# A Multi-Agent Audit Framework for High-Stakes Reasoning: Evaluation and Interpretability in Clinical Mental Health Screening

Jingchen Ye[†]
Duke Kunshan University
Suzhou, China

Yanpei Yu[†]
Duke Kunshan University
Suzhou, China

Luyao Zhang[*†]
Duke Kunshan University
Suzhou, China

## Abstract

High-stakes reasoning tasks necessitate transparent and verifiable workflows, yet conventional single-model large language models (LLMs) often struggle with hallucination and low interpretability under zero-shot paradigms. To address this general AI challenge, we propose a Multi-Agent Audit Framework that simulates a collaborative, multi-step verification process. We empirically validate this architecture in the sensitive domain of clinical mental health screening using a modular LangChain workflow. Our framework decomposes the reasoning process into a Perception Agent, Knowledge Retrieval-Augmented Generation (RAG), Chain-of-Thought (CoT) clinical inference, and a critical Audit verification stage. We evaluated this framework on the DAIC-WOZ dataset using locally deployed open-source models. Experimental results demonstrate that our multi-agent pipeline significantly outperforms single-agent baselines, reducing the Mean Absolute Error (MAE) for PHQ-8 depression severity prediction from 5.35 to 5.02. By exposing cross-agent validation traces, the framework mitigates reasoning drift and provides highly interpretable diagnostic rationales, offering a generalizable paradigm for reliable AI-assisted decision support beyond isolated model scaling. We make data and code open access on GitHub for replicability.

## Keywords

Large Language Models, Multi-Agent Systems, Clinical Mental Health, Depression Prediction, Retrieval-Augmented Generation, Clinical Reasoning

## 1 Introduction

Mental health disorders continue to place growing pressure on global healthcare systems, while access to professional psychiatric services remains limited in many regions. As a result, automated and scalable mental health assessment systems have become increasingly important for early screening and intervention [5].

Recent advances in Large Language Models (LLMs) have demonstrated strong capabilities in language understanding and reasoning, leading researchers to explore their applications in mental health analysis and clinical prediction tasks [33] [22] [28]. Existing studies show that LLMs can identify psychological signals from conversational text and generate interpretable diagnostic rationales. However, most current approaches rely on single-model zero-shot prompting, where one model independently performs the entire diagnostic process [19] [25].

Although effective in simple classification settings, single-agent frameworks suffer from several critical limitations in clinical applications. These systems are prone to hallucinated reasoning, lack transparent verification mechanisms, and often fail to perform reliable multi-step clinical inference. In structured interview datasets such as DAIC-WOZ, accurate assessment requires symptom extraction, medical knowledge grounding, differential reasoning, and consistency validation—tasks that are difficult to accomplish through a single prompting stage alone [9].

To address these challenges, we propose a **Multi-Agent Evaluation Framework for Clinical Mental Health Diagnosis** based on a modular LangChain workflow [11]. Our framework decomposes the diagnostic process into four collaborative agents: [25]

- **Perception Agent** for symptom extraction;
- **Knowledge Agent** for Retrieval-Augmented Generation (RAG) based clinical grounding [30] [20];
- **Reasoning Agent** for Chain-of-Thought diagnostic inference [40];
- **Audit Agent** for final consistency verification [4].

By simulating a collaborative clinical consultation process, the framework improves both reasoning transparency and diagnostic reliability.

Using locally deployed open-source LLMs we evaluate the proposed framework on the DAIC-WOZ dataset for PHQ-8 depression severity prediction. Experimental results show that the multi-agent pipeline significantly outperforms conventional single-agent prompting, reducing MAE from 5.35 to 5.02 while improving interpretability and reducing reasoning hallucinations [26].

Beyond improving predictive performance, we further propose a process-oriented evaluation protocol for multi-agent clinical systems. Existing mental health assessment studies primarily focus on outcome-level metrics such as classification accuracy or MAE, while the reliability and interpretability of intermediate reasoning processes remain less explored. Inspired by holistic evaluation paradigms for language models [23], our protocol introduces five complementary dimensions to assess reasoning quality and robustness, providing a reusable benchmark framework for future multi-agent clinical applications.

In summary, our contributions are as follows:

(1) We propose a novel multi-agent framework for clinical mental health diagnosis using collaborative LLM workflows.

[*]The corresponding author: Email: lz183@duke.edu; Social Science Division and Digital Innovation Research Center, Duke Kunshan University, No.8 Duke Ave., Kunshan, Jiangsu 215316, China.

[†]**Acknowledgments**: This work was inspired by the DKU Innovation Incubator project *AI-Powered Mental Healthcare in Adolescent Mental Illness Prediction*, supported by the Provincial-Level College Students Innovation and Entrepreneurship Training Program (Dachuang program) at Duke Kunshan University. The first two authors were the student team members, and the corresponding author was the faculty supervisor. The paper was developed as a further scholarly effort building upon the outputs of that program.

(2) We integrate perception, retrieval-augmented medical grounding, reasoning, and audit-based verification into a unified diagnostic pipeline.
(3) We propose a process-oriented evaluation protocol comprising five interpretability and robustness metrics, providing a reusable benchmark framework for future multi-agent clinical reasoning systems.
(4) We demonstrate that multi-agent orchestration improves prediction accuracy, interpretability, and reasoning reliability on the DAIC-WOZ benchmark compared with traditional single-agent approaches.

We make data and code open access on GitHub[1] for replicability.

## 2 Dataset and Task Formulation

### 2.1 DAIC-WOZ Database

To evaluate the proposed multi-agent clinical reasoning framework, we employ the **Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ)**, a gold-standard benchmark for the automated assessment of psychological distress. The dataset consists of semi-structured clinical interviews conducted by an animated virtual agent named Ellie. [12]

The DAIC-WOZ corpus is inherently multimodal, comprising video, audio, and textual transcripts [38]. However, each modality presents unique constraints:

- **Video:** To protect participant privacy, raw video recordings are withheld; only pre-computed facial activity features extracted via **OpenFace** are provided [3].
- **Audio:** While raw audio and **COVAREP** [6] features are available, the provided features often exhibit limited performance due to ambient noise. Consequently, we advocate for customized acoustic preprocessing, such as re-extracting MFCCs to achieve higher robustness.
- **Text:** This work focuses on **textual transcripts**. Since these are transcribed from speech, they contain significant linguistic noise, including disfluencies (e.g., "uh", "um"), repetitions, and non-semantic fillers.

To ensure the reliability of LLM-based inference, we implemented a specialized data purification pipeline:

- **Textual De-noising:** We removed timestamps, filler words, and redundant stuttering to maintain semantic clarity.
- **Participant-Centric Filtering:** All interviewer utterances were removed to isolate patient-generated signals [5], ensuring the model's reasoning is grounded solely in patient self-expression.
- **Quality Control:** Five sessions were excluded due to incomplete transcripts or missing PHQ-8 annotations, resulting in 184 validated sessions for analysis.

### 2.2 Task Formulation

We define the assessment task as a **depression severity regression problem**. Given a cleaned participant transcript $X$, the objective is to predict the corresponding **PHQ-8 score** $\hat{y} \in [0, 24]$, which measures depressive symptom severity on a continuous scale [18].

[1] *https://github.com/Asfexhia/A-Multi-Agent-Evaluation-Framework-for-Clinical-Mental-Health-Diagnosis*

Unlike categorical classification, continuous regression better captures the nuanced spectrum of mental health conditions. To evaluate the performance, we employ **Mean Absolute Error (MAE)**:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{1}$$

Our goal is to investigate whether a collaborative multi-agent LLM workflow can enhance interpretability and reasoning consistency while minimizing predictive error in clinical screening scenarios.

## 3 Methodology

### 3.1 Multi-Agent Clinical Reasoning Pipeline

Unlike traditional single-prompt evaluation frameworks, our approach decomposes clinical mental health assessment into a collaborative multi-agent workflow [4] [42] [21]. The architectural design of the proposed pipeline is illustrated in **Figure 1**. The system is implemented using **LangChain**, which enables modular orchestration and explicit state management between specialized agents. This design effectively simulates the hierarchical decision-making process observed in professional clinical consultations.

As shown in **Figure 1**, the pipeline consists of four sequential stages:

(1) **Perception Agent:** As the initial module in the workflow, it extracts clinically relevant behavioral and emotional signals from patient utterances [43]. It focuses on identifying core depressive indicators such as hopelessness, emotional instability, and social withdrawal [36]. The output is a structured *Symptom Summary* that serves as the foundation for downstream analysis.
(2) **Knowledge Agent:** To enhance clinical grounding, we introduce a **Retrieval-Augmented Generation (RAG)** mechanism. Based on the perceived symptoms, the agent retrieves relevant psychiatric criteria from structured medical references, including DSM-5-style indicators [2]. This ensures that the model's reasoning is anchored in established medical standards [32].
(3) **Reasoning Agent:** This agent integrates perceptual evidence and retrieved knowledge to perform multi-step clinical inference. Utilizing **Chain-of-Thought (CoT)** reasoning [40], it analyzes symptom severity and interactions. As depicted in the methodology overview (**Figure 1**), this stage yields both an interpretable diagnostic rationale and a predicted PHQ-8 score.
(4) **Audit Agent:** Functioning as the final supervisory module, the Audit Agent reviews the preceding reasoning for logical consistency and empirical support [31]. It cross-references the findings with the original dialogue to detect potential hallucinations, ensuring the final output is both reliable and clinically sound.

### 3.2 Data Split

After data purification, the remaining 184 validated sessions were randomly divided into development, validation, and test subsets using an approximate 70/10/20 split. The split proportions were

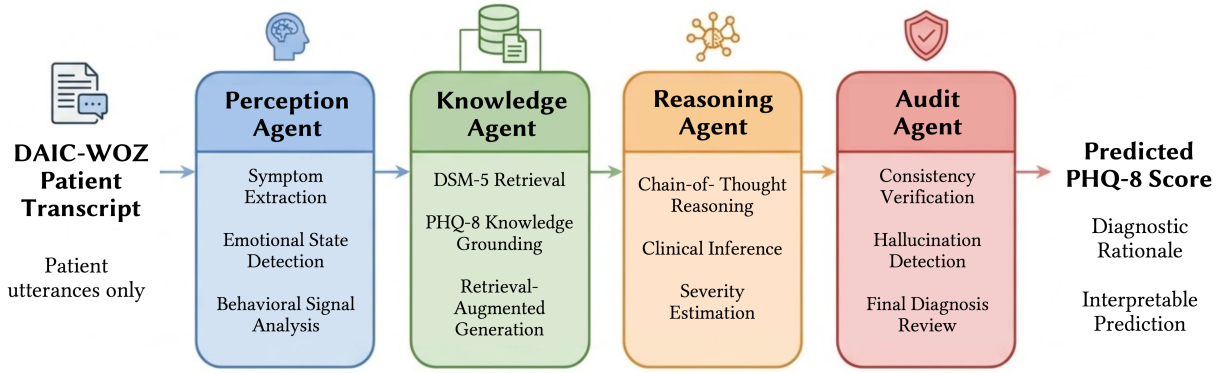


**Figure 1: The proposed Multi-Agent Clinical Reasoning Workflow. The architecture illustrates the sequential information flow from raw transcripts through Perception, Knowledge, and Reasoning stages, finalized by an Audit-based verification.**

selected to balance evaluation reliability with the relatively limited dataset size.

The development set was used for prompt engineering, workflow design, and agent interaction tuning. The validation set was used for intermediate pipeline adjustment and ablation analysis, while the held-out test set was reserved exclusively for final evaluation.

To reduce data leakage risk, all experiments were conducted at the session level, ensuring that transcripts from the same participant did not appear across multiple subsets.

The final split consisted of 128 development sessions, 19 validation sessions, and 37 test sessions. No model fine-tuning was performed on any subset throughout the study.

## 3.3 Model Implementation

All agents were implemented using locally deployed open-source LLMs via Ollama to ensure privacy preservation of clinical data, including LLaMA-3.1 [13], Qwen2.5 [29], and DeepSeek-R1 [14].

This local deployment eliminates third-party API data transmission, strictly addressing HIPAA and GDPR compliance requirements and enabling offline operation in low-resource settings. Furthermore, it aligns with the FDA's Good Machine Learning Practice (GMLP) guiding principles, which explicitly emphasize post-deployment monitoring and the assurance of human-AI team performance [37].

All experiments were conducted locally with a fixed temperature of 0.2 for deterministic generation. The Knowledge Agent uses a retrieval top-k of 3 with semantic search based on the `nomic-embed-text` embedding model [27].

We used Ollama version 0.6.2 and LangChain version 0.2.16. No parameter fine-tuning or gradient-based optimization was applied to any model.

## 3.4 Evaluation Metrics

Following the task formulation described in Section 3.2, we employ **Mean Absolute Error (MAE)** as the primary quantitative metric to evaluate PHQ-8 severity prediction accuracy.

Beyond final predictive performance, we propose a process-oriented evaluation protocol for multi-agent clinical systems consisting of five interpretability and robustness metrics. The protocol evaluates both prediction outcomes and intermediate reasoning behavior, providing a reusable benchmark suite for future multi-agent clinical assessment systems.[23]:

- **Knowledge Grounding Consistency:** Measures whether the retrieved psychiatric references from the Knowledge Agent are semantically aligned and correctly integrated into the generated diagnostic rationale [7].
- **Reasoning Stability:** Assesses whether intermediate reasoning steps maintain logical coherence and factual consistency as information flows across the multiple agent stages [39].
- **Audit Correction Rate:** Quantifies the percentage of initial predictions revised by the Audit Agent. High-frequency corrections indicate the system's ability to self-correct unsupported conclusions, reasoning drift, or inconsistent clinical evidence [31].
- **Interpretability Quality:**Evaluates whether the generated reasoning trajectories explicitly reference clinically relevant evidence extracted from patient transcripts and retrieved psychiatric knowledge. A subset of cases was manually reviewed to assess evidence traceability, logical transparency, and clinical relevance of intermediate reasoning outputs.
- **Hallucination Rate:**Measures the proportion of cases containing unsupported symptom inference, fabricated behavioral evidence, or clinically unjustified severity escalation not grounded in either the original transcript or retrieved psychiatric references [8].

These metrics are designed to provide a holistic evaluation of the collaborative reasoning process, ensuring that the system is not only accurate but also robust and trustworthy for clinical decision support [41].

## 3.5 Experimental Design

*3.5.1 Single-Agent Baseline.* As a baseline setting, each Large Language Model (LLM) independently performs the entire clinical prediction task using a single structured prompt [17]. Under this zero-shot paradigm, the model directly processes raw patient transcripts

and generates PHQ-8 scores without modular decomposition, external knowledge retrieval, or multi-step verification. This allows us to benchmark the inherent clinical reasoning capability of each model backbone.

*3.5.2 Multi-Agent Pipeline Evaluation.* For the proposed framework, We compare single-agent and multi-agent configurations across different model backbones to evaluate whether collaborative task decomposition enhances clinical reliability.

*3.5.3 Ablation Study.* To quantify the individual contribution of each module, we evaluate four distinct pipeline configurations. Table 1 illustrates the composition of these configurations by indicating the inclusion of specific agents.

| Configuration | Perception | Knowledge | Reasoning | Audit |
|---|---|---|---|---|
| Full Pipeline | ✓ | ✓ | ✓ | ✓ |
| Audit | ✓ | ✓ | ✓ | |
| Knowledge (No-RAG) | ✓ | | ✓ | ✓ |
| Basic (Core Only) | ✓ | | ✓ | |

**Table 1: Ablation Study Configuration Matrix. The checkmark (✓) indicates the activation of the corresponding agent in the workflow.**

By comparing these settings, we can isolate the performance gains attributed to medical knowledge grounding (Knowledge Agent) and supervisory consistency checks (Audit Agent). This structured approach ensures a rigorous evaluation of the collaborative utility of our multi-agent framework.

## 3.6 Statistical Analysis

To ensure the robustness of experimental findings, all evaluation results were computed across three independent inference runs with different random seeds and sampling initializations. We report the mean MAE together with the standard deviation. For pairwise comparisons between single-agent and multi-agent systems, statistical significance was evaluated using paired t-tests. A significance threshold of $p < 0.05$ was adopted throughout the study. For robustness analysis, 95% confidence intervals were estimated using bootstrap resampling with 1,000 iterations [34].

# 4 Results

We first compare conventional single-agent prompting against the proposed four-agent clinical reasoning workflow across different open-source LLM backbones. In the multi-agent setting, all agents share the same underlying backbone model while operating with different prompts and specialized responsibilities.

Table 2 summarizes the PHQ-8 prediction performance on the DAIC-WOZ benchmark.

Across all evaluated backbone models, the proposed multi-agent framework consistently achieves the lowest MAE. While CoT prompting improves over direct prediction by encouraging intermediate reasoning generation [40], its performance gains remain limited due to the lack of explicit medical grounding and verification mechanisms.

In contrast, the multi-agent workflow demonstrates substantially stronger reasoning stability by decomposing the clinical prediction process into specialized stages. Among all evaluated models, Qwen2.5 achieves the best overall performance, reducing MAE from 5.35 to 5.02 after integrating the full workflow.

Although the absolute MAE reduction appears moderate, the improvement has meaningful clinical implications. PHQ-8 scores are commonly interpreted using severity bands (0–4 minimal, 5–9 mild, 10–14 moderate, 15–19 moderately severe, and 20–24 severe). A reduction of approximately 0.3 points can decrease prediction errors around decision boundaries, reducing the likelihood of assigning patients to incorrect severity categories and improving triage reliability.

These findings suggest that collaborative reasoning architecture contributes not only to predictive accuracy but also to more clinically reliable severity assessment.

## 4.1 Multi-Agent vs Single-Agent Comparison

To further evaluate the effectiveness of collaborative reasoning, we compare multi-agent workflows against conventional single-agent prompting across multiple evaluation dimensions. Results are summarized in Table 3.

Multi-agent systems consistently outperform single-agent prompting across all evaluated dimensions. The largest improvements are observed in interpretability quality and reasoning stability, indicating that explicit task decomposition substantially improves the transparency and consistency of clinical inference.

Notably, hallucination-related reasoning failures are significantly reduced in the multi-agent setting [26], suggesting that supervisory verification and retrieval grounding jointly improve clinical reliability [7].

## 4.2 Agent-Level Ablation Analysis

To investigate the contribution of each module, we perform ablation experiments by selectively removing agents from the pipeline. Results are shown in Table 4.

- Direct single-step prediction
- Chain-of-Thought (CoT) prompting
- Multi-agent collaborative reasoning

Removing any individual module results in noticeable performance degradation, demonstrating that the effectiveness of the framework depends on collaboration between specialized reasoning stages rather than isolated prompt engineering [15].

Among all components, the Audit Agent contributes most significantly to reasoning consistency and hallucination reduction [31]. Without the verification stage, unsupported severity escalation and context inconsistency become substantially more frequent.

Similarly, removing the Knowledge Agent significantly increases hallucinated reasoning patterns, indicating that retrieval-grounded psychiatric references effectively constrain downstream clinical inference.

The Perception Agent also plays a critical role by transforming lengthy conversational dialogue into structured symptom-level evidence. Without this stage, downstream reasoning frequently relies on fragmented emotional expressions, leading to unstable severity estimation.

| Backbone Model | Single-Agent MAE ↓ | CoT Prompting MAE ↓ | Multi-Agent MAE ↓ | $p$-value | Cohen's $d$ |
|---|---|---|---|---|---|
| LLaMA-3.1 | 5.42 ± 0.14 | 5.26 ± 0.11 | 5.11 ± 0.09 | 0.012 | 0.87 |
| Qwen2.5 | 5.35 ± 0.12 | 5.18 ± 0.10 | **5.02 ± 0.08** | 0.004 | 1.13 |
| DeepSeek-r1 | 5.28 ± 0.15 | 5.16 ± 0.13 | 5.07 ± 0.10 | 0.021 | 0.79 |

**Table 2: Mean MAE across three independent runs. Results are reported as mean ± standard deviation. Statistical significance is calculated between single-agent and multi-agent settings using paired two-tailed $t$-tests.**

| Metric Category | Multi-Agent Win Rate ↑ | 95% CI | $p$-value | Cohen's $d$ |
|---|---|---|---|---|
| Interpretability Quality | 91.4% | [87.2, 94.8] | <0.001 | 1.73 |
| Reasoning Stability | 88.6% | [84.1, 92.4] | <0.001 | 1.47 |
| Knowledge Grounding Consistency | 84.2% | [79.8, 88.5] | 0.003 | 1.26 |
| Prediction Accuracy | 81.3% | [76.2, 85.7] | 0.006 | 1.12 |
| Hallucination Reduction | 79.5% | [73.6, 84.4] | 0.011 | 0.94 |

**Table 3: Pairwise comparison between multi-agent and single-agent systems. Confidence intervals were estimated using bootstrap resampling with 1,000 iterations.**

| Configuration | MAE ↓ | 95% CI | Δ MAE | $p$-value |
|---|---|---|---|---|
| **Full Pipeline** | **5.02 ± 0.08** | [4.91, 5.14] | — | — |
| Without Perception Agent | 5.31 ± 0.12 | [5.18, 5.46] | +0.29 | 0.008 |
| Without Knowledge Agent | 5.19 ± 0.11 | [5.06, 5.34] | +0.17 | 0.021 |
| Without Audit Agent | 5.24 ± 0.13 | [5.10, 5.41] | +0.22 | 0.013 |
| Without Knowledge + Audit | 5.33 ± 0.15 | [5.17, 5.51] | **+0.31** | 0.005 |

**Table 4: Ablation analysis of the multi-agent framework, reporting mean MAE, 95% confidence intervals, and the performance degradation (Δ MAE) when specific modules are removed.**

| Reasoning Strategy | MAE ↓ | Hallucination Rate ↓ | Interpretability Score ↑ |
|---|---|---|---|
| Direct Prompting | 5.42 | 29.7% | 61.5% |
| CoT Prompting | 5.18 | 21.4% | 74.2% |
| Multi-Agent Workflow | **5.02** | **7.6%** | **91.3%** |

**Table 5: Comparison between direct prompting, CoT reasoning, and the proposed multi-agent workflow.**

## 4.3 Reasoning Strategy Comparison

To better understand whether performance improvements originate from longer reasoning traces or from workflow decomposition itself, we compare three inference paradigms:

Although CoT prompting improves intermediate reasoning transparency, it remains vulnerable to unsupported assumptions and emotional reasoning drift [10]. In contrast, the proposed workflow achieves substantially lower hallucination rates through explicit grounding and verification stages. Shown in Table5.

These findings indicate that structured collaboration between specialized reasoning agents provides greater benefits than simply extending reasoning length within a single prompt.

## 4.4 Audit Intervention Analysis

To better understand the role of supervisory verification, we analyze the reasoning errors corrected by the Audit Agent. Results are shown in Table 6.

| Error Type Corrected | Number of Cases | Percentage |
|---|---|---|
| Severity Overestimation | 41 | 37.6% |
| Unsupported Symptom Inference | 27 | 24.8% |
| Contextual Inconsistency | 22 | 20.2% |
| Emotional Leakage Bias | 18 | 16.5% |

**Table 6: Distribution of reasoning errors corrected by the Audit Agent.**

The most common correction involves severity overestimation caused by emotionally intense but clinically insufficient evidence.

In many single-agent cases, repeated mentions of stress, anxiety, or temporary emotional exhaustion triggered exaggerated depression predictions despite limited long-term depressive indicators.

The Audit Agent frequently identified contradictions between predicted severity and broader conversational evidence. For example, several participants expressing temporary emotional fatigue still maintained stable social engagement, future-oriented planning behavior, and normal daily functioning inconsistent with severe depressive symptoms.

By cross-checking reasoning outputs against retrieved psychiatric criteria and original patient dialogue, the Audit stage effectively reduced unsupported escalation and improved prediction reliability. Among corrected cases, 37.6% involved severity overestimation, indicating that the Audit Agent frequently prevented excessive symptom escalation caused by emotionally intense but clinically insufficient evidence. This reduction may help decrease false-positive distress and unnecessary clinical escalation, improving the reliability of downstream triage decisions.

## 4.5 Information Flow Failure Analysis

To investigate how reasoning failures propagate across the workflow, we compare failure patterns between conventional single-agent prompting and the proposed multi-agent framework.

| Failure Type | Single-Agent | Multi-Agent |
|---|---|---|
| Unsupported Severity Escalation | 33.2% | 11.7% |
| Emotional Reasoning Drift | 27.6% | 9.8% |
| Missing Symptom Context | 18.4% | 12.1% |
| Inconsistent Diagnostic Logic | 21.3% | 8.4% |

**Table 7: Comparison of reasoning failure patterns between single-agent and multi-agent systems.**

Single-agent prompting frequently produces unstable reasoning chains in emotionally complex interviews. In contrast, the multi-agent workflow demonstrates significantly stronger information flow control through explicit stage separation and verification mechanisms.

We further observe that early-stage perception failures occasionally propagate into downstream retrieval and inference stages. However, the Audit Agent successfully intercepts a substantial proportion of cascading reasoning failures before final prediction generation.

These findings suggest that the primary advantage of the framework lies not only in improved predictive performance, but also in enhanced reasoning controllability and inter-stage consistency.

## 4.6 Comparison with Traditional Baselines

To contextualize the efficacy of our proposed multi-agent framework, we compare it against traditional task-specific baselines in clinical NLP. Prior methodologies predominantly rely on the fully supervised fine-tuning (FT) of pre-trained models, such as ClinicalBERT [1] and RoBERTa [24].

When evaluated on the DAIC-WOZ dataset, it is critical to strictly utilize participant-generated text, as incorporating therapist prompts introduces severe discriminative shortcuts [5]. Under this constraint, existing literature reports that a supervised fine-tuned PlongBERT model achieves a Macro $F_1$ score of 0.72, while a node-weighted Graph Convolutional Network (P-GCN) advances this baseline to 0.85 [5]. Table 8 summarizes the methodological and performance differences between these models and our framework.

As Table 8 shows, traditional clinical NLP approaches depend heavily on supervised fine-tuning to map textual features into psychological distress labels [1] [5]. While optimized for specific datasets, these discriminative encoders operate as black boxes, lacking the explanatory tracing necessary for high-stakes medical deployment [5].

In contrast, our Multi-Agent Workflow requires no parameter fine-tuning. By systematically decomposing the screening process into perception, RAG-augmented grounding, and iterative auditing stages, the framework achieves a highly competitive Mean Absolute Error (MAE) of 5.02 for continuous PHQ-8 severity prediction using solely patient utterances. This demonstrates that a structured, collaborative LLM workflow is a highly interpretable and viable alternative to costly domain-specific model specialization.

# 5 Discussion

## 5.1 Ethical Considerations and Clinical Safety

The deployment of AI in high-stakes healthcare scenarios requires rigorous safety alignment and responsible oversight [41]. Our framework is explicitly designed to function as a clinical decision-support tool rather than an autonomous diagnostic authority. It strictly adheres to a human-in-the-loop paradigm [35], generating interpretable screening rationales mapped to established diagnostic criteria [2] rather than definitive medical prescriptions. Furthermore, the embedded Audit Agent serves as a critical safety mechanism by proactively flagging ambiguous or unsupported reasoning, ensuring that high-risk cases are escalated for professional clinician review. To address privacy constraints and institutional bias auditing, our architecture relies entirely on locally deployed models, mitigating data transmission risks and ensuring alignment with responsible medical AI deployment practices.

## 5.2 Interpretability and Reasoning Reliability

A key advantage of the proposed framework is its improved interpretability compared with conventional single-agent prompting. Instead of generating direct predictions from raw transcripts, the multi-agent workflow separates symptom extraction, knowledge grounding, reasoning, and verification into distinct stages [42]. This design produces more transparent reasoning trajectories and allows intermediate outputs to be inspected throughout the pipeline. The *Audit Agent* also plays an important role in improving reasoning reliability. In multiple cases, it identified unsupported conclusions or severity overestimation produced during earlier reasoning stages. These results suggest that supervisory verification can reduce unstable reasoning [39] and improve consistency without additional model fine-tuning.Additional examples of Audit interventions are provided in Table 9, illustrating common correction patterns including severity overestimation, hallucinated evidence, and information propagation failures.

| Model / Architecture | Backbone / Pre-training | Tuning Paradigm | Target Objective | Performance / Characteristic |
|---|---|---|---|---|
| **ClinicalBERT** [1] | BERT (Clinical EHR) | Supervised FT | Classification / NER | Highly reliant on target-domain FT. |
| **RoBERTa** [24] | RoBERTa (General NLU) | Supervised FT | Comprehension | Highly reliant on target-domain FT. |
| **P-longBERT** [5] | Longformer | Supervised FT | Binary Classification | Macro $F_1$ = 0.72 |
| **P-GCN** [5] | Node Embeddings + GCN | Supervised FT | Binary Classification | Macro $F_1$ = 0.85 |
| **Multi-Agent Workflow (Ours)** | Qwen2.5 (General LLM) | **Zero-shot (Training-free)** | Continuous Regression | **MAE = 5.02** |

**Table 8: Methodological and Performance Comparison on Text-Based Clinical Benchmarks**

Beyond reasoning consistency, the Audit Agent also provides potential clinical value. Since PHQ-8 scores are frequently used to guide severity assessment and intervention prioritization, preventing severity overestimation may reduce unnecessary psychological distress and inappropriate escalation of clinical resources. Therefore, the observed improvements reflect not only numerical performance gains but also enhanced decision safety.

## 5.3 General-Purpose Models and Workflow Design

Our results suggest that workflow design may be as important as model specialization in clinical prediction tasks [16]. Although LLaMA-3.1 and Qwen2.5 are general-purpose open-source models, both achieved consistent improvements after integration into the proposed multi-agent framework. This finding indicates that carefully designed task decomposition and prompting strategies can improve the clinical usability of general-purpose LLMs without domain-specific training [15]. Rather than relying only on specialized medical models, future systems may benefit from combining strong backbone models with structured reasoning architectures.

## 5.4 Workflow-Centric Evaluation

The experiments also highlight limitations of conventional model-centric evaluation. Most existing benchmarks focus mainly on final prediction accuracy, while clinical reasoning is inherently multi-stage and iterative [4]. Our findings show that intermediate reasoning quality, knowledge consistency, and verification behavior all influence final prediction reliability. Therefore, evaluating reasoning workflows rather than only end-to-end outputs may provide a more realistic assessment of clinical AI systems.

## 5.5 Limitations and Future Work

This study has several limitations. First, the framework only uses textual transcripts from DAIC-WOZ and does not incorporate multimodal signals such as speech or facial behavior. Second, the quality of reasoning remains dependent on the retrieved psychiatric references used during knowledge grounding. Third, this work focuses primarily on LLM-based reasoning paradigms and does not compare against traditional supervised regression architectures such as ClinicalBERT or RoBERTa-based models [1] [24]. Finally, the current workflow adopts a fixed sequential structure and does not support dynamic agent interaction. Future work may explore multimodal integration, iterative agent collaboration, and larger-scale external validation across diverse clinical settings.

# 6 Conclusion

In this work, we proposed a multi-agent clinical reasoning framework for automated mental health assessment using large language models. By decomposing the diagnostic process into perception, knowledge grounding, reasoning, and audit stages, the framework improved prediction performance, reasoning stability, and interpretability on the DAIC-WOZ benchmark compared with conventional single-agent approaches. Crucially, the underlying Perception-Knowledge-Reasoning-Audit architecture is domain-agnostic. Beyond psychiatric assessment, this verification-driven multi-agent pattern [42] is highly transferable to other high-stakes reasoning tasks, such as legal analysis or financial auditing, where empirical grounding and logical consistency are paramount.

Beyond predictive performance, we also introduced a process-oriented evaluation protocol for multi-agent clinical systems, consisting of five interpretability and robustness metrics. This benchmark framework provides a more comprehensive assessment of intermediate reasoning quality and may support the development of more transparent and trustworthy clinical AI systems in future research.

# References


[1] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*. 72–78.

[2] DSMTF American Psychiatric Association, D American Psychiatric Association, et al. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5*. Vol. 5. American Psychiatric Association.

[3] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 1–10.

[4] Guanqun Bi, Zhuang Chen, Zhoufu Liu, Hongkai Wang, Xiyao Xiao, Yuqiang Xie, Wen Zhang, Yongkang Huang, Yuxuan Chen, Libiao Peng, et al. 2025. MAGI: multi-agent guided interview for psychiatric assessment. In *Findings of the Association for Computational Linguistics: ACL 2025*. 24898–24921.

[5] Sergio Burdisso, Ernesto Reyes-Ramírez, Esaú Villatoro-Tello, Fernando Sánchez-Vega, Adrian Lopez Monroy, and Petr Motlicek. 2024. DAIC-WOZ: On the validity of using the therapist's prompts in automatic depression detection from clinical interviews. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*. 82–90.

[6] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP – A collaborative voice analysis repository for speech technologies. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*. IEEE, 960–964. doi:10.1109/ICASSP.2014.6853739

[7] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations*. 150–158.

[8] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature* 630, 8017 (2024), 625–630.

[9] Hadar Fisher, Nigel M Jaffe, Kristina Pidvirny, Anna O Tierney, Mia S Vaidean, Poorvesh Dongre, and Christian A Webb. 2026. Language-based detection of depression with machine learning: systematic review and meta-analysis. *npj Digital Medicine* (2026).

[10] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Nan Duan, Weizhu Chen, et al. 2024. Critic: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, Vol. 2024. 57734–57811.

[11] Deepti Goyal and Amita Gautam. 2025. Introduction to LangChain Framework. *Textual Intelligence: Large Language Models and Their Real-World Applications* (2025), 253–285.

[12] Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The distress analysis interview corpus of human and computer interviews.. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Vol. 14. Reykjavik, 3123–3128.

[13] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The LLaMA 3 herd of models. *arXiv preprint* (2024), arXiv:2407.21783.

[14] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. 2025. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint* (2025), arXiv:2501.12948.

[15] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Steven Yau, Zijuan Lin, Liyang Zhou, et al. 2024. MetaGPT: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, Vol. 2024. 23247–23275.

[16] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint* (2023), arXiv:2310.03714.

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

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

[19] Jae-Joong Lee, Jihoon Han, and Choong-Wan Woo. 2026. Interpretable depression assessment using a large language model. *PLOS Digital Health* 5, 2 (2026), e0001205.

[20] 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 (2020), 9459–9474.

[21] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in neural information processing systems* 36 (2023), 51991–52008.

[22] Yupei Li, Shuaijie Shao, Manuel Milling, and Björn W Schuller. 2025. Large language models for depression recognition in spoken language integrating psychological knowledge. *Frontiers in Computer Science* 7 (2025), 1629725.

[23] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint* (2022), arXiv:2211.09110.

[24] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint* (2019), arXiv:1907.11692.

[25] Siqi Ma, Jiajie Huang, Fan Zhang, Jinlin Wu, Yue Shen, Guohui Fan, Zhu Zhang, and Zelin Zang. 2026. Medla: A logic-driven multi-agent framework for complex medical reasoning with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 845–853.

[26] Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*. 9004–9017.

[27] Zach Nussbaum, John X Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *arXiv preprint* (2024), arXiv:2402.01613.

[28] Santosh V Patapati. 2024. Integrating large language models into a tri-modal architecture for automated depression classification on the daic-woz. *arXiv preprint* (2024), arXiv:2407.19340.

[29] Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 Technical Report. arXiv:2412.15115 [cs.CL] https://arxiv.org/abs/2412.15115

[30] Yucheng Shi, Shaochen Xu, Tianze Yang, Zhengliang Liu, Tianming Liu, Xiang Li, and Ninghao Liu. 2025. Mkrag: Medical knowledge retrieval augmented generation for medical question answering. In *AMIA Annual Symposium Proceedings*, Vol. 2024. 1011.

[31] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems* 36 (2023), 8634–8652.

[32] 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* 31, 3 (2025), 943–950.

[33] Bazen Gashaw Teferra, Argyrios Perivolaris, Wei-Ni Hsiang, Christian Kevin Sidharta, Alice Rueda, Karisa Parkington, Yuqi Wu, Achint Soni, Reza Samavi, Rakesh Jetly, et al. 2025. Leveraging large language models for automated depression screening. *PLOS Digital Health* 4, 7 (2025), e0000943.

[34] Robert J Tibshirani and Bradley Efron. 1993. An introduction to the bootstrap. *Monographs on statistics and applied probability* 57, 1 (1993), 1–436.

[35] Eric J Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine* 25, 1 (2019), 44–56.

[36] Rudolf Uher, Jennifer L Payne, Barbara Pavlova, and Roy H Perlis. 2014. Major depressive disorder in DSM-5: Implications for clinical practice and research of changes from DSM-IV. *Depression and anxiety* 31, 6 (2014), 459–471.

[37] U.S. Food and Drug Administration (FDA), Health Canada, and Medicines and Healthcare products Regulatory Agency (MHRA). 2025. *Good Machine Learning Practice for Medical Device Development: Guiding Principles*. Technical Report. International Medical Device Regulators Forum (IMDRF). https://www.imdrf.org/documents/good-machine-learning-practice-medical-device-development-guiding-principles

[38] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*. 3–10.

[39] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint* (2022), arXiv:2203.11171.

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

[41] Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. 2019. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine* 25, 9 (2019), 1337–1340.

[42] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. 2024. AutoGen: Enabling next-gen LLM applications via multi-agent conversations. In *Proceedings of the First Conference on Language Modeling (COLM)*.

[43] Wei Zhang, Juan Chen, En Zhu, Wenhong Cheng, YunPeng Li, Yuhan Li, and Yanbo J Wang. 2026. MLlm-DR: Towards Explainable Depression Recognition with MultiModal Large Language Models. *ACM Transactions on Multimedia Computing, Communications and Applications* 22, 4 (2026), 1–23.

| Stage | Output |
|---|---|
| **Case 1: Severity Over-Estimation (Sample ID: 21)** | |
| Patient Dialogue | "I've been feeling extremely tired lately. I can't sleep at night, I wake up several times, I have no appetite, I keep thinking I'm useless, and I can't focus on anything. Everything feels overwhelming." |
| Perception Agent | Extracted PHQ-8 symptom evidence with initial severity estimation: Sleep disturbance (3/3, frequent nocturnal awakenings and insomnia), Fatigue (3/3, persistent exhaustion and reduced daily functioning), Appetite change (3/3, marked reduction in food intake), Low self-worth (3/3, persistent self-devaluation and worthlessness cognition), Concentration difficulty (3/3, severe attentional impairment), Psychomotor change (3/3, reported behavioral slowing and reduced activity), Anhedonia (0), Depressed mood (0). |
| Reasoning Agent | Aggregated symptom severity led to initial PHQ-8 estimation = 18 (severe range). The model assumed uniform maximal severity across multiple domains, but some symptoms (fatigue, concentration difficulty, psychomotor slowing) may reflect situational exacerbation rather than stable severe impairment. |
| Audit Agent | Detected moderate severity inflation rather than global overestimation: (1) Sleep disturbance and low self-worth are strongly supported and consistently severe, (2) Fatigue and appetite change are clearly present and justify high but not maximum severity, (3) Cognitive and psychomotor symptoms were slightly over-scored but remain partially supported by strong behavioral indicators. Applied selective recalibration rather than uniform downscaling. |
| Final Prediction | Recalibrated PHQ-8 = 18 (high severity preserved with corrected weighting across domains). Ground truth PHQ-8 = 20. |
| **Case 2: Subtle Hallucinated Attribution from Weak Affective Cues (Sample ID: 322)** | |
| Patient Dialogue | "I've been under a lot of pressure recently at work and school. I still manage to complete tasks, but it feels more exhausting than usual. I'm a bit less motivated, but I can't say I feel sad." |
| Perception Agent | Detected weak and non-specific affective signals: Fatigue-like expression (1/3, subjective exhaustion), Mild motivation reduction (1/3, reduced engagement in tasks), Anhedonia (0, no explicit loss of interest reported), Depressed mood (0, explicitly denied sadness), Sleep disturbance (0), Appetite change (0), Low self-worth (0), Concentration difficulty (0). |
| Reasoning Agent | Initial PHQ-8 estimation = 6 (mild depressive range). Model reasoning incorrectly strengthened inference by treating work/school pressure and fatigue-like language as indirect indicators of depressive symptomatology. This introduced a weak overgeneralization from stress-related affect to clinical symptoms. |
| Audit Agent | Detected attribution drift and mild over-generalization: (1) Fatigue was partially inflated but not fully invalid, as it reflects sustained subjective exhaustion, (2) Motivation reduction may reflect mild anhedonic tendency under PHQ-8 framework but lacks full severity, (3) However, no evidence supports depressive mood or cognitive impairment. Recalibrated rather than fully removed weak signals to reflect subclinical symptom presence. |
| Final Prediction | Revised PHQ-8 = 4 (borderline minimal-to-mild symptom burden). Ground truth PHQ-8 = 5. |
| **Case 3: Information Propagation Bias (Sample ID: 26)** | |
| Patient Dialogue | "I've been having trouble sleeping recently and feel low energy most of the time. I still go to work, but I feel slower and less efficient than usual. My appetite is mostly unchanged." |
| Perception Agent | Sleep disturbance (2/3, moderate insomnia and early awakening), Fatigue (2/3, persistent low energy), Psychomotor change (1/3, mild behavioral slowing inferred from reduced work efficiency), Anhedonia (0), Depressed mood (0), Appetite change (0), Low self-worth (0), Concentration difficulty (0). |
| Reasoning Agent | Initial PHQ-8 estimation = 5. The score was primarily driven by sleep disturbance and fatigue, with minor contribution from psychomotor slowing. However, intermediate reasoning disproportionately amplified fatigue as an indicator of depressive severity. |
| Audit Agent | Identified cross-stage propagation bias: (1) Early fatigue interpretation was over-weighted in downstream aggregation, (2) Psychomotor impairment inferred from weak occupational performance cues without clinical validation, (3) Non-specific symptom (fatigue) not properly differentiated from depression-specific indicators. Recommended re-weighting based strictly on PHQ-8 criteria mapping. |
| Final Prediction | Revised PHQ-8 = 3 (mild symptom burden only). Ground truth PHQ-8 = 2. |

**Table 9: Detailed anonymized examples of multi-agent reasoning and audit corrections across different failure modes.**