A Systematic Evaluation of LLM Strategies for Mental Health Text Analysis: Fine-tuning vs. Prompt Engineering vs. RAG

Arshia Kermani, Veronica Perez-Rosas, Vangelis Metsis
Department of Computer Science
Texas State University
San Marcos, TX 78666, USA
{arshia.kermani, vperezr, vmetsis}@txstate.edu

Abstract

This study presents a systematic comparison of three approaches for the analysis of mental health text using large language models (LLMs): prompt engineering, retrieval augmented generation (RAG), and fine-tuning. Using LLaMA 3, we evaluate these approaches on emotion classification and mental health condition detection tasks across two datasets. Fine-tuning achieves the highest accuracy (91% for emotion classification, 80% for mental health conditions) but requires substantial computational resources and large training sets, while prompt engineering and RAG offer more flexible deployment with moderate performance (40-68% accuracy). Our findings provide practical insights for implementing LLM-based solutions in mental health applications, highlighting the trade-offs between accuracy, computational requirements, and deployment flexibility.

1 Introduction

The increasing prevalence of mental health conditions, coupled with limited access to mental health professionals, has created an urgent need for scalable approaches to mental health assessment and support. Traditional diagnostic methods in this area often rely heavily on clinical interviews and self-reported questionnaires, which can be time-consuming, subject to human bias, and limited in their reach (Chung and Teo, 2023). Recent advances in large language models (LLMs) present promising opportunities to enhance mental health assessment through automated analysis of text-based data.

LLMs have demonstrated remarkable capabilities in understanding and generating human language, with recent models like GPT-4, LLaMA 2 (Touvron et al., 2023), and their derivatives achieving unprecedented performance across various natural language processing tasks (Brown et al., 2020). In the medical domain specifically, LLMs have shown potential in tasks ranging from clinical decision support to patient education and medical documentation (Thirunavukarasu et al., 2023). However, their application to mental health assessment presents unique challenges due to the nuanced nature of emotional expression and the critical importance of accuracy in clinical contexts.

Previous research has explored various approaches to leverage LLMs for mental health applications. Studies have investigated the use of zero-shot and few-shot prompt strategies for mental health text classification (Lamichhane, 2023), achieving moderate success in tasks such as detecting stress and depression. Other work has examined the potential of fine-tuned models for specific mental health tasks (Ezerceli and Dehkharghani, 2024), demonstrating improved performance through domain adaptation. However, there remains a significant gap in understanding the relative efficacy of different LLM deployment strategies for mental health assessment tasks.

Our study addresses this gap by conducting a systematic comparison of three distinct approaches for mental health text classification: prompt engineering (including both zero-shot and few-shot variants), retrieval augmented generation (RAG), and fine-tuning. We evaluated these approaches using two complementary datasets: the DAIR-AI Emotion dataset, comprising 20,000 tweets labeled with six basic emotions, and the Reddit SuicideWatch and Mental Health Collection (SWMH), which contains 54,412 posts related to various mental health conditions.

This work makes several key contributions to the field:

1. We provide the first comprehensive comparison of prompt engineering, RAG, and fine-tuning approaches for mental health text classification, offering insights into their relative strengths and limitations.

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025), pages 172–180
May 3, 2025 ©2025 Association for Computational Linguistics

2. We demonstrate the effectiveness of LLaMA 3-based models for mental health assessment tasks, achieving accuracy rates of up to 91% on emotion classification and 80% on mental health condition classification through fine-tuning.

3. We present practical insights into the implementation challenges and resource requirements of each approach, informing future applications in clinical settings.

Our findings have important implications for the development of automated mental health assessment tools, suggesting that while fine-tuning achieves the highest accuracy, both prompt engineering and RAG offer viable alternatives with different trade-offs in terms of computational resources and deployment flexibility. These results contribute to the broader goal of developing reliable, scalable tools to support mental health professionals and improve access to mental health assessment.

In the following sections, we present related work and background (Section 2), detail our methodology (Section 3), present our experimental results (Section 4), discuss their limitations (Section 7), and conclude in (Section 5).

2 Related Work

The intersection of large language models (LLMs) and mental health assessment represents a rapidly evolving field with significant potential for improving healthcare delivery. This section examines the current state of LLMs in healthcare applications and their specific developments in mental health contexts.

2.1 Large Language Models in Healthcare

Recent advances in LLMs have transformed their potential applications in healthcare (He et al., 2023). These models have demonstrated capabilities ranging from clinical decision support and medical documentation to patient education and healthcare communication (Thirunavukarasu et al., 2023). The emergence of domain-specific medical LLMs, such as Med-PaLM 2 and Clinical-Camel, has further enhanced their utility in healthcare settings by incorporating specialized medical knowledge and terminology (Singhal et al., 2025).

The use of LLMs in healthcare applications typically follows three main strategies: fine-tuning existing models, prompt engineering, and retrieval-augmented generation (RAG). Fine-tuning has shown particular promise in specialized medical tasks, with models achieving performance comparable to healthcare professionals in diagnostic scenarios (Singhal et al., 2025). Prompt engineering approaches have shown effectiveness in zero-shot and few-shot learning contexts, allowing flexible deployment without extensive retraining (Liu et al., 2023). RAG methods have emerged as a promising approach for grounding LLM responses with domain knowledge, thereby reducing hallucination and improving reliability (Lewis et al., 2020; Gao et al., 2023).

2.2 Mental Health Text Analysis

Mental health assessment presents unique challenges for automated analysis due to the subtle nature of emotional expression and the critical importance of accurate interpretation. Traditional approaches to the analysis of mental health text have relied on rule-based systems and classical machine learning techniques, often struggling to capture the nuanced context necessary for an accurate assessment (Kazdin, 2011).

Recent work has begun to explore the potential of LLMs for mental health applications. Studies have shown promising results in the detection of signs of depression, anxiety, and suicidal ideation from social media posts (Ma et al., 2024). The evolution of LLM capabilities has particular relevance for mental health research. Recent studies have shown that advanced LLM versions can provide human-level interpretations in qualitative coding tasks (Dunivin, 2024) and achieve accuracy comparable to mental health professionals in certain diagnostic contexts (Kim et al., 2024).

When applied to qualitative analysis, LLMs have demonstrated the ability to perform various analytical approaches, including thematic analysis, content analysis, and grounded theory, as validated by human experts (Xiao et al., 2023b; Rasheed et al., 2024). This suggests potential for enhancing, rather than replacing, traditional qualitative analysis methods in mental health research. However, these applications pose important challenges, including the need for accurate predictions given the critical nature of mental health assessment, concerns about privacy and data security, and the importance of maintaining therapeutic alliance in clinical settings (Byers et al., 2023).

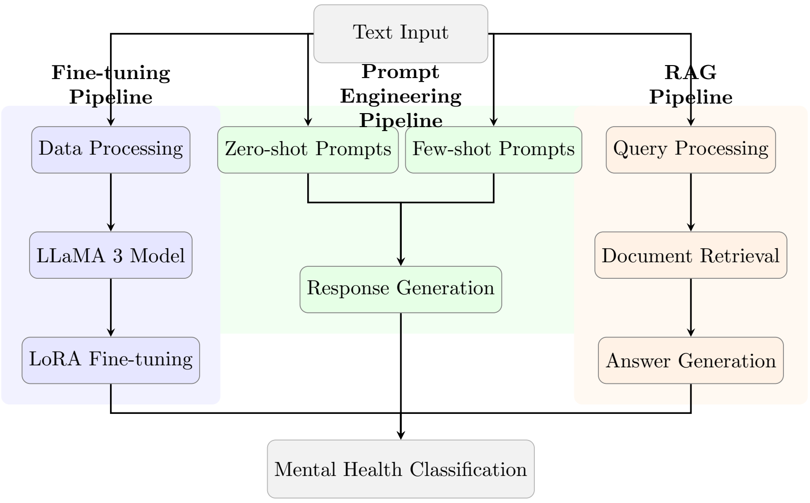

Figure 1: Overview of our experimental framework comparing three LLM deployment approaches (fine-tuning, prompt engineering, and RAG) for mental health text analysis. Each approach processes the same input data through distinct pipelines, enabling systematic comparison of their effectiveness.

2.3 Evaluation Frameworks

The evaluation of LLMs in mental health applications requires careful consideration of both model performance and clinical utility. Although traditional metrics such as accuracy and F1 scores provide important quantitative measures, they must be contextualized within the broader requirements of mental health assessment. Recent work has highlighted the importance of developing comprehensive evaluation frameworks that consider not only classification accuracy but also the ability of the model to provide interpretable and clinically relevant output (Xu et al., 2024).

The existing literature shows particular gaps in understanding the effectiveness of different LLM deployment strategies for mental health applications. Although studies have examined individual approaches, comprehensive comparisons of fine-tuning, prompt engineering, and RAG methods in mental health contexts remain limited. This gap is particularly significant given the practical considerations, such as the availability of resources and the computing requirements involved in deploying these different approaches in clinical settings.

Furthermore, the evaluation of LLMs in mental health applications must consider ethical implications and potential biases (Chancellor et al., 2019; Gallegos et al., 2024). This includes ensuring that models do not perpetuate existing biases in mental health diagnosis and that they maintain appropriate boundaries in therapeutic contexts. These considerations inform both the choice of evaluation metrics and the interpretation of results in mental health applications of LLMs.

3 Methodology

This study implements and evaluates three distinct approaches for LLMs in mental health text analysis: fine-tuning, prompt engineering, and retrieval-augmented generation (RAG). We utilize the LLaMA 3 model architecture, specifically the 8B parameter version, as our base model across all experiments to ensure fair comparison. Figure 1 presents the overall pipeline of our experimental framework. As shown in the Figure, our prompt engineering approach investigates both zero-shot and few-shot learning capabilities of the LLaMA 3 model for mental health text analysis.

3.1 Datasets

Our evaluation employs two complementary datasets that capture different aspects of mental health and emotional expression in text.

DAIR-AI Emotion Dataset The DAIR-AI (Saravia et al., 2018) dataset comprises `20,000` tweets labeled with one of six fundamental emotions: *joy*, *sadness*, *anger*, *fear*, *love*, or *surprise*. During our experiments, we maintain the original paper's data split: training set: 16,000 samples (80%); validation set: 2,000 samples (10%); test set: 2,000 samples (10%). Table 1 shows a sample of the DAIR-AI dataset.

Sample Text	Label
i can go from feeling so hopeless to so damned hopeful just from being around someone who cares and is awake	sadness
im grabbing a minute to post i feel greedy wrong	anger
i am ever feeling nostalgic about the fireplace i will know that it is still on the property	love
ive been taking or milligrams or times recommended amount and ive fallen asleep a lot faster but i also feel like so funny	surprise
i feel as confused about life as a teenager or as jaded as a year old man	fear
i have been with petronas for years i feel that petronas has performed well and made a huge profit	joy

Table 1: Samples and labels from the DAIR-AI Dataset.

Sample Text	Label
Wanting to skip my exam on Saturday because I'm so tired and mentally fried that a few days off might help.	Anxiety
Do other bipolar folks have problems with substance abuse? I've had overdoses and ended up in the ICU, and now I take my meds as prescribed.	Bipolar
Anonymous Entry: plz be nice. I've become a deteriorated husk of a person—hopefully this is my last moment of self-awareness.	Depression
I'm pretty sure my friend is suicidal; he keeps saying self-hating things like "I'm just a little emo prick." What do I do?	Suicide Watch

Table 2: Samples and labels from the SWMH Dataset.

Reddit SuicideWatch and Mental Health Collection (SWMH) The SWMH (Ji et al., 2021) dataset contains 54,412 Reddit posts that discuss various mental health conditions. Each post is labeled with one of the following categories: *depression*, *anxiety*, *bipolar disorder*, or *suicidal ideation*. The dataset is divided by the authors that published it as follows: training set: 34,824 samples (64%); validation set: 8,706 samples (16%); test set: 10,882 samples (20%). Table 2 presents a sample of the SWMH dataset.

3.1.1 Data Preprocessing

We conduct a preprocessing step on both datasets to ensure data quality and standardization. 1) Removal of URLs, user mentions, and special characters. 2) Standardization of text encoding to UTF-8. 3) Truncation of texts exceeding the model's maximum token limit (2048 tokens). 4) Verification of label consistency and removal of any samples with ambiguous or missing labels.

3.2 Experimental Setup

Our experiments are run on A100 GPU with 83.48 GB of RAM and 200 GB of disk space on Google Colab Pro+. We use the 8B parameter version of LLaMA 3 (Grattafiori et al., 2024), applying 4-bit quantization to optimize memory usage while preserving model performance. The base model configuration includes a `float16` precision for computational efficiency and a LLaMA tokenizer with right-padding and end-of-sequence tokens. During fine-tuning, we used the following hyperparameters: learning rate: 2e-4 with cosine schedule; weight decay: 0.001; batch size: 1 per device; gradient accumulation steps: 8; training epochs: 1; maximum steps: -1.

Classification evaluations are performed using F1 score, precision, and recall as our main metrics.

3.3 Fine-tuning

Our fine-tuning approach adapts the LLaMA 3 model to the specific requirements of mental health text analysis while maintaining computational efficiency. Figure 2 illustrates the fine-tuning architecture and process flow. To address the computational challenges of fine-tuning large models, we employed Low-Rank Adaptation (LoRA) (Hu et al., 2021). This approach significantly reduced the number of trainable parameters while maintaining model performance. Our LoRA configuration included a rank of 64 and an alpha scaling factor of 16.

3.4 Zero-Shot Prompting

We use zero-shot prompting to classify text without providing prior examples.

We used the following prompt template for both DAIR-AI and SWHM datasets, adjusting for the corresponding labels.

```
Analyze the emotional content in the following
text and classify it into exactly one of these
categories: joy, sadness, anger, fear, love,
or surprise. Provide only the category label as
output.

Text: [input_text]
```

3.5 Few-Shot Prompting

In our few-shot approach, we first select *random* examples and their corresponding labels from the training set and provide them as additional input

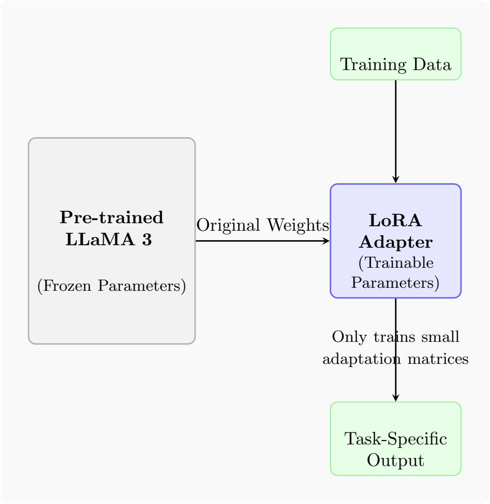

Figure 2: Architecture of our fine-tuning implementation, showing the integration of LoRA for parameter-efficient adaptation, the training process flow, and the evaluation pipeline.

to guide the model's reasoning. We included two examples for each label, ensuring balanced representation across classes. Example prompt:

```
Task: Classify the emotional content of text into
one of these categories: joy, sadness, anger, fear,
love, or surprise.

Example 1:
Text: "Finally got my dream job after months of
trying!"
Emotion: joy

Example 2:
Text: "I miss my old friends so much it hurts."
Emotion: sadness

Example 3:
Text: "How dare they treat people this way!"
Emotion: anger

[Additional examples...]

Now classify this text:
[input_text]
```

3.6 Retrieval-Augmented Generation (RAG)

The success of RAG models hinges on their capability to locate and retrieve pertinent examples and on the LLM's proficiency in effectively using the retrieved information. We believe that this is particularly helpful for the classification of mental health text, where additional context and examples can better inform the model. Since RAG can operate with much fewer training examples than is usually required to fine-tune an LLM model on a specific task, we consider RAG as a middle ground between few-shot prompting and fine-tuning.

We implement a RAG model that incorporates relevant contextual information derived from the training dataset during inference time. Figure 3 presents the overall architecture of our RAG implementation. It retrieves relevant examples from a knowledge base to be used during inference time to inform the model's decision, which are then added as part of the generation input.

3.6.1 Knowledge Base Construction

While implementing our model, we constructed a specialized knowledge base to support the retrieval process for each dataset:

Embedding Generation: We utilized the BAAI/bge-small-en-v1.5 (Xiao et al., 2023a) model to generate dense vector representations of training examples. The resulting embeddings are added to a vector database for storage and retrieval.

Vector Database: We use ChromaDB (Contributors, 2025) as our vector database. Our configuration includes cosine similarity as the distance metric, HNSW (Hierarchical Navigable Small World) as the indexing method, and category labels and source information as metadata storage.

3.6.2 Retrieval Process

During retrieval, we start by embedding the input query with the BAAI/bge-small-en-v1.5 model, then we selected the top-k nearest neighbors considering diverse examples across categories, finally, we form an unified context using the retrieved examples. The retriever returns the three most similar documents for each query, balancing context richness with computational efficiency.

Generation Component: The generation component combines the retrieved context with the input text to produce classification decisions. We use the following prompt template:

```
Review the following examples and context:

[Retrieved Context Documents]

Based on these examples, classify the emotional
content of the following text into one of these
categories: joy, sadness, anger, fear, love,
or surprise. Provide only the category label.

Text to classify: [input_text]
```

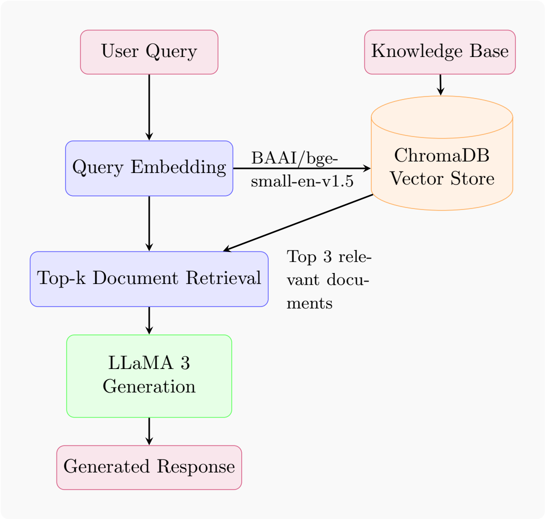

Figure 3: Architecture of our RAG model, illustrating the flow from input processing through retrieval and generation stages. The diagram shows how the system integrates embedded knowledge retrieval with LLM-based classification.

Method	DAIR-AI Emotion		SWMH	
	Accuracy	F1	Accuracy	F1
Fine-tuning	**91%**	**0.87**	**80%**	**0.81**
Zero-shot	49%	0.38	68%	0.67
Few-shot	39%	0.30	45%	0.57
RAG	47%	0.32	56%	0.45

Table 3: Performance comparison across methods and datasets. F1-scores (macro) provide a balanced measure of performance across all categories.

4 Results

Our experiments show significant performance variations across fine-tuning, prompt engineering, and retrieval augmented generation (RAG) approaches. Figure 4 presents the comparative performance in all methods.

Fine-tuning achieves the best performance, with accuracies of 91% and 80% on the DAIR-AI Emotion and SWMH datasets. Notably, zero-shot prompting emerged as the second-best performing approach, reaching 49% and 68% on each dataset, surpassing both the few-shot prompting and RAG. This suggests that carefully crafted prompts can effectively leverage the model's pre-trained knowledge for mental health text analysis, even without additional examples or context. However, we should clarify that in this study, we only used simple prompts as in the example shown in section 3.4 and kept them consistent throughout all experiments.

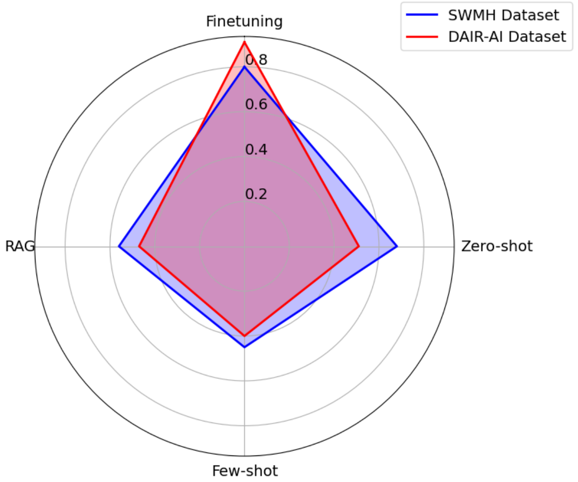

Figure 4: Performance comparison across different approaches for both datasets. The graph shows accuracy scores for fine-tuning, zero-shot prompting, few-shot prompting, and RAG methods.

Table 3 summarizes the classification performance in all methods and datasets. The performance disparity between the different approaches is particularly noteworthy, with the gap being more pronounced in emotion classification compared to the detection of mental health conditions.

Compared to prior work using relation networks (Ji et al., 2021), our approach demonstrates a 15.3% absolute improvement in classification accuracy and a significant boost in F1-score on the SWMH dataset. While the baseline relied on handcrafted sentiment and topic modeling features for classification, our fine-tuned LLaMA 3 model effectively captures the intricate language nuances in mental health discourse, yielding superior predictive performance. Furthermore, our zero-shot prompting approach (68% accuracy) surpassed the baseline's performance, suggesting that LLMs can generalize mental health-related text classification without requiring domain-specific feature engineering.

The greater advantage of the fine-tuning approach in the DAIR-AI Emotion dataset compared to the SWMH can be partially attributed to the dataset sizes (54.4K vs. 20K). A larger training set enables for a more effective fine-tuning, whereas this advantage diminishes and may even be reversed with smaller training sets.

4.1 Analysis of Best-Performing Methods

A more detailed evaluation of fine-tuning, zero-shot prompting, few-shot prompting, and RAG reveals distinct patterns in their effectiveness across different classification tasks. Table 4 presents a comprehensive comparison of these approaches in all classification categories.

The fine-tuned model demonstrated varying levels of performance across emotion categories. It achieved exceptional results for basic emotions such as *joy* and *sadness* (F1-scores of 0.94 and 0.95), followed by a strong performance for *anger* and *fear* (both 0.89). More complex emotional states proved more challenging, with *love* achieving an F1-score of 0.81 and *surprise* showing the lowest performance at 0.72. For mental health conditions, the model achieved the highest performance in detecting *anxiety* and *bipolar* disorder (F1-scores of 0.86 and 0.85, respectively) while maintaining robust performance for *depression* detection (F1: 0.79).

Zero-shot prompting showed notably strong performance in mental health condition detection, particularly for *depression* and *anxiety* (F1-scores of 0.70 and 0.74). However, its performance on emotion classification varied considerably. While achieving moderate results for *joy* and *sadness* (F1-scores of 0.56 and 0.58), it struggled significantly with more nuanced emotions like *love* and *surprise* (F1-scores of 0.25 and 0.26). The approach showed particularly low recall for *fear* detection despite high precision, indicating a conservative classification pattern for this category.

4.2 Analysis of Less Successful Methods

The evaluation of RAG and few-shot prompting revealed important insights about their practical limitations in mental health text analysis. Table 5 presents the key performance metrics for these approaches.

The RAG system achieved moderate performance levels (47% and 56% accuracy in DAIR-AI and SWMH, respectively), with effectiveness heavily dependent on retrieval quality. Performance was strongest when highly relevant context was successfully retrieved (64% accuracy) but dropped significantly with lower-quality retrievals (31% accuracy). Few-shot prompting showed unexpectedly lower performance compared to zero-shot approaches, suggesting that example-based prompting may introduce conflicting patterns that complicate the classification task in mental health contexts.

Our findings indicate that while RAG and few-shot prompting offer benefits in terms of interpretability and flexibility, their current implementations face significant challenges in achieving reliable performance for mental health text analysis task (Chung et al., 2023).

5 Conclusion

This study provided a systematic comparison of fine-tuning, prompt engineering, and retrieval augmented generation for mental health text classification. Fine-tuning showed superior performance, achieving 91% accuracy in emotion classification and 80% in the detection of mental health conditions, although at the cost of significant computational requirements. Zero-shot prompting emerged as a viable alternative, particularly for mental health condition detection (68% accuracy), suggesting that carefully designed prompts can effectively leverage pre-trained knowledge when fine-tuning is not feasible. However, both RAG and few-shot prompting showed limited effectiveness, with performance heavily dependent on retrieval quality and example selection.

These findings have important implications for developing automated mental health assessment tools. While fine-tuned models show promise for reliable screening applications, their varying performance across different emotional states and mental health conditions suggests current approaches may be better suited for initial assessment rather than definitive diagnosis.

Future research directions include the investigation of hybrid approaches that combine the strengths of multiple methods, the development of more efficient fine-tuning techniques, and the exploration of ways to improve the detection of nuanced psychological states. In addition, more work is needed to validate these approaches in clinical settings and across diverse populations.

6 Ethical Considerations

This study follows ethical guidelines on data usage, model reliability, and the responsible deployment of large language models (LLMs) for mental health evaluation. The datasets used in this research, DAIR-AI Emotion and SWMH, are publicly available, ensuring transparency and reproducibility. The SWMH dataset consists of publicly shared Reddit posts, while the DAIR-AI Emotion dataset contains labeled social media text. No personally identifiable information (PII) was processed, and no direct engagement with individuals was conducted.

Automated systems carry the risk of misclassifi-

Dataset	**Category**	**Fine-tuning**		**Zero-shot**		**Few-shot**		**RAG**	
		F1	Prec/Rec	F1	Prec/Rec	F1	Prec/Rec	F1	Prec/Rec
DAIR-AI	Joy	0.94	0.94/0.93	0.56	0.80/0.43	0.35	0.35/0.35	0.44	0.82/0.30
	Sadness	0.95	0.95/0.94	0.58	0.47/0.75	0.24	0.20/0.30	0.27	0.35/0.22
	Anger	0.89	0.88/0.91	0.46	0.39/0.57	0.36	0.30/0.45	0.29	0.41/0.23
	Fear	0.89	0.89/0.88	0.17	0.88/0.09	0.29	0.25/0.35	0.31	0.45/0.24
	Love	0.81	0.80/0.82	0.25	0.26/0.25	0.22	0.20/0.25	0.30	0.44/0.23
	Surprise	0.72	0.73/0.71	0.26	0.24/0.27	0.34	0.30/0.40	0.31	0.44/0.24
	Average	0.87	0.86/0.87	0.38	0.40/0.36	0.30	0.27/0.33	0.32	0.45/0.25
SWMH	Depression	0.79	0.78/0.80	0.70	0.59/0.84	0.52	0.74/0.40	0.40	0.45/0.36
	Anxiety	0.86	0.87/0.86	0.74	0.90/0.63	0.56	0.78/0.44	0.61	0.72/0.53
	Bipolar	0.85	0.87/0.83	0.62	0.88/0.48	0.63	0.80/0.53	0.40	0.45/0.36
	Suicide	0.75	0.75/0.75	0.61	0.68/0.56	0.55	0.64/0.48	0.39	0.44/0.35
	Average	0.81	0.82/0.81	0.67	0.70/0.64	0.57	0.75/0.47	0.45	0.52/0.40

Table 4: Detailed performance metrics for Fine-tuning, Zero-shot, Few-shot, and RAG approaches across all categories. Precision/Recall values are presented as Prec/Rec.

Method	**DAIR-AI**		**SWMH**	
	Acc.	Top Category	Acc.	Top Category
RAG	47%	Joy (0.44)	56%	Anxiety (0.61)
Few-shot	39%	Anger (0.36)	45%	Bipolar (0.63)

Table 5: Performance summary of RAG and few-shot approaches. The top Category shows the highest F1 score achieved for any single category.

cation, especially in sensitive areas such as depression and suicidal ideation. Any potential application of these models outside of research settings would require extensive validation, supervision by clinical professionals, and adherence to ethical and regulatory standards to avoid misinformation or unintended consequences.

7 Limitations

This study demonstrated the potential of large language models for psychological assessments but also showed a few limitations. Fine-tuning a model as extensive as LLaMA-3 8B required significant computational resources. This dependency on high-end resources limits the accessibility of our approach for researchers with constrained computational capacities. Furthermore, the models were trained and evaluated on the DAIR-AI Emotion and SWMH datasets, which, while diverse, may not fully capture the complexity and variability of real-world psychological text data. This could restrict the generalizability of the findings to other domains, languages, or text formats, e.g., short vs. long text. Additionally, our study does not address the practical integration of these tools into clinical workflows, which would require collaboration with domain experts and rigorous validation.

Addressing these limitations in future research could improve the accessibility, generalizability, and ethical applicability of LLM-based psychological assessment tools.

References

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

Morgan Byers, Mark Trahan, Erica Nason, Chinyere Eigege, Nicole Moore, Micki Washburn, and Vangelis Metsis. 2023. Detecting intensity of anxiety in language of student veterans with social anxiety using text analysis. *Journal of Technology in Human Services*, 41:1–21.

Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent MB Silenzio, and Munmun De Choudhury. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 79–88.

J. Chung and J. Teo. 2023. Single classifier vs. ensemble machine learning approaches for mental health prediction. *Brain Informatics*, 10.

Neo Christopher Chung, George Dyer, and Lennart Brocki. 2023. Challenges of large language models for mental health counseling. *Preprint*, arXiv:2311.13857.

Chroma Contributors. 2025. Chroma: The ai-native open-source embedding database. Accessed: 2025-02-12.

Zackary Okun Dunivin. 2024. Scalable qualitative coding with llms: Chain-of-thought reasoning matches human performance in some hermeneutic tasks. *arXiv preprint arXiv:2401.15170*.

Ozay Ezerceli and Rahim Dehkharghani. 2024. Mental disorder and suicidal ideation detection from social media using deep neural networks. *Journal of Computational Social Science*, pages 1–31.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2023. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv preprint arXiv:2310.05694*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Shaoxiong Ji, Xue Li, Zi Huang, and Erik Cambria. 2021. Suicidal ideation and mental disorder detection with attentive relation networks. *Neural Computing and Applications*.

Alan E. Kazdin. 2011. Conceptualizing the challenge of reducing interpersonal violence and other public health problems: Behavioral and mental health disorders. *American Psychologist*, 66(7):621–639.

Jiyeong Kim, Kimberly G Leonte, Michael L Chen, John B Torous, Eleni Linos, Anthony Pinto, and Carolyn I Rodriguez. 2024. Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder. *NPJ Digital Medicine*, 7(1):193.

Bishal Lamichhane. 2023. Evaluation of chatgpt for nlp-based mental health applications. *Preprint*, arXiv:2303.15727.

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.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yingzhuo Ma, Yi Zeng, Tong Liu, Ruoshan Sun, Mingzhao Xiao, and Jun Wang. 2024. Integrating large language models in mental health practice: a qualitative descriptive study based on expert interviews. *Frontiers in Public Health*, 12:1475867.

Zeeshan Rasheed, Muhammad Waseem, Aakash Ahmad, Kai-Kristian Kemell, Wang Xiaofeng, Anh Nguyen Duc, and Pekka Abrahamsson. 2024. Can large language models serve as data analysts? a multi-agent assisted approach for qualitative data analysis. *arXiv preprint arXiv:2402.01386*.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Hugo Touvron, Louis Martin, Kevin Stone, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023a. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Ziang Xiao, Xingdi Yuan, Q Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023b. Supporting qualitative analysis with large language models: Combining codebook with gpt-3 for deductive coding. In *Companion proceedings of the 28th international conference on intelligent user interfaces*, pages 75–78.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32.