psychological knowledge adviser** in this project.

Juyi Lin, Ph.D. Student in Electrical and Computer Engineering, Northeastern University, gave suggestions on future optimization.

Lei Lu, Ph.D. Student in Electrical and Computer Engineering, Northeastern University, gave suggestions on RAG systems.

C Example of Proposed System Dialogue

User: "There are times where it honest to god feels like I'm trapped inside of a cage in my own head and someone else is at the controls. I should clean the house today, nope...[Description of ADHD, a real human description of ADHD from Reddit]"

System Follow-up: "1. Do you experience any repetitive thoughts or actions that feel necessary to perform, even if they don't make sense to you?

2. How often do you find yourself unable to initiate or complete tasks despite having the intention or desire to do so?

3. Have you noticed any patterns in your sleep, such as difficulty falling asleep, staying asleep, or irregular sleep schedules?

4. Can you describe any physical sensations or involuntary movements that occur before you feel unable to control your actions?

5. Is there a specific time of day when these feelings of being "trapped" are more intense or frequent?"

User should response as comprehensive as possible.

Diagnostic: ADHD, with some reasoning and suggestions.

We would also see which agent takes over the system, and details of voting at the back-end.