

Standardization of Psychiatric Diagnoses — Role of Fine-tuned LLM Consortium and OpenAI-gpt-oss Reasoning LLM Enabled Decision Support System

Eranga Bandara^a, Ross Gore^a, Atmaram Yarlagadda^b, Anita H. Clayton^d, Preston Samuel^c, Christopher K. Rhea^a, Sachin Shetty^a

^a*Old Dominion University, Norfolk, VA, USA*

^b*McDonald Army Health Center, Newport News, VA, USA*

^c*Blanchfield Army Community Hospital, Fort Campbell, KY, USA*

^d*Department of Psychiatry and Neurobehavioral Sciences,
University of Virginia School of Medicine, Charlottesville, VA, USA*

Abstract

The diagnosis of most mental disorders, including psychiatric evaluations, primarily depends on dialogues between psychiatrists and patients. This subjective process can lead to variability in diagnoses across clinicians and patients, resulting in inconsistencies and challenges in achieving reliable outcomes. To address these issues and standardize psychiatric diagnoses, we propose a Fine-Tuned Large Language Model (LLM) Consortium and OpenAI-gpt-oss Reasoning LLM-enabled Decision Support System for the clinical diagnosis of mental disorders. Our approach leverages fine-tuned LLMs trained on conversational datasets involving psychiatrist–patient interactions focused on mental health conditions (e.g., depression). These models are capable of identifying mental disorders with high accuracy based on natural language input. The diagnostic predictions from individual models are aggregated through a consensus-based decision-making process, refined by the OpenAI-gpt-oss reasoning LLM. We propose a novel method for deploying LLM agents that orchestrate seamless communication between the LLM

Email addresses: cmedawer@odu.edu (Eranga Bandara), rgore@odu.edu (Ross Gore), atmaram.yarlagadda.civ@health.mil (Atmaram Yarlagadda), AHC8V@uvahealth.org (Anita H. Clayton), preston.1.samuel.mil@health.mil (Preston Samuel), crhea@odu.edu (Christopher K. Rhea), sshetty@odu.edu (Sachin Shetty)

consortium and the reasoning LLM, ensuring transparency, reliability, and responsible AI across the entire diagnostic workflow. Each LLM in the consortium was fine-tuned using the Unislot library on Google Colab’s Tesla GPUs. To optimize performance on consumer-grade hardware, we employed Low-Rank Adaptation with 4-bit quantization (QLoRA). Experimental results demonstrate the transformative potential of combining fine-tuned LLMs with a reasoning model to create a robust and highly accurate diagnostic system for mental health assessment. A prototype of the proposed platform, integrating three fine-tuned LLMs with the OpenAI-gpt-oss reasoning LLM, was developed in collaboration with the U.S. Army Medical Research Team in Norfolk, Virginia, USA. To the best of our knowledge, this work represents the first application of a fine-tuned LLM consortium integrated with a reasoning LLM for clinical mental health diagnosis—paving the way for next-generation AI-powered eHealth systems aimed at standardizing psychiatric diagnoses.

Keywords: Psychiatric Diagnosis, LLM-Reasoning, OpenAI-gpt-oss, LLM, Llama-3, Responsible AI

1. Introduction

The diagnosis of mental disorders remains one of the most complex and nuanced challenges in clinical medicine. Unlike many physical illnesses, which can be confirmed through objective tests such as blood panels, imaging, or genetic screening, psychiatric diagnoses rely heavily on subjective assessment [1]. The current diagnostic process is predominantly based on conversations between psychiatrists and patients, guided by standardized criteria outlined in manuals such as the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) [2]. During clinical interviews, psychiatrists assess the patient’s reported symptoms, behavioral cues, emotional responses, and historical context to reach a diagnosis. However, this approach introduces significant variability and subjectivity. Diagnoses can differ between clinicians assessing the same patient, and even the same clinician may interpret similar symptoms differently across patients [3]. This variability can be attributed to differences in clinical experience, interpersonal dynamics, cultural context, and interpretation of symptom severity or duration [4]. Moreover, time constraints, communication barriers, and unconscious bias can further compromise diagnostic consistency. These challenges contribute

to misdiagnosis, delayed treatment, and inconsistent care, ultimately affecting patient outcomes [5].

In recent years, advances in artificial intelligence, particularly LLMs [6, 7], have demonstrated exceptional capabilities in natural language understanding and reasoning. These developments open new avenues for enhancing the diagnostic process in psychiatry by introducing data-driven, scalable, and interpretable decision support systems. In this context, we propose a Fine-Tuned LLM Consortium and OpenAI-gpt-oss Reasoning LLM [8, 9, 10, 11] enabled decision support system designed to assist and augment the clinical diagnosis of mental disorders. The LLMs in the system are trained/fine-tuned on curated conversational datasets that simulate or replicate psychiatrist–patient interactions related to mental health evaluations, such as those for depression, anxiety, PTSD, and schizophrenia [12, 13]. By fine-tuning a set of LLMs on these datasets, the system learns to recognize patterns indicative of specific mental disorders. Predictions from these models are then aggregated through a consensus-based reasoning process, led by OpenAI-gpt-oss, a dedicated reasoning LLM that evaluates and refines the diagnostic output. The specialized orchestration LLM agent [14, 15, 16] coordinates communication between components, ensuring an end-to-end workflow that is accurate, interpretable, and aligned with DSM-5 diagnostic standards [17]. To enable deployment in resource-constrained environments, each model is fine-tuned using the Unslot library on Google Colab GPUs [18] with Low-Rank Adaptation and 4-bit quantization (QLoRA) [19], ensuring high performance with minimal computational overhead. The proposed approach aims not to replace clinicians but to support them with a robust, evidence-based, and transparent tool that improves diagnostic precision and consistency. The following are our main contributions of this research.

1. Fine-tuning a consortium of LLMs to analyze conversational data and predict diagnoses of mental disorders.
2. Incorporating the OpenAI-gpt-oss reasoning model to provide the final diagnosis based on the LLM consortium’s predictions.
3. Automating the end-to-end diagnostic decision-making process for mental disorders through an LLM consortium integrated with the OpenAI-gpt-oss model, orchestrated by LLM agents to ensure transparency, reliability, and adherence to Responsible AI principles.
4. Implementing the prototype of the platform, integrating three LLMs with OpenAI-gpt-oss reasoning LLM, in collaboration with the U.S.

Army Medical Research Team in Norfolk, Virginia, USA.

The remainder of the paper is organized as follows: Section 2 introduces the core technologies that underpin the proposed AI-assisted psychiatric diagnostic platform. Section 3 details the overall system architecture, highlighting the integration of large language models and reasoning engines. Section 4 outlines the platform’s core functionalities and operational workflow. Section 5 presents implementation details and evaluates the system’s performance across diagnostic tasks. Section 6 reviews related work and contextualizes our approach within the landscape of LLM-based medical diagnosis systems. Finally, Section 7 concludes the paper and discusses potential directions for future research and clinical deployment.

2. Background

This section provides a foundational overview of the core technologies underpinning the proposed AI-assisted psychiatric diagnostic platform. In particular, we highlight advancements in Large Language Models (LLMs), reasoning-capable LLMs, fine-tuning techniques, and the emerging paradigm of AI agents.

2.1. Large Language Models (LLMs)

Large Language Models (LLMs) are advanced deep neural networks trained on extensive text corpora to comprehend, generate, and reason using natural language. They form the backbone of modern Natural Language Processing (NLP) systems [20] and have demonstrated exceptional performance across a wide range of tasks, including text summarization, machine translation, dialogue generation, and question answering [6].

Several prominent LLMs such as OpenAI’s GPT [21], Meta’s Llama [22, 7], Mistral [23], and Alibaba’s Qwen2 [24] are available in both proprietary and open-source formats. Open-source LLMs offer significant advantages for healthcare applications, including transparency, customizability, and lower deployment costs. For instance, Llama-3[22, 7] is valued for its compact architecture and strong performance, even on resource-constrained systems. Mistral[25] features optimized attention mechanisms for faster inference, while Qwen2 [24] is designed for multilingual and on-device use, making it suitable for scalable, privacy-preserving deployments.

2.2. Reasoning LLMs

While foundational LLMs excel in pattern recognition and natural language generation, they often lack the capacity for structured, multi-step reasoning. Reasoning LLMs[8] address this limitation by being specifically designed or fine-tuned to synthesize diverse inputs, resolve conflicting information, and support logical decision-making processes. Unlike traditional LLMs that primarily rely on next-token prediction, reasoning models simulate higher-order cognitive functions akin to human deductive reasoning[21, 26].

OpenAI-gpt-oss [9, 27] is a reasoning LLM designed to perform advanced evaluative and comparative tasks across multiple inputs. Unlike traditional generative LLMs that focus on single-output prediction, OpenAI-gpt-oss is capable of synthesizing responses, resolving contradictions, and applying logical inference to arrive at consistent, well-reasoned conclusions. It excels in tasks involving multi-model output reconciliation, ranking, and consensus generation. These capabilities make it particularly suitable for applications that require structured reasoning, such as diagnostic decision support, content validation, and multi-agent coordination, where interpretability and reliability are critical.

2.3. LLM Fine-tuning

Fine-tuning is a technique for adapting pre-trained LLMs to specific downstream tasks or domains. It involves retraining the model on curated datasets that reflect the target domain’s language, structure, and semantics, allowing the model to produce outputs more aligned with specialized applications such as clinical diagnostics [28, 29].

To optimize for efficiency and scalability of fine-tuning, Low-Rank Adaptation (LoRA)[30] is commonly used. LoRA introduces trainable low-rank matrices into the transformer architecture, enabling task-specific adaptation while significantly reducing the number of trainable parameters. In resource-constrained environments, Quantized LoRA (QLoRA)[19] provides an even more memory-efficient approach by quantizing model weights to 4-bit representations. QLoRA retains most of the performance benefits of full-precision fine-tuning while dramatically reducing memory and compute requirements. Together, these techniques make fine-tuning large models feasible on modest hardware, supporting wider adoption of LLMs in specialized domains.

Several open-source libraries facilitate efficient fine-tuning workflows for LLMs. Unslloth[31], for example, enables high-speed, memory-efficient fine-tuning of models such as LLaMA, Mistral, and Qwen using LoRA and

QLoRA techniques. It is optimized for both consumer-grade GPUs (e.g., NVIDIA RTX 3090) and cloud-based environments, including TPU-enabled platforms like Google Colab[18]. Successful fine-tuning of large models generally requires GPUs with ample VRAM and compute capabilities. High-performance GPUs such as the NVIDIA A100 (40GB/80GB) and H100 are well-suited for large-scale training workloads, while more accessible GPUs like the NVIDIA RTX 3090/4090 and Tesla T4 provide sufficient resources for small to medium-scale fine-tuning and prototyping [32].

2.4. AI Agents and Agentic AI

AI agents are autonomous computational entities designed to perform complex tasks by interacting with data sources, machine learning models, and external APIs within dynamic or uncertain environments. When these agents are powered by LLMs, they are referred to as LLM agents, capable of interpreting natural language instructions, generating structured outputs, managing tasks, and coordinating actions across digital ecosystems [14, 15].

Agentic AI extends this concept by organizing multiple LLM agents into collaborative, role-specialized systems that demonstrate advanced capabilities such as long-term planning, self-reflection, adaptive behavior, and multi-agent coordination [15, 33]. These systems operate through agent hierarchies or workflows in which each agent performs a specific role, such as prompt engineering, retrieval, inference, evaluation, or integration. The modularity of agentic architectures enhances scalability, interpretability, and reusability, making them particularly suitable for domains requiring structured reasoning, task delegation, and reliable decision support.

3. System Architecture

Figure 1 describes the architecture of the platform. The proposed platform is composed of 4 layers: 1) Data Lake layer, 2) LLM Agent Layer, 3) LLM Layer, and 4) OpenAI-gpt-oss Reasoning Layer. Below is a brief description of each layer.

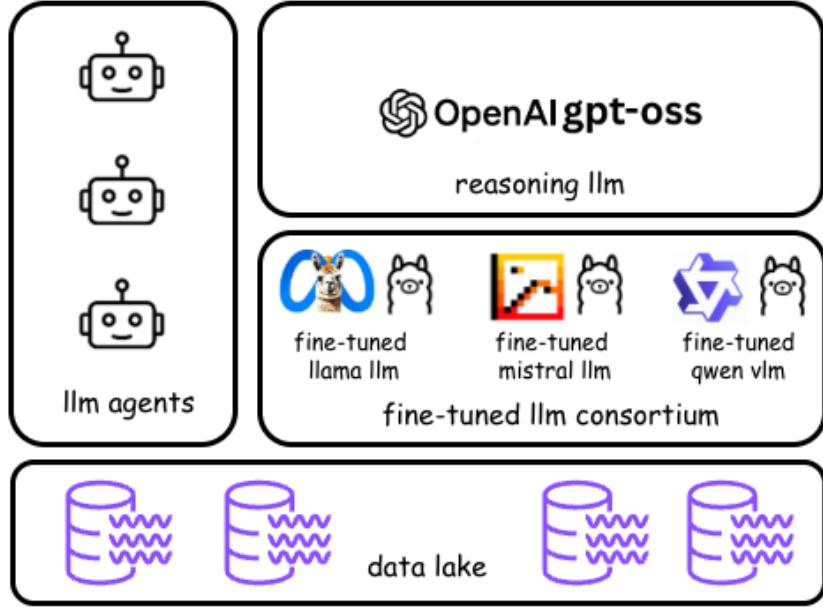


Figure 1: Platform architecture.

3.1. Data Lake Layer

The Data Lake layer serves as the foundational infrastructure for managing and storing extensive conversational datasets that simulate or replicate psychiatrist–patient interactions relevant to mental health evaluations. This centralized repository is designed to support the training and fine-tuning [34] of LLMs for predictive psychiatric diagnosis. It hosts a diverse collection of large-scale, labeled datasets comprising transcribed conversations, symptom narratives, and corresponding clinician-verified diagnoses—aligned with standardized diagnostic frameworks such as the DSM-5 [35, 36]. These datasets are critical for enabling the LLMs to learn clinically relevant patterns, understand nuanced symptom expressions, and associate them with specific mental disorders. By capturing a wide range of linguistic, contextual, and diagnostic variability, the Data Lake layer empowers the platform to develop robust, generalizable models capable of supporting accurate and consistent mental health assessments across diverse patient populations [37, 38].

3.2. LLM Agent Layer

The LLM agent layer functions as the orchestration and automation core of the platform, enabling seamless integration and coordination across the

Data Lake, fine-tuned LLMs, and the OpenAI-gpt-oss reasoning engine. The LLM agents act as orchestrators, which are responsible for all custom prompt engineering required to facilitate effective communication between components, ensuring that diagnostic predictions are generated, aggregated, and refined in a coherent and structured manner. Specifically, the LLM agents dynamically construct prompts using patient-clinician conversational data retrieved from the Data Lake [39, 40]. These prompts are used to query the ensemble of fine-tuned LLMs, each of which outputs preliminary diagnostic assessments based on the detected symptoms and inferred clinical patterns. The agents then aggregate these outputs and format them into a consolidated, structured prompt tailored for the OpenAI-gpt-oss reasoning LLM [21]. The OpenAI-gpt-oss model leverages its advanced reasoning capabilities to evaluate and synthesize the collective outputs of the LLM consortium, ultimately generating a refined and consistent diagnostic prediction. By adapting prompts to match the input requirements and context of each model, the LLM Agent Layer ensures optimal information flow and model interoperability [14]. This orchestrated process not only enhances diagnostic accuracy and consistency but also enables a fully automated, end-to-end AI-driven diagnostic workflow, as illustrated in Figure 2.

3.3. LLM Layer

The LLM Layer serves as the analytical core of the platform, enabling the system to interpret psychiatrist–patient conversations and generate accurate diagnostic predictions. This layer comprises a consortium of fine-tuned LLMs, each trained on domain-specific conversational datasets involving mental health evaluations [28]. These models are specialized to recognize linguistic and behavioral patterns indicative of various psychiatric conditions, such as depression, anxiety, PTSD, and schizophrenia [41]. The fine-tuned LLMs are deployed and managed using Ollama[42, 43], a lightweight framework optimized for efficient inference and deployment of LLMs on consumer-grade hardware. This ensures the platform can scale effectively and maintain high performance even under resource constraints. As illustrated in Figure 2, the LLM Agent Layer interfaces with the LLM consortium through Ollama’s API, orchestrating prompt generation, model invocation, and response handling. By leveraging multiple specialized models within the consortium, the LLM Layer enhances diagnostic robustness through diversity in model reasoning, ultimately supporting a more comprehensive and consistent assessment of patient mental health based on natural language conversations.

3.4. OpenAI-gpt-oss Reasoning LLM Layer

The OpenAI-gpt-oss Reasoning Layer embodies the platform’s advanced cognitive and decision-making capabilities, leveraging state-of-the-art reasoning language models to synthesize diagnostic insights. The OpenAI-gpt-oss Reasoning LLM acts as the cognitive and synthesis engine of the platform. It is responsible for the high-level reasoning, integration, and refinement of system modeling predictions derived from the LLM consortium.

Within the platform, OpenAI-gpt-oss serves as the final decision-making engine. It receives diagnostic predictions from the consortium of fine-tuned LLMs and performs structured reasoning to evaluate, cross-validate, and refine these outputs [44]. By synthesizing diverse model perspectives, OpenAI-gpt-oss determines the most consistent and clinically aligned diagnostic outcome, ensuring accuracy, coherence, and alignment with DSM-5 diagnostic criteria [36]. The LLM Agent Layer facilitates this process by aggregating the preliminary predictions and formatting them into structured, context-aware prompts tailored for OpenAI-gpt-oss. This enables the reasoning LLM to process heterogeneous inputs and deliver a final, consensus-driven diagnosis. By integrating probabilistic reasoning and consistency checks, the OpenAI-gpt-oss Reasoning Layer plays a pivotal role in enhancing the reliability, transparency, and clinical relevance of AI-assisted psychiatric diagnosis.

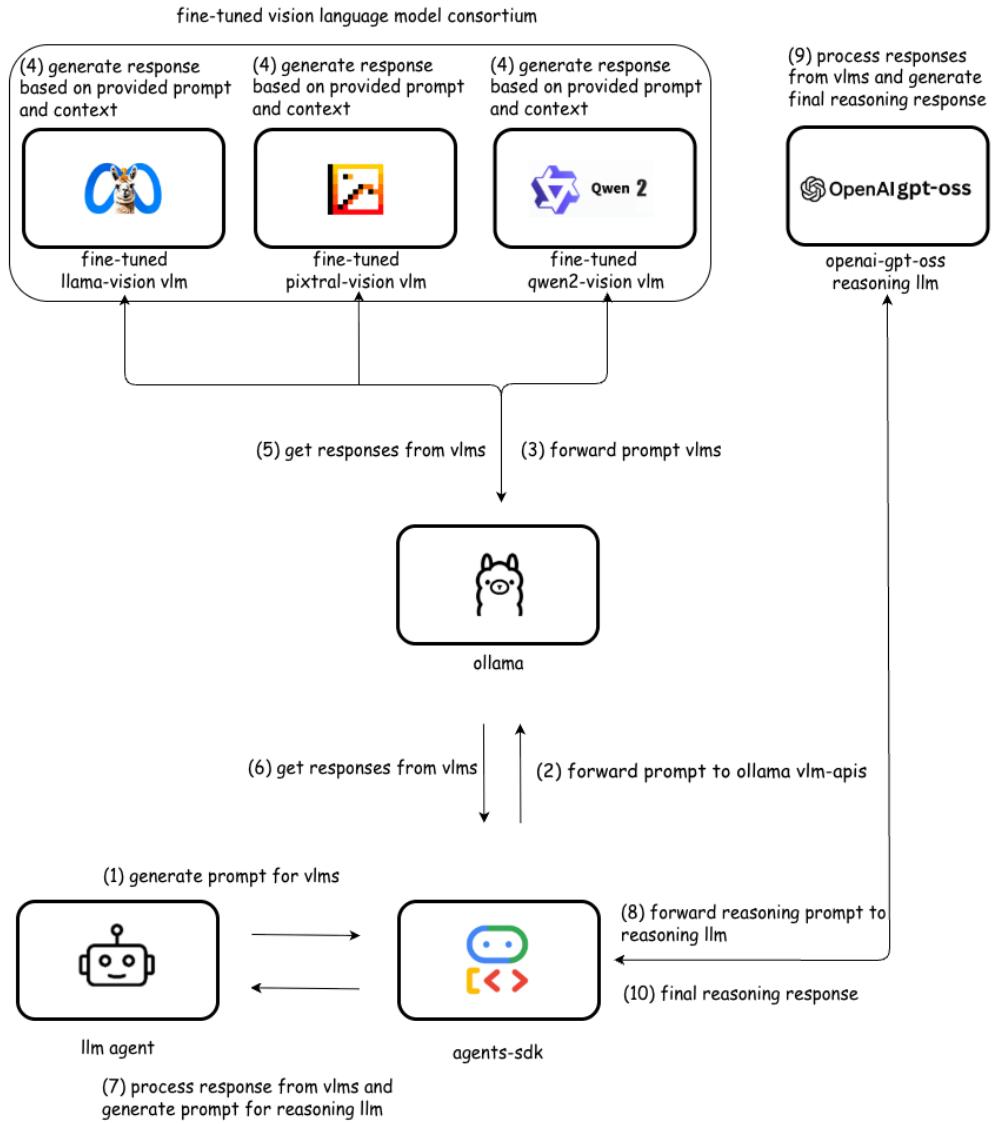


Figure 2: LLM integration flow with Ollama LLM-API

4. Platform Functionality

There are four main functionalities of the platform: 1) Data Lake Setup, 2) LLM Fine-Tuning, 3) Diagnosis prediction of Fine-tuned LLMs, and 4) Final Diagnosis Prediction by OpenAI-gpt-oss reasoning LLM. This section goes into the specifics of these functions.

4.1. Data Lake Setup

The first step in the platform’s workflow involves the setup of the Data Lake, which forms the foundational layer for storing, managing, and accessing large-scale conversational datasets. These datasets simulate or replicate psychiatrist–patient interactions focused on mental health evaluations and serve as the primary training resource for fine-tuning the language models used in diagnosis prediction [45]. The Data Lake primarily contains transcribed conversations between clinicians and patients, along with their corresponding diagnoses provided by licensed psychiatrists. These diagnoses are annotated in accordance with standardized diagnostic frameworks, such as the DSM-5, ensuring clinical relevance and consistency [46]. This centralized repository enables the platform to efficiently organize diverse, high-quality training data necessary for modeling diagnostic reasoning and symptom interpretation.

By providing robust, scalable, and secure data infrastructure, the Data Lake supports the development of fine-tuned LLMs capable of understanding complex psychiatric narratives, detecting subtle symptom patterns, and generating clinically accurate predictions. It is a critical enabler of the platform’s end-to-end AI-driven psychiatric diagnosis capabilities.

4.2. LLM Fine-Tuning

The second step in the platform workflow involves fine-tuning LLMs using the curated and pre-processed data stored in the Data Lake. This stage is crucial for transforming general-purpose models into specialized diagnostic agents capable of interpreting psychiatrist–patient conversations and identifying symptoms of mental disorders. Multiple state-of-the-art models, including Llama-3[22, 7], Mistral[25], and Qwen2[24], are fine-tuned on this domain-specific dataset to adapt them to the unique linguistic and contextual characteristics of psychiatric evaluations. The structure and composition of the dataset used for fine-tuning are illustrated in Figure3.

The fine-tuning process is carried out using the Unsloth library[31], which facilitates efficient large-scale model adaptation. To ensure that models are deployable on consumer-grade hardware without compromising performance, the fine-tuning process incorporates Quantized Low-Rank Adapters (QLoRA) [19] with 4-bit quantization as depicted in Figure 4. This optimization significantly reduces memory and computational requirements, making the models suitable for real-time inference and edge deployment.

Upon completion, the fine-tuned and quantized models are deployed via Ollama [42, 11], a lightweight framework optimized to manage and run LLMs

efficiently. These specialized models form the diagnostic core of the platform, each capable of analyzing psychiatric dialogue and producing preliminary mental disorder predictions based on learned diagnostic patterns and criteria aligned with DSM-5.

conversation string · lengths	reasoning string · classes	diagnosis string · classes
171•176 35%	The patient... 21%	Panic Diso... 21%
Patient: I've been feeling really down for weeks and I...	The patient reports a persistent depressed mood,...	Major Depressive Disorder (DSM-5 296.21)
Patient: I've been feeling really down for weeks and I...	The patient reports a persistent depressed mood,...	Major Depressive Disorder (DSM-5 296.21)
Patient: I suddenly feel like I can't breathe and my heart races. Doctor: Do these happen without warning? Patient: Yes, and I worry about having another attack. (Sample 2)	The patient describes recurrent unexpected panic attacks and persistent concern about having more attacks, consistent with Panic Disorder as per DSM-5.	Panic Disorder (DSM-5 300.01)
Patient: I worry about everything constantly. Even...	The patient exhibits excessive worry lasting over six months,...	Generalized Anxiety Disorder (DSM-5 300.02)
Patient: I suddenly feel like I can't breathe and my heart...	The patient describes recurrent unexpected panic attacks and...	Panic Disorder (DSM-5 300.01)

Figure 3: The format of the dataset used to fine-tune the LLMs.

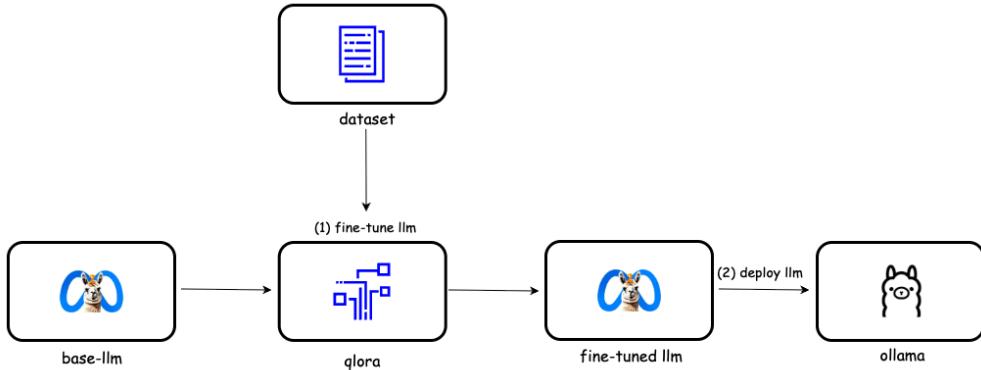


Figure 4: Fine-tune LLMs with Qlora and deploy with Ollama.

4.3. Diagnosis Prediction by Fine-tuned LLMs

Following the fine-tuning process, the next phase of the platform involves generating preliminary diagnoses of mental disorders using the consortium

of fine-tuned LLMs. When new conversational data is received between a psychiatrist and a patient, the platform’s LLM Agents initiate diagnostic analysis by interfacing with fine-tuned models through the Ollama API [42]. To facilitate accurate and context-aware predictions, the LLM Agent employs custom prompt engineering, embedding the relevant conversation data and structured diagnostic context into customized prompts for each model [47]. These prompts are carefully designed to align with the LLM input expectations and to provide adequate clinical representation, such as duration of symptoms, emotional tone, and functional impact, based on the diagnostic criteria of DSM-5. Each fine-tuned model then analyzes the provided input, extracts potential symptoms, and produces its own diagnostic prediction. These individual outputs are collected by the LLM Agent, which organizes them into a structured format for downstream processing. This step ensures that the specialized capabilities of fine-tuned LLM are fully utilized to provide rich, reliable, and interpretable diagnostic insights into potential mental health conditions.

By allowing multiple independent evaluations throughout the model consortium, this layer improves diagnostic diversity, robustness, and the general ability of the platform to generalize in diverse patient presentations.

4.4. Final Diagnosis Prediction by OpenAI-gpt-oss Reasoning LLM

To ensure the highest level of diagnostic accuracy, reliability, and clinical validity, the platform employs a consensus-based decision-making mechanism to generate the final diagnosis. Rather than relying on the output of a single model, the platform aggregates diagnostic predictions from multiple fine-tuned LLMs within the consortium. These individual outputs are then evaluated, compared, and synthesized by OpenAI-gpt-oss, a specialized reasoning LLM designed to perform advanced analytical inference [44, 16]. As a core component of the architecture, OpenAI-gpt-oss acts as an intelligent adjudicator, capable of contextualizing, validating, and refining the predictions provided by the underlying LLMs. Using its advanced reasoning capabilities, OpenAI-gpt-oss identifies the most consistent and clinically appropriate diagnostic result from a diverse set of model-generated insights.

To enable this reasoning process, the LLM Agent constructs custom, structured prompts by embedding and organizing the outputs from the fine-tuned models. These prompts, as illustrated in Figure 5, provide OpenAI-gpt-oss with a unified view of candidate diagnoses, associated symptoms, and contextual cues. The reasoning LLM processes this composite input and

produces a final diagnosis that aligns with the DSM-5 criteria and reflects a well-supported clinical interpretation [36]. This consensus-driven architecture significantly improves the robustness and generalizability of diagnostic predictions by mitigating the limitations of individual models and reducing variability. By orchestrating this process through a transparent, explainable pipeline, the platform not only increases trustworthiness but also establishes a replicable framework for AI-assisted psychiatric evaluation.

The integration of ensemble-based inference with symbolic reasoning marks a transformative shift in mental health diagnostics, offering a scalable and interpretable decision support tool for clinicians. It demonstrates the potential of combining large-scale language understanding with structured reasoning to improve clinical decision-making in complex, subjective domains such as mental health.

```

1 prompt = ChatPromptTemplate(
2     messages=[
3         SystemMessagePromptTemplate.from_template(
4             "You are a clinical AI expert specializing in diagnostic reasoning for psychiatric and mental
5             health disorders. "
6             "You will receive diagnostic predictions generated by multiple fine-tuned large language models
7             (LLMs) based on a conversation between a psychiatrist and a patient. "
8             "Your task is to critically evaluate these model predictions and determine the most clinically
9             accurate and DSM-5-aligned final diagnosis."
10            ),
11            HumanMessagePromptTemplate.from_template(
12                "Here are the diagnosis predictions generated by the LLMs for the following
13                psychiatrist-patient conversation:\n"
14                "- **Mistral model**: {prediction_1}\n"
15                "- **Qwen2 model**: {prediction_2}\n"
16                "- **Llama-3 model**: {prediction_3}\n\n"
17                "Based on these predictions, please analyze the consistency, symptom alignment, and diagnostic
18                criteria, and provide the most accurate and well-reasoned final diagnosis for the patient's
19                mental disorder."
20            )
21        ]
22    )

```

Figure 5: Prompt for OpenAI-gpt-oss reasoning LLM for final prediction reasoning.

5. Implementation and Evaluation

The implementation of the proposed platform was conducted in collaboration with the U.S. Army Medical Research team in Newport News, Virginia, USA. The LLM layer and reasoning Layer comprises of three fine-tuned LLMs, including Llama-3 [7], Mistral [25], and Qwen [24], and OpenAI-gpt-oss [8, 27] reasoning LLM. The LLM Agent Layer was implemented using OpenAI Agents SDK [48] and Google Agent Development Kit [49], enabling

secure orchestration, transparent auditability, and decentralized control of the LLM interactions.

Fine-tuning was conducted using the Unsloth library [31] on Google Colab, leveraging both NVIDIA A100 GPUs and Tesla TPUs [18] to support efficient and scalable training cycles. The original dataset consisted of approximately 2,000 annotated records, each containing a psychiatrist–patient conversation, the doctor’s diagnostic reasoning, and the corresponding final diagnosis. These records were compiled from multiple data sources, as illustrated in Figure 3.

The Unsloth framework requires the input data to be structured in a conversational format [31]. To meet this requirement, the dataset was pre-processed and transformed into the required schema, shown in Figure 6. Each training sample includes the fields: instruction (providing context or a prompt to the LLM), content (representing the main conversational input), and text (containing the model’s expected diagnostic reasoning output). The dataset was partitioned into training, validation, and testing subsets using a 2/3, 1/6, 1/6 split, respectively. The training process was completed in approximately 1,627 seconds (27.12 minutes). Peak memory reservation during training was 14.605 GB, with actual memory utilization reaching 5.853 GB, equivalent to 39.69% of the reserved memory and 99.03% of the peak allocation. These results demonstrate that fine-tuning large language models for psychiatric diagnosis using structured conversational data can be performed efficiently on moderate-scale datasets, even with limited hardware resources. This underscores the practicality and accessibility of applying LLMs in specialized domains such as mental health.

After fine-tuning, the models were quantized using QLoRA [19], a process that enables efficient operation on consumer-grade hardware. This optimization was critical for deploying the fine-tuned models on Ollama, a framework designed for the lightweight yet high-performance execution of LLMs. Based on the predictions of the LLMs, OpenAI-gpt-oss LLM makes the final diagnosis of the mental illness. Custom prompts are used to instruct the OpenAI-gpt-oss Reasoning LLM to understand the context of the prediction. Based on the provided context and the predictions of the LLMs, the model makes the final diagnosis prediction. Platform performance is evaluated in three main areas: 1) Evaluation of LLM fine-tuning, 2) Diagnostic Performance of Fine-Tuned LLM Consortium, and 3) Diagnostic Reasoning Performance of the OpenAI-gpt-oss LLM.

```

1+ {
2   "Question": "Patient: I keep getting flashbacks from the accident. Doctor: Do you avoid reminders of the
   event? Patient: Yes, I avoid driving and loud noises.",
3   "Complex_CoT": "Let's analyze this conversation step by step. The patient reports experiencing flashbacks
   of a traumatic accident. This aligns with the DSM-5 criterion for intrusive symptoms associated with
   Post-Traumatic Stress Disorder (PTSD). The doctor inquires about avoidance behaviors, and the patient
   confirms avoiding both driving and loud noises—these are classic avoidance symptoms. In addition to
   intrusion and avoidance, PTSD also involves persistent arousal and negative alterations in cognition
   or mood. Although not fully elaborated in this snippet, the presence of both flashbacks and avoidance
   is strongly suggestive of PTSD. According to DSM-5, these are key diagnostic features for PTSD under
   code 309.81.",
4   "Response": "Post-Traumatic Stress Disorder (DSM-5 309.81)",
5   "text": "Below is an instruction that describes a task, paired with an input that provides further context
   . \nWrite a response that appropriately completes the request. \nBefore answering, think carefully
   about the question and create a step-by-step chain of thoughts to ensure a logical and accurate
   response.\n\n### Instruction:\nYou are a medical expert with advanced knowledge in psychiatric
   diagnostics and DSM-5 criteria. \nPlease evaluate the following psychiatrist-patient conversation and
   determine the most accurate mental disorder diagnosis.\n\n### Question:\nPatient: I keep getting
   flashbacks from the accident. Doctor: Do you avoid reminders of the event? Patient: Yes, I avoid
   driving and loud noises.\n\n### Response:\n\nLet's analyze this conversation step by step. The
   patient reports experiencing flashbacks of a traumatic accident. This aligns with the DSM-5 criterion
   for intrusive symptoms associated with Post-Traumatic Stress Disorder (PTSD). The doctor inquires
   about avoidance behaviors, and the patient confirms avoiding both driving and loud noises—these are
   classic avoidance symptoms. In addition to intrusion and avoidance, PTSD also involves persistent
   arousal and negative alterations in cognition or mood. Although not fully elaborated in this snippet,
   the presence of both flashbacks and avoidance is strongly suggestive of PTSD. According to DSM-5,
   these are key diagnostic features for PTSD under code 309.81.\n</think>\nPost-Traumatic Stress
   Disorder (DSM-5 309.81)< | end_of_sentence | >"
6 }

```

Figure 6: The required data format of the unsloth library to fine-tune the LLM.

5.1. Evaluation of LLM Fine-Tuning

This evaluation focuses on measuring the effectiveness of the fine-tuning process in improving the diagnostic accuracy of LLMs within the platform. Specifically, we evaluated the performance of the fine-tuned Llama-3 model in its ability to identify psychiatric symptoms and produce accurate diagnoses based on conversational data between psychiatrists and patients [50].

Throughout the fine-tuning process, we continuously monitored critical metrics—specifically, training loss and validation loss—to assess the model’s learning dynamics and generalization ability [28]. As shown in Figure 9, the training loss (Figure 7) and validation loss (Figure 8) both exhibit a steep decline during the initial training steps, indicating rapid learning and effective assimilation of domain-specific patterns. The validation loss continues to decrease smoothly over time, stabilizing around step 25, which suggests improved generalization to unseen samples. Meanwhile, the training loss decreases more aggressively and plateaus slightly earlier, signaling convergence. Figure 9 provides an integrated visualization of both metrics along with the area between the curves, which quantifies the generalization gap.

This shaded area (2.41) highlights the difference between training and validation performance. The relatively narrow and consistently shrinking gap further confirms the model’s ability to generalize well without overfitting. These trends collectively indicate that the fine-tuning process was effective and stable, enabling the LLM to adapt precisely to the psychiatric diagnostic domain while maintaining performance on unseen conversational data.

Figure 10 captures multiple key training dynamics, including the loss difference, loss ratio, and loss derivatives over training steps, offering valuable insights into the model’s convergence behavior and generalization performance. The consistently positive loss difference (validation loss exceeding training loss) suggests signs of overfitting, especially at steps with noticeable spikes. The loss ratio, ranging from 1.0 to 3.0, highlights varying degrees of generalization, where a lower ratio reflects better alignment between training and validation performance. Additionally, the loss derivatives reveal rapid initial improvements followed by smaller, oscillating changes, indicating stabilization or saturation in the learning process [51].

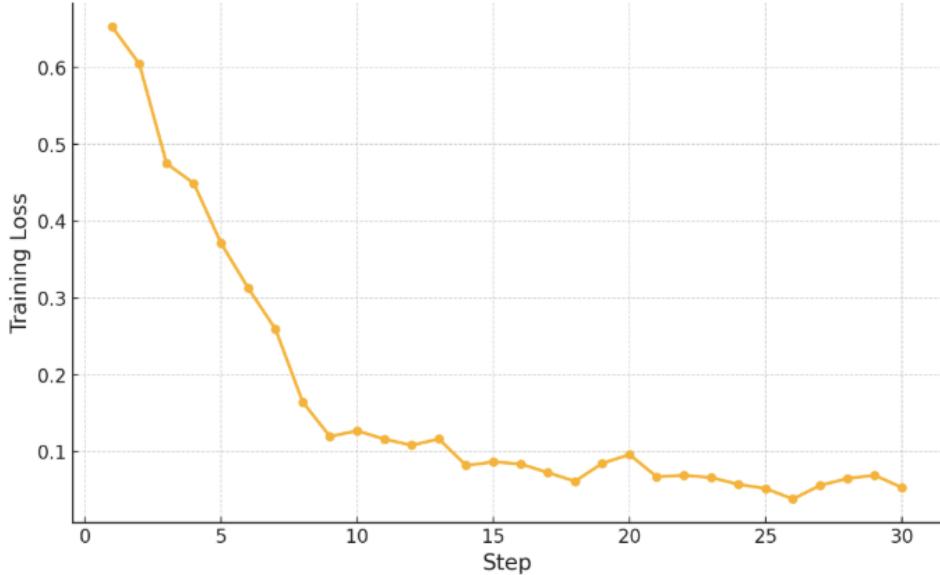


Figure 7: Training loss during fine-tuning of the Llama-3 LLM

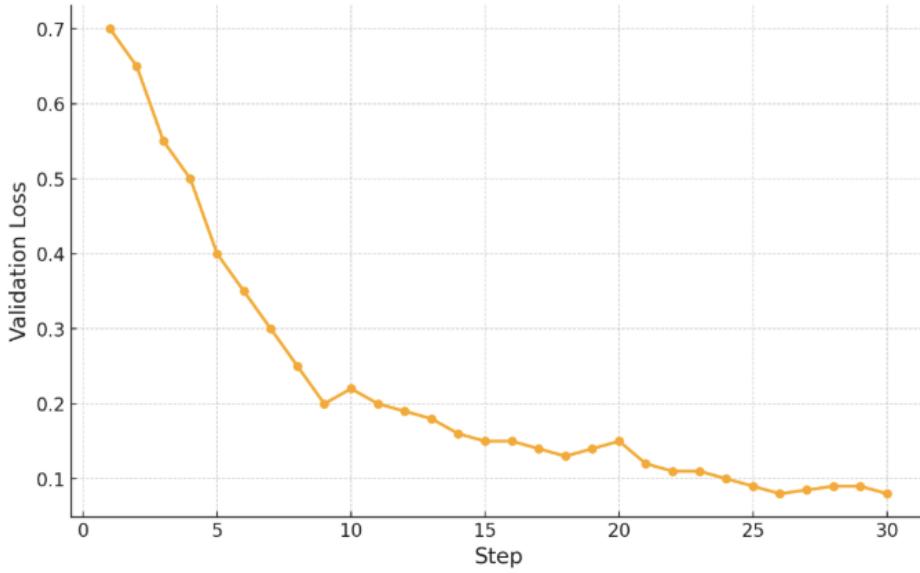


Figure 8: Validation loss during fine-tuning of the Llama-3 LLM

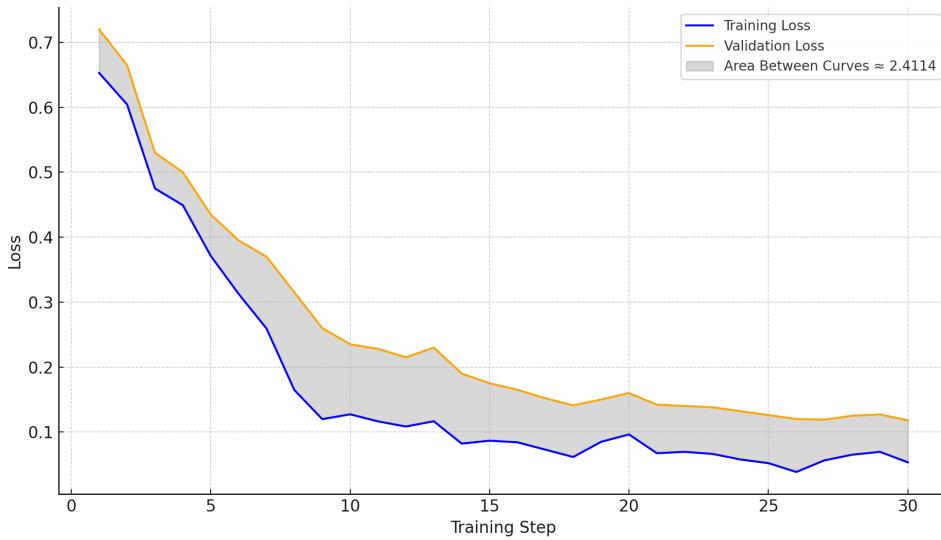


Figure 9: Training vs. Validation Loss and Area Between Curves during Fine-Tuning of the Llama-3 LLM

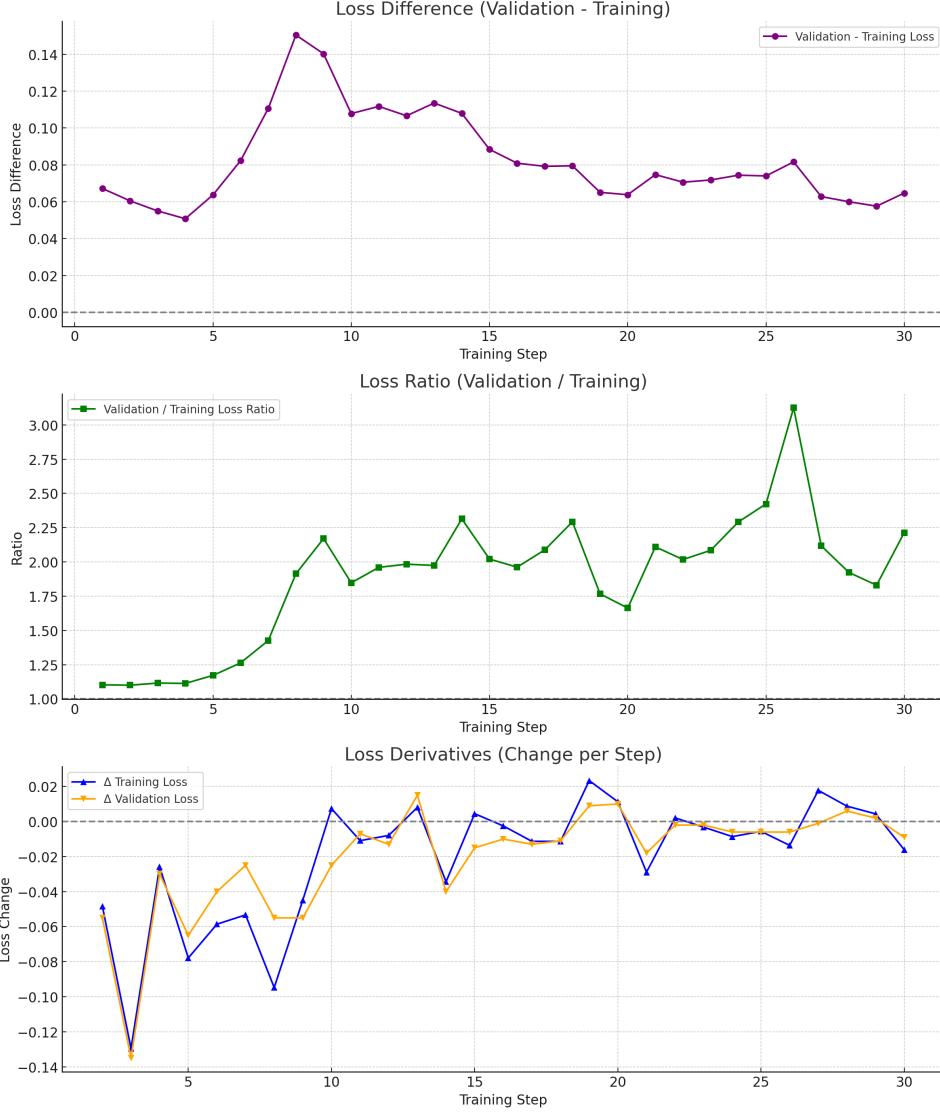


Figure 10: Ratio of training to validation loss during the fine-tuning of the Llama-3 LLM.

5.2. Diagnostic Performance of Fine-Tuned LLM Consortium

Following the training phase, we assessed the predictive performance of the fine-tuned models in the context of psychiatric diagnosis. This evaluation involved comparing real diagnostic annotations—based on psychiatrist–patient conversations—with the predictions generated by both the baseline (pre-trained) LLMs and their fine-tuned counterparts.

Figure 11 and Figure 12 present the diagnostic predictions made by the Llama-3 model [22, 7] before and after fine-tuning for two psychiatric conditions: Major Depressive Disorder (DSM-5 296.21) and Bipolar I Disorder, Current Episode Manic (DSM-5 296.41) [52]. Before fine-tuning, the model generated verbose but loosely structured outputs, relying on high-level symptom descriptions without explicitly mapping them to DSM-5 diagnostic codes [2]. While it identified relevant symptoms (e.g., euphoric mood, insomnia, and loss of appetite), it did not consistently align them with standardized diagnostic criteria. After fine-tuning on a domain-specific dataset, the model demonstrated significantly improved diagnostic precision. It accurately produced concise and clinically valid diagnoses with correct DSM-5 codes, directly inferred from the contextual patient-physician conversations. This highlights the effectiveness of targeted fine-tuning in enhancing clinical reasoning and diagnostic accuracy in mental health applications.

Conversation	Patient: I've been feeling really down for weeks and I don't enjoy anything anymore. Doctor: How is your sleep and appetite? Patient: I can't sleep and I've lost my appetite.
Real Observation	Major Depressive Disorder (DSM-5 296.21)
Llama-3 Prediction (Before Fine-tuning)	<p>**Major Depressive Disorder (MDD)**</p> <p>Criterion A: Five or more symptoms during the same two-week period, at least one of which is (1) depressed mood most of the day, nearly every day; and Criterion B: The symptoms cause clinically significant distress or impairment in social, occupational, or other areas of functioning.</p> <p>Given the patient's reports of feeling "really down" for weeks, lack of interest in activities, insomnia, and loss of appetite, I believe MDD is a strong contender. The patient's symptoms align with those listed in the DSM-5 criteria for MDD:</p>
Llama-3 Prediction (After Fine-tuning)	Major Depressive Disorder - DSM-5 296.21

Figure 11: Diagnostic output of the Llama-3 LLM for symptoms indicative of Major Depressive Disorder (DSM-5 296.21)

Conversation	Patient: I feel on top of the world, barely sleep, and started multiple projects. Doctor: Are your thoughts racing? Patient: Yes, I can't stop talking or thinking.
Real Observation	Bipolar I Disorder, Current Episode Manic (DSM-5 296.41)
Llama-3 Prediction (Before Fine-tuning)	<p>**Patient's symptoms:**</p> <ol style="list-style-type: none"> 1. Feeling "on top of the world" (euphoric mood) 2. Difficulty sleeping (insomnia) 3. Starting multiple projects (hyperactivity) 4. Thoughts racing (rapid thoughts) 5. Can't stop talking or thinking (pressure of speech) <p>Based on these symptoms, I predict that the patient may be experiencing **Mania**, specifically **Bipolar**.</p>
Llama-3 Prediction (After Fine-tuning)	Bipolar I Disorder - DSM-5 296.41

Figure 12: Diagnostic output of the Llama-3 LLM for symptoms consistent with Bipolar I Disorder, Current Episode Manic (DSM-5 296.41)

Figure 14 and Figure 13 illustrate the diagnostic predictions produced by the Mistral model [23] for Panic Disorder and Post-Traumatic Stress Disorder (PTSD) [53]. Before fine-tuning, Mistral was able to recognize relevant clinical symptoms from the conversation and partially align them with DSM-5 criteria. For instance, in the panic disorder case, the model identified key features such as shortness of breath, rapid heart rate, and anticipatory anxiety about future attacks. Similarly, for PTSD, it recognized trauma exposure, intrusive flashbacks, and avoidance behaviors, citing appropriate DSM-5 criteria. However, after fine-tuning, the model demonstrated improved diagnostic specificity and accuracy. It produced concise and direct classifications that fully matched the DSM-5 diagnostic codes (e.g., DSM-5 300.01 for Panic Disorder and DSM-5 309.81 for PTSD) [36]. This confirms that fine-tuning enhanced Mistral’s capacity for structured clinical reasoning and reliable psychiatric diagnosis from patient-doctor dialogues.

Conversation	Patient: I suddenly feel like I can't breathe and my heart races. Doctor: Do these happen without warning? Patient: Yes, and I worry about having another attack.
Real Observation	Panic Disorder (DSM-5 300.01)
Mistral Prediction (Before Fine-tuning)	<p>The symptoms mentioned in the conversation are:</p> <ol style="list-style-type: none"> 1. Sudden feelings of not being able to breathe (shortness of breath) 2. Rapid heart rate 3. Attacks occurring without warning 4. Patient worrying about having another attack <p>These symptoms align with the diagnostic criteria for Panic Disorder, as specified in the DSM-5:</p>
Mistral Prediction (After Fine-tuning)	Panic Disorder - DSM-5 300.01

Figure 13: Diagnostic output of the Mistral LLM for symptoms consistent with Panic Disorder (DSM-5 300.01)

Conversation	Patient: I keep getting flashbacks from the accident. Doctor: Do you avoid reminders of the event? Patient: Yes, I avoid driving and loud noises. (Sample 19)
Real Observation	Post-Traumatic Stress Disorder (DSM-5 309.81)
Mistral Prediction (Before Fine-tuning)	<p>Based on the provided conversation, the patient's symptoms of recurring flashbacks related to a traumatic event and avoidance behavior (avoiding reminders of the event and driving) could suggest Post-Traumatic Stress Disorder (PTSD).</p> <p>The patient meets criteria 1 (exposure to a traumatic event) and 2 (persistent re-experiencing of the event through flashbacks), as well as criterion 4 (alterations in arousal and reactivity) based on their avoidance behavior. However, it is essential to remember that a definitive diagnosis requires a comprehensive evaluation by a mental health professional. The provided conversation should be considered only as a preliminary indication of potential PTSD.</p>
Mistral Prediction (After Fine-tuning)	Post-Traumatic Stress Disorder - DSM-5 309.81

Figure 14: Diagnostic output of the Mistral LLM for symptoms consistent with Post-Traumatic Stress Disorder (DSM-5 309.81)

Figure 16 and Figure 15 present the diagnostic outputs generated by the Qwen2 model [54] for Schizophrenia and Generalized Anxiety Disorder, respectively [55]. Prior to fine-tuning, the model identified relevant symptoms and attempted to match them to DSM-5 criteria. However, the predictions were often verbose, interpretative, and occasionally lacked clinical precision. After fine-tuning on the psychiatric dataset, the Qwen2 model accurately mapped symptom clusters to their corresponding DSM-5 diagnoses

with greater clarity and conciseness, yielding outputs that are more aligned with psychiatric clinical standards.

Conversation	Patient: I hear voices that others can't hear. Doctor: Do you avoid certain places due to fear? Patient: Yes, I often avoid public spaces because I feel people are watching or trying to harm me.
Real Observation	Schizophrenia (DSM-5 295.90)
Qwen-2 Prediction (Before Fine-tuning)	**Patient's symptoms:** 1. Hearing voices that others can't hear (auditory hallucinations) 2. Avoiding certain places due to fear and perceived threat (avoidant behavior) * Hearing voices that others can't hear is a classic symptom of auditory hallucinations, which is characteristic of Schizophrenia. * Avoiding public spaces due to perceived threat and fear suggests delusions (fixed, false beliefs) and paranoia, which are also common in Schizophrenia.
Qwen-2 Prediction (After Fine-tuning)	Schizophrenia - DSM-5 295.90

Figure 15: Diagnostic output of the Qwen2 LLM for symptoms consistent with Schizophrenia (DSM-5 295.90)

Conversation	Patient: I worry about everything constantly. Even small things stress me out. Doctor: How long has this been going on? Patient: Over six months. I can't focus and feel tense all the time.
Real Observation	Generalized Anxiety Disorder (DSM-5 300.02)
Qwen-2 Prediction (Before Fine-tuning)	The symptoms described by the patient—constant worry, stress over small things, difficulty focusing, and tension—are consistent with Generalized Anxiety Disorder (GAD). According to the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), GAD is characterized by excessive anxiety and worry about multiple events or activities that occur more days than not for at least six months. The individual finds it difficult to control this worry, and they may experience symptoms such as restlessness, fatigue, difficulty concentrating, irritability, muscle tension, and sleep disturbances. To meet the criteria for GAD, an individual must display excessive anxiety or worry that is not proportionate to the actual threat posed by the situation, and these symptoms must cause significant distress or impairment in social, occupational, or other important areas of functioning (American Psychiatric Association, 2013). The patient's report of ongoing symptoms for over six months, as well as their difficulty focusing and feeling tense all the time, suggest that they may meet the criteria for GAD. It is essential to note that a professional diagnosis can only be made after a thorough evaluation by a mental health provider.
Qwen-2 Prediction (After Fine-tuning)	Generalized Anxiety Disorder

Figure 16: Diagnostic output of the Qwen2 LLM for symptoms indicative of Generalized Anxiety Disorder (DSM-5 300.02)

These results demonstrate that the fine-tuned models consistently produce predictions that closely align with clinically validated diagnoses, show-

ing improved precision, consistency, and interpretability. Compared to their baseline versions, the fine-tuned LLMs exhibit a substantial increase in diagnostic accuracy, underscoring the effectiveness of task-specific fine-tuning in mental health applications. These findings validate the utility of LLMs as decision-support tools in AI-assisted psychiatric diagnosis.

5.3. Diagnostic Reasoning Performance of the OpenAI-gpt-oss LLM

In this evaluation, we assessed the diagnostic reasoning performance of the OpenAI-gpt-oss reasoning LLM by comparing its final diagnosis predictions with those generated independently by multiple fine-tuned LLMs. The goal was to evaluate OpenAI-gpt-oss’s ability to synthesize diverse diagnostic outputs and determine the most accurate and clinically appropriate outcome. Figure 17 illustrates a comparative analysis of the diagnostic predictions from the Llama-3, Mistral, and Qwen2 models, alongside the final output produced by OpenAI-gpt-oss. The figure highlights the reasoning model’s ability to interpret and reconcile varying predictions, applying structured clinical logic to arrive at a DSM-5-aligned diagnosis. The results demonstrate that OpenAI-gpt-oss significantly enhances diagnostic reliability by evaluating the coherence and clinical relevance of the LLM consortium’s outputs. This consensus-driven reasoning step adds an important layer of interpretability and robustness, reinforcing the value of integrating a dedicated reasoning LLM within the framework for AI-assisted psychiatric diagnosis.

Conversation	Patient: I keep getting flashbacks from the accident. Doctor: Do you avoid reminders of the event? Patient: Yes, I avoid driving and loud noises. (Sample 19)
Real Observation	Post-Traumatic Stress Disorder (DSM-5 309.81)
Prediction of Fine-tuned Llama-3 LLM	Post-Traumatic Stress Disorder
Prediction of Fine-tuned Mistral LLM	Post-Traumatic Stress Disorder (DSM-5 309.8)
Prediction of Fine-tuned Qwen-2 LLM	Unknown
Reasoning of OpenAI-o3 LLM	<p>Okay, let's tackle this mental disorder diagnosis problem. The user provided a doctor-patient conversation and predictions from three different LLMs. The task is to come up with a final diagnosis based on these predictions.</p> <p>First, I need to understand the conversation. The patient mentions having flashbacks from an accident and avoiding driving and loud noises. The doctor is asking about avoidance behaviors. The key symptoms here are flashbacks (intrusive memories) and avoidance of reminders. These are classic signs of PTSD.</p> <p>Looking at the model predictions: Llama-3 and Mistral both say PTSD, with Mistral even adding the DSM-5 code. Qwen says "Unknown." Why would Qwen say that? Maybe the model is more cautious or requires more symptoms to make a diagnosis. Let me recall the DSM-5 criteria for PTSD. The main criteria are exposure to trauma, intrusive symptoms (like flashbacks), avoidance, negative alterations in cognition/mood, and arousal/reactivity symptoms. The patient here has exposure (accident), intrusive memories, and avoidance. So they meet at least three criteria. But maybe Qwen is looking for more symptoms, like hypervigilance or sleep issues, which aren't mentioned here.</p> <p>But the two other models are confident. Llama-3 and Mistral's agreement strengthens the PTSD diagnosis. Qwen's uncertainty might be due to the absence of other symptoms in the conversation. However, according to DSM-5, you don't need all symptoms, just a subset from each category. The patient here has the necessary symptoms for a diagnosis.</p> <p>I should also check if there's any other possible disorder. For example, acute stress disorder is similar but occurs within a month of the trauma. Since the conversation doesn't specify the timeframe, but PTSD typically lasts longer. Also, the patient's avoidance and flashbacks are more aligned with PTSD.</p> <p>The final diagnosis should be PTSD, given the majority agreement and symptom alignment. Qwen's uncertainty might suggest the need for more information, but based on what's present, PTSD fits best.</p>

Figure 17: Final diagnostic reasoning produced by the OpenAI-gpt-oss LLM through multi-model consensus.

6. Related Work

Recent advances have seen LLMs and multimodal AI systems increasingly applied across a wide range of healthcare domains. These efforts collectively highlight the expanding capabilities of LLMs in clinical reasoning, decision support, and medical knowledge retrieval. Despite these advancements, most prior work remains focused on general medicine or physical health domains, with limited specialization in psychiatric evaluation or adherence to standardized diagnostic frameworks such as the DSM-5. Furthermore, while in-

dividual models often perform well in isolated tasks, they rarely incorporate structured reasoning or multi-model coordination—critical components for nuanced psychiatric assessments. Table 1 presents a comparative analysis of these prior systems in relation to our proposed framework. In the subsections that follow, we provide a detailed overview of each relevant system’s architecture, capabilities, and relevance to psychiatric diagnostic support.

6.1. *Med-PaLM*

The Med-PaLM family of models, including Med-PaLM and Med-PaLM-2, developed by Google Research, represents some of the most advanced and widely evaluated LLMs for clinical and biomedical tasks [56]. These models are fine-tuned versions of the PaLM (Pathways Language Model) foundation model, tailored to the medical domain through instruction tuning using a combination of publicly available medical datasets, expert-annotated clinical examples, and proprietary benchmarks.

Med-PaLM-2 demonstrated expert-level competency, achieving over 85% accuracy on U.S. Medical Licensing Examination (USMLE)-style multiple-choice questions, including datasets such as MedQA, MedMCQA, and PubMedQA. Notably, in head-to-head evaluations, physicians preferred Med-PaLM-2’s long-form answers to those written by human experts across multiple dimensions, including factual correctness, comprehensiveness, and safety. The model also performed well on several open-ended generation tasks, such as consumer health question answering and summarization of clinical information.

In addition to the text-only versions, a multimodal extension called Med-PaLM-M was developed, enabling the model to jointly reason over textual and visual inputs. This includes radiology images, dermatology photos, pathology slides, and ophthalmic scans. Evaluated on the MultiMedBench benchmark, Med-PaLM-M demonstrated state-of-the-art performance across 14 diverse tasks encompassing classification, retrieval, and open-ended question answering with multimodal inputs. The model is capable of interpreting complex visual patterns and aligning them with clinical narratives, allowing for integrated diagnostic reasoning that mirrors specialist workflows.

Furthermore, the Med-PaLM series introduces an ethical framework and rigorous safety assessments for AI in healthcare. This includes human evaluation protocols to assess bias, hallucination risks, and alignment with clinical guidelines. Despite its strengths, limitations remain—such as interpretability challenges and dependence on high-quality supervised data—underscoring

the need for hybrid frameworks that can further ensure trust and safety in clinical applications.

6.2. LLM for Differential Diagnosis (DDx)

The “LLM for DDx” framework [57] investigates the application of LLMs to the task of generating differential diagnoses from clinical vignettes, including both structured patient data and free-text clinical notes. By leveraging prompt engineering techniques, the system guides the LLM to output ranked diagnostic hypotheses based on symptomatology, patient history, and clinical context. This framework adopts few-shot or zero-shot prompting, often augmented with chain-of-thought reasoning or self-consistency sampling, to simulate a clinician’s diagnostic reasoning. The model is evaluated on publicly available datasets such as MedQA, MIMIC-III case notes, and Clinical Case Challenge benchmarks. It demonstrates promising capabilities in recognizing comorbidities, parsing temporal disease progression, and identifying rare or underrepresented conditions—tasks that often challenge traditional rule-based decision support systems. One of the key strengths of this approach lies in its adaptability: the model can generalize across specialties (e.g., internal medicine, neurology, pediatrics) without needing task-specific retraining. However, the system does not integrate multimodal data sources (e.g., lab results, imaging), and lacks structured inference mechanisms like probabilistic reasoning or iterative refinement via specialized LLMs. As such, its diagnostic accuracy can be sensitive to prompt phrasing and input variability, necessitating human-in-the-loop oversight for safe clinical deployment. Despite these limitations, the work highlights the potential of LLMs to augment differential diagnosis in resource-constrained or high-ambiguity settings, offering a foundation for future systems that incorporate more structured reasoning pipelines and multimodal data fusion.

6.3. Me-Llama

Me-Llama [37] adapts the Llama architecture for biomedical applications through a two-phase process: (1) continual pretraining on large-scale biomedical corpora—including PubMed articles, clinical guidelines, and de-identified electronic health records (EHRs); and (2) instruction tuning using curated clinical dialogues and task-specific prompts aligned with workflows typical in patient care. The model is optimized for downstream tasks such as clinical note summarization, medical question answering, and evidence extraction from unstructured clinical documents. Benchmarks show that Me-Llama

outperforms general-purpose LLMs on several biomedical NLP tasks, including the BioASQ challenge, PubMedQA, and MedNLI. Despite being limited to text-only inputs (i.e., it does not incorporate imaging, waveform, or wearable sensor data), Me-Llama exhibits a strong semantic understanding of biomedical terminology, abbreviations, and guideline-referenced medical reasoning. Its lightweight fine-tuning also makes it suitable for deployment in edge healthcare systems, such as point-of-care mobile applications and hospital EMR-integrated tools. Overall, Me-Llama demonstrates the feasibility and performance benefits of domain-specific LLM adaptation within clinical language environments.

6.4. DrHouse

DrHouse [45] introduces an advanced virtual provider assistant system that integrates LLM-based diagnostic reasoning with real-time physiological data collected from consumer-grade wearable sensors (e.g., smart watches, sleep trackers, fitness monitors). The system continuously monitors health signals such as heart rate variability, blood oxygen saturation (SpO_2), sleep quality, step count, and circadian rhythm alignment, enabling longitudinal health tracking and context-aware decision-making. In contrast to static, prompt-only systems, DrHouse engages users through multi-turn, adaptive dialogues that emulate the back-and-forth of a clinical consultation. The LLM component dynamically adjusts its diagnostic hypotheses using an iterative concurrent reasoning framework—recalculating disease likelihoods as new data or clarifications are introduced. Additionally, DrHouse retrieves up-to-date medical knowledge from external expert databases such as UpToDate, PubMed abstracts, and clinical guidelines, thereby ensuring that recommendations remain grounded in the latest medical evidence.

The system employs a dual-loop architecture that balances real-time inference with long-term patient modeling. It maintains an evolving profile of each user’s medical baseline and incorporates Bayesian-style updates for probabilistic diagnosis refinement. DrHouse was evaluated on both public benchmark datasets (e.g., MedQA, Symptoma) and proprietary longitudinal datasets derived from wearable telemetry and self-reported health surveys. Quantitatively, the model achieved up to an 18.8% improvement in diagnostic accuracy compared to baseline LLMs without sensor fusion. Qualitatively, user studies showed that 91.7% of patients found the interaction intuitive and trustworthy, while 75% of clinicians expressed confidence in its ability to support primary care triage. These results underscore the promise of

multimodal, sensor-aware agentic systems for scalable, AI-assisted frontline healthcare.

6.5. CDSS

CDSS [12] introduces a novel Clinical Decision Support System tailored for mental health diagnostics by synergistically combining LLMs with constraint logic programming (CLP). The system ingests natural language diagnostic manuals—specifically DSM-5-TR and ICD-11 CDDR—and uses an LLM to transform each diagnostic criterion into logic rules (e.g., Datalog clauses) [12]. These candidate rules are then vetted and refined by domain experts to ensure clinical fidelity before being executed by an off-the-shelf CLP engine to derive patient-specific diagnoses based on structured patient data.

In empirical evaluations, the hybrid CDSS is compared against two baselines: an “LLM-only” approach that directly generates diagnostic outputs, and an intermediate LLM-to-CLP pipeline without expert oversight. Results indicate that only the expert-validated pipeline consistently produces diagnoses aligned with official manuals, highlighting the necessity of human-in-the-loop rule verification to prevent hallucinations and maintain interpretability. The authors also emphasize operational benefits: the logic rules are transparent and inspectable, facilitating clinician trust and auditability. Moreover, the approach addresses critical ethical concerns by avoiding the direct ingestion of sensitive patient data into the LLM—patient records are instead processed via the CLP engine, mitigating privacy and safety risks associated with raw LLM consumption.

This work represents a significant advancement in mental health AI, as it operationalizes a structured, interpretable, and demonstrably safe CDSS anchored in expert-validated logic and modular LLM capabilities—marking a concrete step toward real-world psychiatric diagnostic tools.

6.6. Weda-GPT

Weda-GPT [34] is a culturally-informed clinical decision-support system that leverages fine-tuned Llama-3 models to provide diagnostic assistance and therapeutic recommendations within indigenous and traditional medicine frameworks. Designed specifically for use in the Indonesian archipelago, Weda-GPT incorporates linguistic, cultural, and epistemological knowledge derived from indigenous medical texts, oral traditions, and local practitioner expertise.

The system is built using a multi-stage fine-tuning pipeline: starting from a Llama-3 base model, it is further adapted with region-specific datasets encompassing herbal pharmacology, traditional syndromic classifications, and culturally embedded health beliefs. Special emphasis is placed on aligning model outputs with culturally appropriate terminology and explanatory models, enabling the system to provide contextually sensitive health advice.

Weda-GPT has been evaluated through case-based testing and participatory design sessions involving local healers and community health workers. The results show that the model effectively maps patient symptoms to culturally relevant diagnoses and treatments, including herbal prescriptions and ritual-based healing practices. Moreover, users reported a high degree of trust and interpretability in the system’s responses, in part due to its capacity to explain recommendations in locally meaningful terms.

Although Weda-GPT does not operate within Western psychiatric diagnostic categories such as DSM-5, it highlights the broader potential of LLMs to support non-Western health systems and pluralistic medical epistemologies. Its design underscores the importance of cultural adaptation and domain-specific alignment when deploying AI in diverse global health contexts. As such, Weda-GPT serves as a complementary model to Western-centric clinical decision systems, demonstrating the scalability and flexibility of LLM-based healthcare tools across sociocultural boundaries.

Table 1: LLM-based Medical Diagnosis Framework Comparison

Platform	Domain	Fine-tuning Support	Running LLM	Vision LM Support	Reasoning LLM Support	LLM Consortium Support
Psychiatric-Diagnoses	Psychiatric	✓	Llama-3, Mistral, Qwen-2	✗	✓	✓
Med-PaLM [56]	General medicine	✓	PaLM	✗	✗	✗
LLM for DDX [57]	General medicine	✓	Not specified	✗	✓	✗
Me-LLaMA [37]	General medicine	✓	Llama	✗	✗	✗
CDSS [12]	Mental Health	✗	GPT-4	✗	✗	✗
DrHouse [45]	General medicine	✓	Not specified	✓	✓	✗
Weda-GPT [34]	Indigenous Medicine	✓	Llama-3	✗	✗	✗

7. Conclusions and Future Work

In this paper, we present an AI-assisted diagnostic framework that integrates a consortium of fine-tuned LLMs with a reasoning LLM (OpenAI-gpt-oss) to improve the accuracy, consistency, and transparency of psychiatric diagnosis while upholding Responsible AI principles. Recognizing the inherent

subjectivity and variability in traditional mental health assessments—often based on unstructured clinical interviews—we proposed a novel architecture that leverages conversational data, custom prompt engineering, and multi-model consensus reasoning to replicate and improve upon clinical diagnostic workflows. The platform is structured into four key layers: the Data Lake Layer for managing annotated psychiatrist–patient dialogues; the fine-tuned LLM Layer for training models specialized in symptom analysis; the LLM Agent Layer for orchestrating model interactions and prompt generation; and the OpenAI-gpt-oss Reasoning Layer, which synthesizes model outputs into a final, reliable diagnosis aligned with DSM-5 criteria. Our approach demonstrates that AI systems, when trained and orchestrated properly, can support mental health professionals by offering data-driven insights and reducing diagnostic variability. The use of low-rank adapters and quantization techniques further enables efficient deployment on consumer-grade hardware, making the system accessible in real-world clinical and remote care settings. To the best of our knowledge, this research represents the first end-to-end integration of fine-tuned large language models (LLMs) with a reasoning engine to standardize psychiatric diagnoses. It lays the foundation for future advancements in AI-assisted eHealth systems, where intelligent agents can augment clinical decision-making while preserving interpretability and ethical responsibility. Future work will focus on clinical validation, multilingual adaptation, and integration with multimodal inputs—such as voice, facial expressions, and affective signals—to enhance diagnostic depth, contextual understanding, and empathy.

Acknowledgements

This work was supported in part by the DoD Center of Excellence in AI and Machine Learning (CoE-AIML) under Contract Number W911NF-20-2-0277 with the U.S. Army Research Laboratory.

References

- [1] C. Yin, F. Li, S. Zhang, Z. Wang, J. Shao, P. Li, J. Chen, and X. Jiang, “Mdd-5k: A new diagnostic conversation dataset for mental disorders synthesized via neuro-symbolic llm agents,” *arXiv preprint arXiv:2408.12142*, 2024.

- [2] D. A. Regier, E. A. Kuhl, and D. J. Kupfer, “The dsm-5: Classification and criteria changes,” *World psychiatry*, vol. 12, no. 2, pp. 92–98, 2013.
- [3] M. Rahsepar Meadi, T. Sillekens, S. Metselaar, A. van Balkom, J. Bernstein, and N. Batelaan, “Exploring the ethical challenges of conversational ai in mental health care: Scoping review,” *JMIR Mental Health*, vol. 12, p. e60432, 2025.
- [4] K. A. Smith, C. Bleas, M. Faurholt-Jepsen, J. Firth, T. Van Daele, C. Moreno, P. Carlbring, U. W. Ebner-Priemer, N. Koutsouleris, H. Riper *et al.*, “Digital mental health: challenges and next steps,” *BMJ Ment Health*, vol. 26, no. 1, 2023.
- [5] S. Bracko and A. Čelofiga, ““it’s a psychiatric patient”: Misdiagnosing of somatic symptoms in patients with mental disorders due to stigma and inadequate diagnostic treatment,” *Archives of Psychiatry Research: An International Journal of Psychiatry and Related Sciences*, vol. 60, no. 1., pp. 62–66, 2024.
- [6] S. Arora, B. Yang, S. Eyuboglu, A. Narayan, A. Hojel, I. Trummer, and C. Ré, “Language models enable simple systems for generating structured views of heterogeneous data lakes,” *arXiv preprint arXiv:2304.09433*, 2023.
- [7] M. Becattini, R. Verdecchia, and E. Vicario, “Sallma: A software architecture for llm-based multi-agent systems.”
- [8] Y. Zhang, S. Mao, T. Ge, X. Wang, A. de Wynter, Y. Xia, W. Wu, T. Song, M. Lan, and F. Wei, “Llm as a mastermind: A survey of strategic reasoning with large language models,” *arXiv preprint arXiv:2404.01230*, 2024.
- [9] S. Agarwal, L. Ahmad, J. Ai, S. Altman, A. Applebaum, E. Arbus, R. K. Arora, Y. Bai, B. Baker, H. Bao *et al.*, “gpt-oss-120b & gpt-oss-20b model card,” *arXiv preprint arXiv:2508.10925*, 2025.
- [10] R. Gore, E. Bandara, S. Shetty, A. E. Musto, P. Rana, A. Valencia-Romero, C. Rhea, L. Tayebi, H. Richter, A. Yarlagadda *et al.*, “Proof-of-tbi-fine-tuned vision language model consortium and openai-o3 reasoning llm-based medical diagnosis support system for mild traumatic brain injury (tbi) prediction,” *arXiv preprint arXiv:2504.18671*, 2025.

- [11] E. Bandara, A. Hass, S. Shetty, R. Mukkamala, R. Gore, A. Rahman, and S. H. Bouk, “Deep-stride: Automated security threat modeling with vision-language models,” in *2025 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, 2025, pp. 1–7.
- [12] B. H. Kim and C. Wang, “Large language models for interpretable mental health diagnosis,” *arXiv preprint arXiv:2501.07653*, 2025.
- [13] K. D. Kannan, S. K. Jagatheesaperumal, R. N. Kandala, M. Lotfaliany, R. Alizadehsanid, and M. Mohebbi, “Advancements in machine learning and deep learning for early detection and management of mental health disorder,” *arXiv preprint arXiv:2412.06147*, 2024.
- [14] X. Huang, W. Liu, X. Chen, X. Wang, H. Wang, D. Lian, Y. Wang, R. Tang, and E. Chen, “Understanding the planning of llm agents: A survey,” *arXiv preprint arXiv:2402.02716*, 2024.
- [15] D. B. Acharya, K. Kuppan, and B. Divya, “Agentic ai: Autonomous intelligence for complex goals—a comprehensive survey,” *IEEE Access*, 2025.
- [16] E. Bandara, R. Gore, S. Shetty, R. Mukkamala, C. Rhea, A. Yarlagadda, S. Kaushik, L. De Silva, A. Maznychenko, I. Sokolowska *et al.*, “Standardization of neuromuscular reflex analysis—role of fine-tuned vision-language model consortium and openai gpt-oss reasoning llm enabled decision support system,” *arXiv preprint arXiv:2508.12473*, 2025.
- [17] R. E. Sistad, R. Kimerling, P. P. Schnurr, and M. J. Bovin, “The impact of screening positive for hazardous alcohol use on the diagnostic accuracy of the ptsd checklist for dsm-5 among veterans,” *Journal of Traumatic Stress*, vol. 37, no. 2, pp. 328–336, 2024.
- [18] H. Kimm, I. Paik, and H. Kimm, “Performance comparision of tpu, gpu, cpu on google colaboratory over distributed deep learning,” in *2021 IEEE 14th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC)*. IEEE, 2021, pp. 312–319.
- [19] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, “Qlora: Efficient finetuning of quantized llms,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

- [20] X. Zhang, B. Yu, H. Yu, Y. Lv, T. Liu, F. Huang, H. Xu, and Y. Li, “Wider and deeper llm networks are fairer llm evaluators,” *arXiv preprint arXiv:2308.01862*, 2023.
- [21] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [22] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, “The llama 3 herd of models,” *arXiv preprint arXiv:2407.21783*, 2024.
- [23] B. Wang, S. Wang, and Q. Ouyang, “Probabilistic inference layer integration in mistral llm for accurate information retrieval,” 2024.
- [24] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, “Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution,” *arXiv preprint arXiv:2409.12191*, 2024.
- [25] H. Samo, K. Ali, M. Memon, F. A. Abbasi, M. Y. Koondhar, and K. Dahri, “Fine-tuning mistral 7b large language model for python query response and code generation: A parameter efficient approach,” *VAWKUM Transactions on Computer Sciences*, vol. 12, no. 1, pp. 205–217, 2024.
- [26] E. Bandara, S. H. Bouk, S. Shetty, R. Gore, S. Kompella, R. Mukkamala, A. Rahman, P. Foytik, X. Liang, N. W. Keong, and K. De Zoysa, “Bassallama — fine-tuned meta’s llama llm, blockchain and nft enabled real-time network attack detection platform for wind energy power plants,” in *2025 International Wireless Communications and Mobile Computing (IWCMC)*, 2025, pp. 330–336.
- [27] G. Mondillo, M. Masino, S. Colosimo, A. Perrotta, and V. Frattolillo, “Evaluating ai reasoning models in pediatric medicine: A comparative analysis of o3-mini and o3-mini-high,” *medRxiv*, pp. 2025–02, 2025.
- [28] X. Lin, W. Wang, Y. Li, S. Yang, F. Feng, Y. Wei, and T.-S. Chua, “Data-efficient fine-tuning for llm-based recommendation,” in *Proceed-*

ings of the 47th international ACM SIGIR conference on research and development in information retrieval, 2024, pp. 365–374.

- [29] E. Bandara, S. H. Bouk, S. Shetty, R. Gore, S. Kompella, R. Mukkamala, A. Rahman, P. Foytik, X. Liang, N. W. Keong, and K. De Zoysa, “Vindsec-llama — fine-tuned meta’s llama-3 llm, federated learning, blockchain and pbom-enabled data security architecture for wind energy data platforms,” in *2025 International Wireless Communications and Mobile Computing (IWCMC)*, 2025, pp. 120–126.
- [30] A. Augustin, J. Yi, T. Clausen, and W. Townsley, “A study of lora: Long range & low power networks for the internet of things,” *Sensors*, vol. 16, no. 9, p. 1466, 2016.
- [31] Y. Zheng, R. Zhang, J. Zhang, Y. Ye, Z. Luo, Z. Feng, and Y. Ma, “Llamafactory: Unified efficient fine-tuning of 100+ language models,” *arXiv preprint arXiv:2403.13372*, 2024.
- [32] C. Liao, M. Sun, Z. Yang, J. Xie, K. Chen, B. Yuan, F. Wu, and Z. Wang, “Lohan: Low-cost high-performance framework to fine-tune 100b model on a consumer gpu,” *arXiv preprint arXiv:2403.06504*, 2024.
- [33] E. Bandara, P. Foytik, S. Shetty, R. Mukkamala, A. Rahman, X. Liang, N. W. Keong, and K. D. Zoysa, “Slicegpt – openai gpt-3.5 llm, blockchain and non-fungible token enabled intelligent 5g/6g network slice broker and marketplace,” in *2024 IEEE 21st Consumer Communications & Networking Conference (CCNC)*, 2024, pp. 439–445.
- [34] E. Bandara, P. Foytik, S. Shetty, R. Mukkamala, A. Rahman, X. Liang, N. W. Keong, and K. De Zoysa, “Wedagpt—generative-ai (with custom-trained meta’s llama2 llm), blockchain, self sovereign identity, nft and model card enabled indigenous medicine platform,” in *2024 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2024, pp. 1–6.
- [35] C. Yin, F. Li, S. Zhang, Z. Wang, J. Shao, P. Li, J. Chen, and X. Jiang, “Mdd-5k: A new diagnostic conversation dataset for mental disorders synthesized via neuro-symbolic llm agents,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 24, 2025, pp. 25715–25723.

- [36] M. S. Scheeringa, “Is factor analysis useful for revising diagnostic criteria for ptsd? a systematic review of five issues ten years after dsm-5,” *Journal of Psychiatric Research*, 2024.
- [37] Q. Xie, Q. Chen, A. Chen, C. Peng, Y. Hu, F. Lin, X. Peng, J. Huang, J. Zhang, V. Keloth *et al.*, “Me-llama: Medical foundation large language models for comprehensive text analysis and beyond,” 2024.
- [38] E. Bandara, P. Foytik, S. Shetty, and A. Hassanzadeh, “Generative-ai(with custom-trained meta’s llama2 llm), blockchain, nft, federated learning and pbom enabled data security architecture for metaverse on 5g/6g environment,” in *2024 IEEE 21st International Conference on Mobile Ad-Hoc and Smart Systems (MASS)*, 2024, pp. 118–124.
- [39] G. Marvin, N. Hellen, D. Jjingo, and J. Nakatumba-Nabende, “Prompt engineering in large language models,” in *International Conference on Data Intelligence and Cognitive Informatics*. Springer, 2023, pp. 387–402.
- [40] E. Bandara, S. H. Bouk, S. Shetty, S. Roy, R. Mukkamala, A. Rahman, P. Foytik, X. Liang, N. W. Keong, and K. De Zoysa, “Llama-recipe — fine-tuned meta’s llama llm, pbom and nft enabled 5g network-slice orchestration and end-to-end supply-chain verification platform,” in *2025 IEEE 22nd Consumer Communications & Networking Conference (CCNC)*, 2025, pp. 1–6.
- [41] Y. Auxéméry, “Post-traumatic psychiatric disorders: Ptsd is not the only diagnosis,” *La Presse Médicale*, vol. 47, no. 5, pp. 423–430, 2018.
- [42] T. Reason, E. Benbow, J. Langham, A. Gimblett, S. L. Klijn, and B. Malcolm, “Artificial intelligence to automate network meta-analyses: Four case studies to evaluate the potential application of large language models,” *PharmacoEconomics-Open*, pp. 1–16, 2024.
- [43] E. Bandara, S. Shetty, R. Mukkamala, A. Rahman, P. Foytik, X. Liang, K. De Zoysa, and N. W. Keong, “Devsec-gpt — generative-ai (with custom-trained meta’s llama2 llm), blockchain, nft and pbom enabled cloud native container vulnerability management and pipeline verification platform,” in *2024 IEEE Cloud Summit*, 2024, pp. 28–35.

- [44] J. Wang, “A tutorial on llm reasoning: Relevant methods behind chatgpt o1,” *arXiv preprint arXiv:2502.10867*, 2025.
- [45] B. Yang, S. Jiang, L. Xu, K. Liu, H. Li, G. Xing, H. Chen, X. Jiang, and Z. Yan, “Drhouse: An llm-empowered diagnostic reasoning system through harnessing outcomes from sensor data and expert knowledge,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 4, pp. 1–29, 2024.
- [46] M. Böther, D. Graur, X. Yao, and A. Klimovic, “Decluttering the data mess in llm training,” in *2nd Workshop on Hot Topics in System Infrastructure (HotInfra 2024)*. HotInfra 2024, 2024.
- [47] B. Perak, S. Beliga, and A. Meštrović, “Incorporating dialect understanding into llm using rag and prompt engineering techniques for causal commonsense reasoning,” in *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, 2024, pp. 220–229.
- [48] E. Chen, C. Lin, X. Tang, A. Xi, C. Wang, J. Lin, and K. R. Koedinger, “Vtutor: An open-source sdk for generative ai-powered animated pedagogical agents with multi-media output,” *arXiv preprint arXiv:2502.04103*, 2025.
- [49] A. Yehudai, L. Eden, A. Li, G. Uziel, Y. Zhao, R. Bar-Haim, A. Cohan, and M. Shmueli-Scheuer, “Survey on evaluation of llm-based agents,” *arXiv preprint arXiv:2503.16416*, 2025.
- [50] T. Eelbode, P. Sinonquel, F. Maes, and R. Bisschops, “Pitfalls in training and validation of deep learning systems,” *Best Practice & Research Clinical Gastroenterology*, vol. 52, p. 101712, 2021.
- [51] Y. Wang, J. Wei, C. Y. Liu, J. Pang, Q. Liu, A. P. Shah, Y. Bao, Y. Liu, and W. Wei, “Llm unlearning via loss adjustment with only forget data,” *arXiv preprint arXiv:2410.11143*, 2024.
- [52] R. S. McIntyre, J. K. Soczynska, D. S. Cha, H. O. Woldeyohannes, R. S. Dale, M. T. Alsuwaidan, L. A. Gallaugher, R. B. Mansur, D. J. Muzina, A. Carvalho *et al.*, “The prevalence and illness characteristics of dsm-5-defined “mixed feature specifier” in adults with major depressive

- disorder and bipolar disorder: results from the international mood disorders collaborative project,” *Journal of Affective Disorders*, vol. 172, pp. 259–264, 2015.
- [53] S. Wichmann, C. Kirschbaum, C. Böhme, and K. Petrowski, “Cortisol stress response in post-traumatic stress disorder, panic disorder, and major depressive disorder patients,” *Psychoneuroendocrinology*, vol. 83, pp. 135–141, 2017.
 - [54] M. Xiang, R. Fernando, and B. Wang, “On-device qwen2. 5: Efficient llm inference with model compression and hardware acceleration,” *arXiv preprint arXiv:2504.17376*, 2025.
 - [55] C. Wanderley Espinola, J. C. Gomes, J. Mônica Silva Pereira, and W. P. dos Santos, “Detection of major depressive disorder, bipolar disorder, schizophrenia and generalized anxiety disorder using vocal acoustic analysis and machine learning: an exploratory study,” *Research on Biomedical Engineering*, vol. 38, no. 3, pp. 813–829, 2022.
 - [56] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, “Large language models encode clinical knowledge,” *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.
 - [57] D. McDuff, M. Schaekermann, T. Tu, A. Palepu, A. Wang, J. Garrison, K. Singhal, Y. Sharma, S. Azizi, K. Kulkarni *et al.*, “Towards accurate differential diagnosis with large language models,” *arXiv preprint arXiv:2312.00164*, 2023.

Authors



Eranga Bandara is a Senior Research Scientist at the Virginia Modeling, Analysis, and Simulation Center (VMASC), Old Dominion University, Virginia, USA. He done his Ph.D. in Computer Science with a specialization in blockchain and distributed systems. Following his doctoral work, he developed a deep interest in neuroscience and is currently pursuing academic studies in the Fundamentals of Neuroscience program at Harvard University. His interdisciplinary research interests span privacy-preserving AI, distributed systems, blockchain architectures, and next-generation wireless networks (5G/6G), with a growing focus on brain modeling, the application of artificial intelligence in neuroscience, and leveraging neuroscience principles to advance artificial intelligence.



Dr. Ross Gore is a Research Associate Professor at the Virginia Modeling, Analysis and Simulation Center (VMASC) at Old Dominion University, Suffolk, Virginia, USA. He earned his Ph.D. (2012) and M.S. (2007) in Computer Science from the University of Virginia and his B.S. (2003) in Computer Science from the University of Richmond. Dr. Gore's research focuses on data science, predictive analytics, and simulation validation, with applications spanning public health, city planning, cybersecurity, and risk assessment. He is particularly interested in leveraging diverse data sources—from mobile devices to social media—to inform critical decision-making. Recent projects include using mobile phone data to inform public health policies during the COVID-19 pandemic, applying social media analytics to explain geographic variations in obesity, and customizing cyber vulnerability assessments—often developed through close collaboration with stakeholders and iterative prototyping.



Dr. Atma Ram Yarlagadda, MD is an experienced psychiatrist and Installation Director of Psychological Health at McDonald Army Health Center, Fort Eustis, Virginia. He earned his medical degree from Danylo Halytsky Lviv State Medical University in 1983 and has over 40 years of clinical practice in military psychiatry. His work focuses on advancing mental health services for military personnel, emphasizing evidence-based assessment and treatment of mood and trauma-related disorders.



Dr. Anita H. Clayton, MD is the Wilford W. Spradlin Professor and Chair of the Department of Psychiatry and Neurobehavioral Sciences at the University of Virginia School of Medicine. She completed her MD and psychiatry residency at UVA, followed by service in the U.S. Navy Medical Corps before joining UVA faculty in 1990. Dr. Clayton is a leading expert in major depressive disorder, women's mental health, sexual dysfunction, and reproductive psychiatry with over 225 peer-reviewed publications. She pioneered validated assessment tools such as the CSFQ, SIDI-F and DSDS and served as president of the International Society for the Study of Women's Sexual Health. She is the current president of the American Society of Clinical Psychopharmacology.



Col. Dr. Preston Samuel L., DO, MS, FAPA currently serves as Commander and CEO of Blanchfield Army Community Hospital, Fort Campbell, Kentucky. A dual board-certified physician in Family Medicine and Psychiatry, he graduated from the Lake Erie College of Osteopathic Medicine and completed residencies at Walter Reed Army Medical Center. His distinguished military career includes leadership roles such as Chief of Behavioral Health and Regional Director of Psychological Health across multiple Army commands.



Dr. Christopher K. Rhea is the Associate Dean for Research & Innovation in the College of Health Sciences at Old Dominion University. He is a recognized expert at the intersection of neuromotor control and advanced technology, using tools such as virtual reality and smartphones to address human health challenges like fall prevention in older adults, concussion assessment, and rehabilitation. His research has been supported by major agencies including the NIH, Department of Defense, US Navy, HRSA, and the Women's Football Foundation, and he is known for building interdisciplinary teams to solve complex health problems.



Dr. Sachin Shetty is an Associate Director in the Virginia Modeling, Analysis and Simulation Center at Old Dominion University and an Associate Professor with the Department of Computational Modeling and Simulation Engineering. Sachin Shetty received his PhD in Modeling and Simulation from the Old Dominion University in 2007. His research interests lie at the intersection of computer networking, network security and machine learning. Recently, he has been involved with developing cyber risk/resilience metrics for critical infrastructure and blockchain technologies for distributed system security. His laboratory has been supported by the National Science Foundation, Air Office of Scientific Research, Air Force Research Lab, Office of Naval Research, Department of Homeland Security, and Boeing. He has published over 150 research articles in journals and conference proceedings and four books. He is the recipient of Commonwealth Cyber Initiative Research Fellow, Fulbright Specialist award, EPRI Cybersecurity Research Challenge award, DHS Scientific Leadership Award and has been inducted in Tennessee State University's million-dollar club.