LLAMADRS: Prompting Large Language Models for Interview-Based Depression Assessment

Anonymous ACL submission

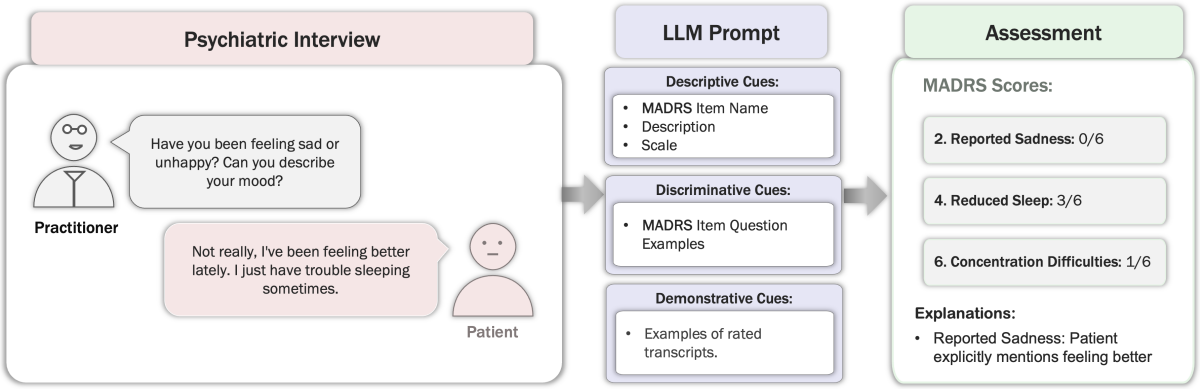


Figure 1: **Overview of the LLAMADRS Framework.** The left panel illustrates a traditional structured clinical interview between a patient and a clinician. The right panel demonstrates the automated depression assessment process using a large language model (**Qwen 2.5 - 72b**), including scoring of MADRS items with explanations for each score.

Abstract

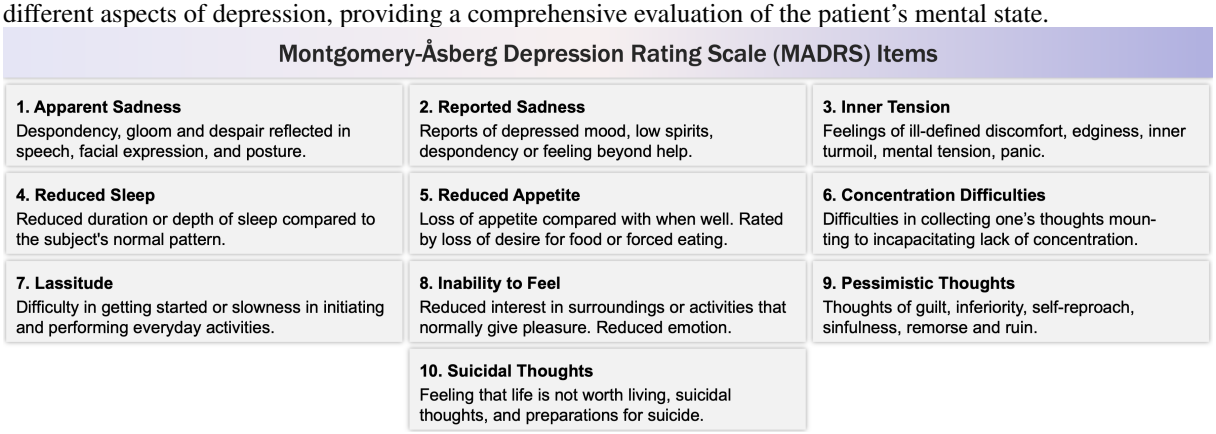
This study introduces LLAMADRS, a novel framework leveraging open-source Large Language Models (LLMs) to automate depression severity assessment using the **Montgomery-Åsberg Depression Rating Scale** (MADRS). We employ a zero-shot prompting strategy with carefully designed cues to guide the model in interpreting and scoring transcribed clinical interviews. Our approach, tested on 236 real-world interviews from the **Context-Adaptive Multimodal Informatics** (CAMI) dataset, demonstrates strong correlations with clinician assessments. The Qwen 2.5–72b model achieves near-human level agreement across most MADRS items, with Intraclass Correlation Coefficients (ICC) closely approaching those between human raters. We provide a comprehensive analysis of model performance across different MADRS items, highlighting strengths and current limitations. Our findings suggest that LLMs, with appropriate prompting, can serve as efficient tools for mental health assessment, potentially increasing accessibility in resource-limited settings. However, challenges remain, particularly in assessing symptoms that rely on non-verbal cues, underscoring the need for multimodal approaches in future work.

1 Introduction

Depression, a leading cause of disability worldwide, affects approximately 280 million people (Institute for Health Metrics and Evaluation, 2023). Accurate and timely assessment is crucial for effective treatment. However, traditional diagnostic methods face significant challenges. These methods, such as structured interviews paired with clinical rating scales like the **Montgomery-Åsberg Depression Rating Scale** (MADRS), require extensive clinician training and are prone to subjective interpretation (Montgomery and Asberg, 1979).

Large Language Models (LLMs) offer a promising alternative to traditional methods. They have the potential to enable more frequent assessments and provide valuable insights into symptom fluctuations and treatment responses (Torous et al., 2021). The application of LLM to the assessment of depression encompasses two critical aspects: the ability to conduct patient interviews and the ability to evaluate the severity of depression based on the content of the interview. Our work concentrates on the latter, harnessing LLMs to automate the assessment process. Specifically, we explore the potential of LLMs to accurately analyze and score patient interviews conducted by human clinicians—a crucial

Figure 2: **Montgomery-Åsberg Depression Rating Scale (MADRS) Items.** The scale includes ten items assessing different aspects of depression, providing a comprehensive evaluation of the patient's mental state.



step towards more efficient and objective psychiatric evaluations.

The evaluation of LLMs in clinical settings has historically presented various challenges, particularly in managing lengthy and complex clinical interview transcripts. However, recent advances in open-source models such as LLAMA 3.1 (Dubey et al., 2024) and Qwen 2.5 (Team, 2024) have enabled zero-shot inference on long-context data of up to **128k** tokens. These developments open new possibilities for analyzing extensive clinical information, potentially enhancing the depth and accuracy of psychiatric evaluations.

Our study leverages the **Context-Adaptive Multimodal Informatics** (CAMI) dataset, comprising authentic clinical interviews annotated by mental health professionals. Unlike previous work that often relied on synthetic or non-clinical data, our use of real-world interviews substantially enhances the study's ecological validity. This approach enables a more rigorous evaluation of LLM applicability in psychiatric settings, grounding our findings in the nuances of actual clinical practice—a crucial advancement over prior research in this field.

In this study, we introduce LLAMADRS, a framework that demonstrates the viability of open-source LLMs for depression assessment through careful prompt engineering. Our zero-shot approach achieves strong correlations with clinician assessments for several MADRS items, particularly those involving concrete symptoms like reduced appetite. However, challenges persist in items requiring visual observation.

Our contributions are as follows:

- **C1:** A structured prompting strategy incorporating descriptive and demonstrative cues that achieves near-human reliability in specific MADRS domains without requiring additional training data.
- **C2:** An empirical demonstration that targeted analysis of symptom-specific interview segments generally outperforms full-transcript processing, with mean absolute error reduced from 4.90 to 3.69 across most assessment domains.
- **C3:** A comprehensive analysis revealing systematic variations in LLM performance across different MADRS items, with strongest performance in concrete symptoms and challenges in observational items.
- **C4:** A statistical analysis identifying significant factors in prediction errors, including rater effects and patient characteristics, providing insights for future clinical implementation.

2 Related Work

The intersection of **Natural Language Processing** (NLP) and mental health has emerged as a significant research domain, propelled by advancements in **Large Language Models** (LLMs). This section surveys relevant literature, highlighting key progress and identifying crucial gaps our research aims to address.

2.1 NLP in Mental Health Assessment

Over the past decade, researchers have extensively explored NLP techniques to identify and predict mental health risks through analysis of textual content and social interaction patterns. Early studies

focused on detecting indicators of mental health issues such as *anxiety* (Shen and Rudzicz, 2017; Saifullah et al., 2021; Ahmed et al., 2022), *depression* (De Choudhury et al., 2013; Eichstaedt et al., 2018; Tsugawa et al., 2015; Park et al., 2021), and *suicidal ideation* (Tadesse et al., 2019; De Choudhury et al., 2016; Coppersmith et al., 2018) by analyzing social media posts and online activity. These studies used various techniques, including content analysis and sentiment analysis, to identify linguistic markers of psychopathology (Chancellor and De Choudhury, 2020; Guntuku et al., 2017).

Although these approaches have shown promise, they often lack the nuance required for clinical applications. For instance, De Choudhury et al. (2013) and Eichstaedt et al. (2018) demonstrated high accuracy in detecting depression from social media posts, but their methods may not generalize well to diverse populations or account for cultural differences in expressing mental health concerns (Guntuku et al., 2017). In addition, a significant limitation of many online studies is the lack of gold standard or clinically verified assessments. Instead, they often rely on rough proxies such as participation in depression forums (Sadeque et al., 2016) or brief symptom self-reportsDe Choudhury et al. (2013), which may not accurately reflect clinical diagnoses.

2.2 Large Language Models in Mental Health Applications

The emergence of instruction-finetuned Large Language Models (LLMs) such as **GPT-4** (Bubeck et al., 2023), **PaLM** (Chowdhery et al., 2022), and **FLAN-T5** (Chung et al., 2024) has opened new frontiers in mental health applications. However, initial evaluations of these models revealed significant challenges. Studies by Yang et al. (2023), Lamichhane (2023), and Amin et al. (2023) assessed **ChatGPT** (GPT-3.5) on various mental health classification tasks. Their findings highlighted limitations in the model's ability to provide consistent, clinically relevant insights, emphasizing the need for cautious interpretation of LLM outputs in mental health contexts.

A comprehensive evaluation by Xu et al. (2024) examined several LLMs—including **Alpaca** (Taori et al., 2023), **FLAN-T5** (Chung et al., 2024), **LLaMA2** (Touvron et al., 2023), **GPT-3.5**, and **GPT-4**—on mental health prediction tasks using online text. This study provided a nuanced view of both the strengths and limitations of these models in mental health applications. Efforts to tailor LLMs specifically for mental health have shown promise. Ji et al. (2022) introduced **MentalBERT** and **MentalRoBERTa**, models pre-trained on mental health-related data. These specialized models outperformed existing clinical models in detecting depression and suicidal ideation from social media content. Similarly, Galatzer-Levy et al. (2023) explored the **Med-PaLM 2** model's capability to predict mental health diagnoses.

Recent studies have expanded LLM applications in mental health, focusing on interpretability and specialized tasks (Yang et al., 2024; Xu et al., 2024; Xiao et al., 2024). Questionnaire-based approaches (Rinaldi et al., 2020; Yadav et al., 2020) have shown promise, with studies employing patient self-report measures like the PHQ-9 (Rosenman et al., 2024) and the **Beck Depression Inventory** (Losada et al., 2019) for depression prediction. While valuable, these self-report measures may lack the nuanced assessment provided by trained clinicians. More recent research has used prompt engineering and LLMs to automate depression severity assessment using clinician-administered instruments like the MADRS (Raganato and Navigli, 2024), which are generally considered higher quality due to the clinician's training and ability to differentiate between similar symptoms.

Our work, LLAMADRS, advances this trajectory by applying open-source LLMs to the CAMI dataset of authentic clinical interviews based on the MADRS. By leveraging this gold-standard, clinician-administered assessment and real-world data, we aim to bridge the gap between theoretical advancements and clinical practice, pushing the field towards automated mental health assessments that maintain the rigor of expert evaluation.

3 Methodology

3.1 Dataset

We use a subset of the **Context-Adaptive Multimodal Informatics** (CAMI) dataset, which contains audio-visual recordings of clinical interviews from patients diagnosed with serious mental illness. The subset consists of 236 semi-structured interviews conducted with 140 patients (57.75% male, 40.14% female, 2.11% other; age range 19-74, mean age 41.5 years). Three trained research assistants administered these 30 minutes interviews. While the interviews incorporated multiple psychiatric assessment scales including the Positive and Negative

Syndrome Scale-6 (PANSS-6) (Kay et al., 1987) and Young Mania Rating Scale (YMRS) (Young et al., 1978), this study focuses exclusively on the Montgomery-Åsberg Depression Rating Scale (MADRS) (Montgomery and Asberg, 1979).

3.2 Data Preprocessing

The audio recordings are transcribed and diarized to separate patient and clinician speech. The preprocessing pipeline begins with WHISPERX (Bakhturina et al., 2023) for audio transcription, followed by **Qwen 2.5 - 72B Instruct** for speaker diarization to categorize utterances as patient or clinician speech. The same model is used to refine and correct transcription inaccuracies.

3.3 Interview Segmentation

For each interview transcript, **Qwen 2.5 - 72B Instruct** performs systematic classification of clinician questions according to the ten MADRS assessment domains. The model maps each question to the relevant assessment item (apparent sadness, reported sadness, inner tension, etc.). Question-response pairs are subsequently extracted and grouped by their MADRS categories, ensuring that each MADRS item assessment is conducted on precisely relevant interview segments.

3.4 MADRS Item Assessment

For each mapped interview segment, **Qwen 2.5 - 72B Instruct** generates zero-shot structured assessments comprising four elements: item score (0-6), justification, supporting utterances, and the most relevant clinical question. The assessment framework employs two complementary prompting components, implemented through a standardized prompt architecture (Figure 4 in Appendix).

Descriptive Cues: Provide MADRS-specific context, including item definitions, evaluation criteria, and standardized examples of assessment questions.

Demonstrative Cues: Present exemplar assessments for each possible score (0-6), featuring annotated clinician-patient exchanges that demonstrate score assignment rationales.

4 Experimental Setup

We run the model inference on each interview transcript individually, ensuring that there is no data leakage between examples. The model outputs the MADRS item scores, explanations, key utterances, and the most relevant questions. Each model was ran 5 times over the full data. Figure 5 in the Appendix provides a detailed illustration of the assessment process, comparing successful and problematic cases.

4.1 Baselines

For comprehensive evaluation, we implement comparative analyses across several dimensions:

Context Scope: We evaluate the efficacy of domain-specific context by comparing two approaches: (1) using mapped interview segments corresponding to individual MADRS items, and (2) processing complete interview transcripts. This comparison assesses whether targeted symptom-specific context enhances assessment precision relative to full-transcript analysis and the model's ability to identify relevant contextual segments.

Model Architecture: We conduct comparative analyses using state-of-the-art language models including LLAMA 3.1 - INSTRUCT (70B) and QWEN 2 - INSTRUCT (72B), benchmarking their performance against our primary QWEN 2.5 - INSTRUCT (72B) implementation.

Parameter Scaling: We analyze the impact of model scale using QWEN 2.5 - INSTRUCT variants (3b, 7b, 14b, 32b, 72b), examining how parameter count influences assessment accuracy and explanation coherence across MADRS domains.

Prompt Engineering: We conduct ablation studies on our assessment framework, independently evaluating the contribution of descriptive and demonstrative cues to assessment quality.

4.2 Statistical Analysis

We employed linear mixed-effects models to analyze our MADRS prediction errors. This approach accounted for the nesting of instances within patients and allowed us to statistically control for patient education and gender, as well as rater differences. We also decomposed the visit number and token count predictors into within-patient and between-patient components, which allowed us to avoid "Simpson's paradox" (Hamaker and Muthén, 2020). The model formulas are specified as:

$$
\begin{aligned}
Y_{ij} &= \beta_{0i} + \beta_1 \mathrm{V}^W_{ij} + \beta_2 \mathrm{T}^W_{ij} + \beta_3 \mathrm{R2}_{ij} \\
&\quad + \beta_4 \mathrm{R3}_{ij} + \varepsilon_{ij} \\
\beta_{0i} &= \gamma_{00} + \gamma_{01} \mathrm{V}^B_i + \gamma_{02} \mathrm{T}^B_i + \gamma_{03} \mathrm{Edu}_i \\
&\quad + \gamma_{04} \mathrm{Male}_i + \gamma_{05} \mathrm{OtherGender}_i + u_i \\
\varepsilon_{ij} &\sim \mathcal{N}(0, \sigma^2_\varepsilon), \quad u_i \sim \mathcal{N}(0, \sigma^2_u)
\end{aligned}
\tag{1}
$$

Table 1: Comprehensive Performance Metrics for MADRS Items and Scoring Methods. Classification metrics (F1 Score, Accuracy) are for a threshold of ≥ 3 for individual items and ≥ 20 for total scores. MAE = Mean Absolute Error. ICC(3,k) represents Two-way mixed, average measures, consistency. All items are scored 0–6. Total scores range from 0–60. **Bold** indicates best performance, *italic* indicates worst performance.

MADRS Item	MAE	R^2	ICC(3,k)	F1 Score	Accuracy	Class Dist.
Apparent Sadness	*0.89 ± 0.01*	*0.45 ± 0.01*	0.83 ± 0.00	0.82 ± 0.00	0.83 ± 0.00	(121, 109)
Reported Sadness	0.72 ± 0.02	0.65 ± 0.01	0.89 ± 0.01	0.87 ± 0.00	0.84 ± 0.00	(90, 140)
Inner Tension	0.55 ± 0.01	0.65 ± 0.01	0.90 ± 0.00	**0.89 ± 0.00**	0.86 ± 0.00	(76, 155)
Reduced Sleep	0.84 ± 0.01	0.47 ± 0.01	*0.83 ± 0.01*	0.80 ± 0.01	0.84 ± 0.01	(143, 88)
Reduced Appetite	**0.38 ± 0.02**	**0.77 ± 0.01**	**0.94 ± 0.00**	0.78 ± 0.01	**0.91 ± 0.00**	(183, 48)
Concent. Diffs.	0.84 ± 0.01	0.53 ± 0.01	0.86 ± 0.00	0.79 ± 0.00	0.80 ± 0.00	(118, 113)
Lassitude	0.65 ± 0.01	0.47 ± 0.02	0.86 ± 0.00	*0.76 ± 0.01*	0.84 ± 0.01	(160, 70)
Inability to Feel	0.70 ± 0.02	0.63 ± 0.01	0.89 ± 0.00	0.81 ± 0.01	0.87 ± 0.01	(153, 78)
Pess. Thoughts	0.64 ± 0.02	0.64 ± 0.01	0.90 ± 0.00	0.79 ± 0.01	0.83 ± 0.01	(141, 90)
Suic. Thoughts	0.57 ± 0.01	0.71 ± 0.01	0.91 ± 0.00	0.84 ± 0.01	0.90 ± 0.01	(156, 75)
Item-wise Scoring	3.69 ± 0.05	**0.84 ± 0.00**	**0.96 ± 0.00**	**0.90 ± 0.00**	**0.88 ± 0.00**	(109, 120)

where:

- Y_{ij}: Absolute prediction error for the LLM model on instance j of patient i.
- V^W, T^W: Within-patient components for visit number and token count (e.g., $V_{ij} - \overline{V}_i$).
- V^B, T^B: Between-patient components for visit number and token count (e.g., $\overline{V}_i$).
- R2, R3: Dummy codes for which rater scored each instance: rater 2 or rater 3 (vs. rater 1).
- Edu: Ordinal variable for patient education.
- Male, OtherGender: Dummy codes for patient gender: male or other (vs. female).
- β_1–β_4: Slopes for within-patient effects.
- γ_{00}: Fixed (or population-level) intercept.
- γ_{01}–γ_{05}: Slopes for between-patient effects.
- ε_{ij}: Level 1 residual error term.
- u_i: Random intercept deviation for patient i.

5 Results

Table 1 presents the comprehensive performance metrics for each MADRS item and the total score. The **Qwen 2.5 - 72b Instruct** model, guided by our prompting strategy, demonstrates strong correlations with clinician assessments across multiple metrics.

5.1 Impact of Context Scope

Figure 3 presents a systematic comparison of error rates between full transcript and item-specific analysis approaches. The results demonstrate consistently lower Mean Absolute Error (MAE) rates for item-specific segmented analysis across most MADRS domains, with a notable exception in the assessment of *Reported Sadness* which seems to benefit from the added context of the full transcript.

Table 2: Impact of Different Prompt Cues on MADRS Score Prediction

Prompt Var.	MAE	
	Full	Section
All Cues	**4.90 ± 0.11**	**3.69 ± 0.05**
No Descr. Cues	5.00 ± 0.12	3.62 ± 0.04
No Cues	5.40 ± 0.13	4.37 ± 0.06
No Dem. Cues	5.60 ± 0.14	3.80 ± 0.07

5.2 Model Performance

The model performs exceptionally well on certain items, particularly *Reduced Appetite* (MAE = 0.38 ± 0.02, R^2 = 0.77 ± 0.01) and *Inner Tension* (MAE = 0.55 ± 0.01, R^2 = 0.65 ± 0.01). Conversely, items like *Apparent Sadness* (MAE = 0.89 ± 0.01, R^2 =

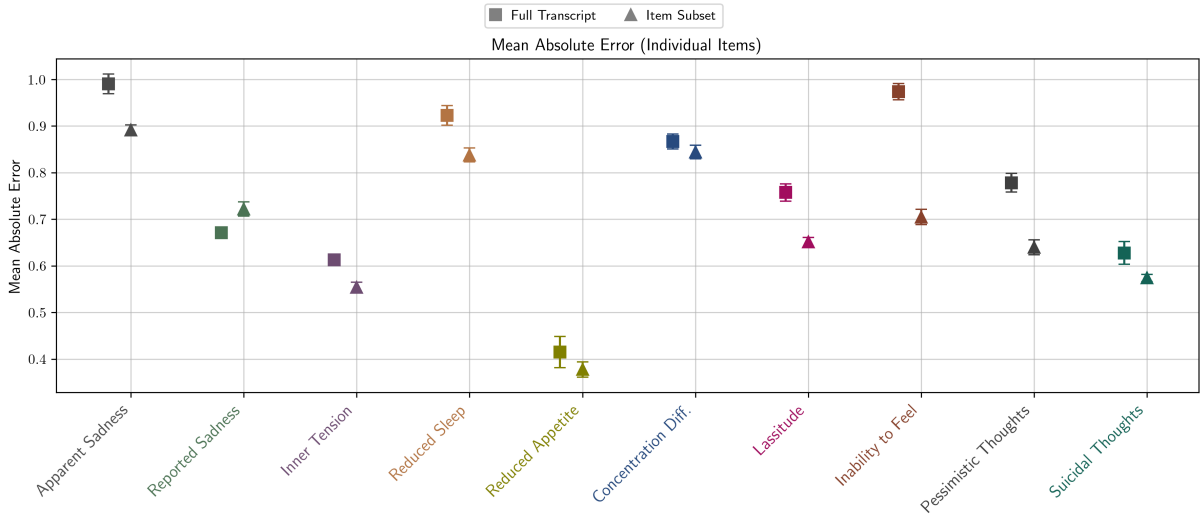


Figure 3: Mean absolute error (MAE) comparison between full transcript and item-specific context analysis across MADRS items, with standard error bars (n=150). Item-specific processing demonstrates reduced error rates relative to full-transcript analysis ($p < 0.01$).

Table 3: Performance Comparison of Large Language Models for MADRS Total Score Prediction

	Cont. Len.	MAE		R^2	
		Full	Segmented	Full	Segmented
Qwen 2.5 Inst. (72B)	**128K**	**4.90 ± 0.11**	**3.69 ± 0.05**	**0.69 ± 0.03**	**0.84 ± 0.00**
Llama 3.1 Inst. (70B)	128K	6.12 ± 0.17	4.86 ± 0.18	0.54 ± 0.05	0.74 ± 0.03
Qwen 2 Inst. (72B)	128K	7.10 ± 0.82	4.40 ± 0.20	0.40 ± 0.16	0.78 ± 0.03
Qwen 2.5 Inst. (32B)	128K	15.55 ± 0.20	3.52 ± 0.17	−3.74 ± 0.08	0.85 ± 0.02
Qwen 2.5 Inst. (14B)	128K	15.61 ± 0.19	3.62 ± 0.17	−3.80 ± 0.05	0.84 ± 0.03
Qwen 2.5 Inst. (7B)	128K	17.36 ± 0.22	4.47 ± 0.19	−4.05 ± 0.04	0.77 ± 0.03
Qwen 2.5 Inst. (3B)	32K	19.40 ± 0.24	7.03 ± 0.21	−4.45 ± 0.06	0.50 ± 0.04
Llama 3.1 Inst. (8B)	128K	19.42 ± 0.17	9.96 ± 0.22	−2.27 ± 0.08	0.06 ± 0.04

Note: MADRS = Montgomery-Åsberg Depression Rating Scale; MAE = Mean Absolute Error. Model parameters (B) are shown in billions. Best performing metrics are highlighted in **bold**. Negative R^2 values indicate poor model fit relative to baseline.

0.45 ± 0.01) and *Reduced Sleep* (MAE = 0.84 ± 0.01, R^2 = 0.47 ± 0.01) show higher error rates and lower correlation with clinician ratings. For the MADRS total score, the **Item-wise** method achieves an MAE of 3.69 ± 0.05 and an R^2 of 0.84 ± 0.00, as shown in Table 1. This performance demonstrates strong correlation with clinician assessments and robust predictive capability. In the next subsections, we use the **item-wise** MADRS prediction as our primary metric for cross-model comparisons and ablation studies.

5.3 Impact of Different Prompt Cues

Table 2 presents comprehensive ablation studies examining each cue type's contribution across both segmented and full transcript analyses. When analyzing full transcripts, removing **Demonstrative Cues** causes the largest performance degradation (MAE = 5.60 ± 0.14), while the absence of descriptive cues shows a more modest impact (MAE = 5.00 ± 0.12). For segmented analysis, the pattern persists but with lower overall error rates: removing demonstrative cues yields MAE = 3.80 ± 0.07, while removing descriptive cues results in MAE = 3.62 ± 0.04. The optimal performance is achieved with all cues present, yielding MAE = 4.90 ± 0.11 for full transcripts and MAE = 3.69 ± 0.05 for segmented analysis.

Table 4: Comparison of ICC values for MADRS items between our study and Iannuzzo et al. (2006)

MADRS Item	Our ICC	Human ICC
MADRS total	0.96	0.98
Appar. sadness	0.83	0.92
Repor. sadness	0.89	0.94
Inner tension	0.90	0.92
Red. sleep	0.83	0.86
Red. appetite	0.94	0.94
Concentration	0.86	0.90
Lassitude	0.86	0.90
Inabil. to feel	0.89	0.94
Pessim. thoughts	0.90	0.93
Suicid. thoughts	0.91	0.97

5.4 Comparison with Other Models

As detailed in Table 3, **Qwen 2.5 - Instruct (72B)** demonstrates superior performance across both analysis approaches. For full transcript analysis, it achieves MAE = 4.90 ± 0.11 and R^2 = 0.69 ± 0.03, while segmented analysis yields improved results with MAE = 3.69 ± 0.05 and R^2 = 0.84 ± 0.00. **Llama 3.1 - Instruct (70B)** achieves full transcript performance of MAE = 6.12 ± 0.17 and R^2 = 0.54 ± 0.05, and segmented analysis results of MAE = 4.86 ± 0.18 and R^2 = 0.74 ± 0.03. Both models leverage a **128K** token context length, with segmented analysis consistently outperforming full transcript analysis across both models.

5.5 Model Size and Performance

As evidenced in Table 3, model performance scales with parameter count. Within the **Qwen 2.5** family, models below 32 billion parameters exhibit markedly degraded performance in full transcript analysis (R^2 = -4.45 ± 0.06 for 3B variant). The smallest architectures demonstrate the poorest metrics, with **Qwen 2.5 - Instruct (3B)** and **Llama 3.1 - Instruct (8B)** yielding MAE values of 19.40 ± 0.24 and 19.42 ± 0.17 respectively.

While segmented analysis partially mitigates these deficits (72B: MAE = 3.69 ± 0.05; 3B: MAE = 7.03 ± 0.21), the performance gap between full transcript and segmented analysis narrows with increased model size, suggesting enhanced capacity for managing extended clinical narratives in larger models. Parameter scaling also correlates with prediction stability, evidenced by decreasing standard deviations in performance metrics.

5.6 Near-Human Level Agreement on MADRS Ratings

Table 4 demonstrates the model's Intraclass Correlation Coefficient (ICC) across MADRS items, revealing a noteworthy pattern when compared to inter-rater reliability benchmarks from Iannuzzo et al. (2006). The model achieves exemplary agreement on *Reduced Appetite* (ICC 0.94 vs. 0.94) and strong performance on *Reported Sadness* (ICC 0.89 vs. 0.94). Notably, as the ICC between our model and research assistants varies across different items, similar variations are observed in the human inter-rater reliability scores from Iannuzzo et al. (2006), with the exception of *Inner Tension* assessment.

These findings underscore a fundamental challenge in psychiatric assessment: the absence of an absolute ground truth against which to measure performance. Disagreements between the model and research assistants may reflect not only algorithmic limitations but also the inherent subjectivity in clinical assessment—a challenge that similarly affects human rater concordance. Despite these measurement challenges, the model achieves both high absolute performance metrics and strong correlation with patterns of human inter-rater reliability, suggesting robust and clinically relevant assessment capabilities.

5.7 Error Analysis

Linear mixed-effects models identified significant predictors of MADRS assessment errors (Table 5). **Rater identity** emerged as a primary predictor of error magnitude, with Rater R2's assessments associated with increased prediction errors for *Inner Tension* (0.81), *Pessimistic Thoughts* (0.85), and *Concentration Difficulties* (0.73), while showing decreased errors for *Reported Sadness* (-0.62). Similarly, Rater R1's assessments corresponded to higher prediction errors across *Inner Tension* (0.57), *Concentration Difficulties* (0.71), *Reduced Appetite* (0.56), and *Lassitude* (0.40). The analysis revealed that higher **between-patient visit numbers** corresponded to reduced errors in *Inability to Feel* assessment (-0.20). Patient characteristics also influenced error patterns: higher **education levels** corresponded to increased errors in *Concen-*

Table 5: Feature Importance Analysis for MADRS Items

MADRS Item	V^B	Edu	Age
Reported Sadness	–	–	0.12
Inner Tension	–	–	–
Reduced Appetite	–	–	–
Concentration Difficulties	–	0.26	–
Lassitude	–	–	–
Inability to Feel	-0.20	–	–

MADRS Item	R1	R2
Reported Sadness	–	-0.62
Inner Tension	0.57	0.81
Reduced Appetite	0.56	–
Concentration Difficulties	0.71	0.73
Lassitude	0.40	–
Pessimistic Thoughts	–	0.85

Note: V^B: Visit Number (Between-Patient), Edu: Education Level, R1: Rater 1, R2: Rater 2. Values indicate feature importance coefficients. '–' indicates non-significant coefficients.

tration Difficulties (0.26), while increased **age** was associated with higher errors in *Reported Sadness* assessment (0.12).

6 Discussion

Our comprehensive analysis of LLAMADRS reveals crucial insights into the potential and limitations of LLM-based depression assessment. We structure our discussion around key empirical findings and their implications for clinical applications.

6.1 Performance Analysis

The superior performance of **Qwen 2.5 - 72B** in segmented analysis (MAE = 3.69 ± 0.05, R^2 = 0.84 ± 0.00) demonstrates the viability of LLM-based depression assessment. This performance level, approaching human rater reliability (ICC = 0.94 for *Reduced Appetite*), suggests potential clinical utility. However, the degraded performance in full transcript analysis (MAE = 4.90 ± 0.11) highlights the importance of structured input processing.

6.2 Architectural and Methodological Insights

Our ablation studies reveal two critical findings. First, the significant impact of demonstrative cues (MAE increase to 5.60 ± 0.14 without them) underscores the importance of example-based guidance in clinical assessment tasks. Second, the clear scaling relationship with model size, particularly in full transcript analysis, suggests that increased parameter count enhances both raw performance and context management capabilities.

6.3 Clinical Assessment Patterns

Performance variation across MADRS items reveals systematic patterns. Strong performance on concrete symptoms (e.g., *Reduced Appetite*, MAE = 0.38 ± 0.02) contrasts with challenges in assessing subjective states (e.g., *Apparent Sadness*, MAE = 0.89 ± 0.01). This pattern aligns with clinical intuition: concrete symptoms typically have clearer linguistic markers and more consistent reporting patterns.

6.4 Implementation Considerations

The substantial performance gap between segmented and full transcript analysis suggests practical implementation strategies. While larger models demonstrate enhanced capability for processing complete interviews, the superior performance of segmented analysis indicates that structured input processing remains beneficial across all model scales. This finding has direct implications for clinical deployment, suggesting a hybrid approach that combines automated segmentation with focused assessment.

7 Conclusion

This study establishes LLAMADRS as a viable framework for automated depression severity assessment using open-source **Large Language Models**. Through systematic evaluation on 236 real-world clinical interviews, we demonstrate that carefully engineered prompting strategies enable **Qwen 2.5–72b** to achieve near-human reliability in specific MADRS domains. The superior performance in concrete symptom assessment validates the potential of LLM-based approaches for clinical applications. Our comprehensive analysis reveals that segmented processing consistently outperforms full-transcript analysis, highlighting the importance of structured input handling in clinical assessments. The clear relationship between model scale and performance, particularly in managing extended clinical narratives, provides crucial insights for future work on this topic.

8 Limitations

Our study faces several key limitations in its current form. The reliance on transcribed text data omits important non-verbal cues crucial for assessing symptoms like Apparent Sadness, where visual and auditory signals play vital roles. Our dataset's focus on inpatient settings may limit generalizability to other contexts. Additionally, the computational requirements of our best-performing models may restrict implementation in resource-constrained settings. Finally, the model may miss subtle clinical nuances that experienced human raters might catch, particularly in complex cases.

9 Ethical Considerations

The deployment of AI systems for mental health assessment requires careful ethical consideration. Our system is designed to support, not replace, clinical decision-making, with final decisions remaining with qualified healthcare professionals. Patient privacy and informed consent are paramount, requiring robust data protection measures and clear communication about the system's role and limitations. While this technology could increase access to mental health assessment in resource-limited settings, care must be taken to ensure it doesn't exacerbate healthcare disparities. Extensive validation across diverse populations remains necessary before clinical deployment.

References

Arfan Ahmed, Sarah Aziz, Carla T Toro, Mahmood Alzubaidi, Sara Irshaidat, Hashem Abu Serhan, Alaa A Abd-Alrazaq, and Mowafa Househ. 2022. Machine learning models to detect anxiety and depression through social media: A scoping review. *Computer Methods and Programs in Biomedicine Update*, page 100066.

Mostafa M Amin, Erik Cambria, and Björn W Schuller. 2023. Will affective computing emerge from foundation models and general ai? a first evaluation on chatgpt. *arXiv preprint arXiv:2303.03186*.

Ekaterina Bakhturina, Chun-Liang Han, and Yulia Tsvetkov. 2023. Whisper-x: Improving speech recognition with visual context. *arXiv preprint arXiv:2303.08288*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):43.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1):128–137.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208.

Isaac R Galatzer-Levy, Daniel McDuff, Vishnu Natarajan, Alan Karthikesalingam, and Matteo Malgaroli. 2023. The capability of large language models to measure psychiatric functioning. *arXiv preprint arXiv:2308.01341*.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Ellen L. Hamaker and Bengt Muthén. 2020. The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological Methods*, 25(3):365–379.

Rebecca W. Iannuzzo, Judith Jaeger, Joseph F. Goldberg, Vivian Kafantaris, and M. Elizabeth Sublette. 2006. Development and reliability of the HAM-D/MADRS Interview: An integrated depression symptom rating scale. *Psychiatry Research*, 145(1):21–37.

Institute for Health Metrics and Evaluation. 2023. Global Health Data Exchange (GHDx). Accessed: 2023-03-04.

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

Stanley R Kay, Abraham Fiszbein, and Lewis A Opler. 1987. The positive and negative syndrome scale (panss) for schizophrenia. *Schizophrenia Bulletin*, 13(2):261–276.

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

D.E. Losada, F. Crestani, and J. Parapar. 2019. Overview of erisk 2019 early risk prediction on the internet. In *Proceedings of the 10th Conference and Labs of the Evaluation Forum (CLEF'19)*.

Stuart A Montgomery and Marie Asberg. 1979. A new depression scale designed to be sensitive to change. *The British journal of psychiatry*, 134(4):382–389.

Minsu Park, David W McDonald, and Meeyoung Cha. 2021. Perception differences between the depressed and non-depressed users in twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1):476–485.

Alessandro Raganato and Roberto Navigli. 2024. everaging prompt engineering and large language models for automating madrs score computation for depression severity assessment. *arXiv preprint arXiv:2309.10975*.

A. Rinaldi, J. Benito, and V. Soler. 2020. Predicting depression through tweets using sentiment analysis. In *CEUR Workshop Proceedings*.

Gony Rosenman, Lior Wolf, and Talma Hendler. 2024. LLM Questionnaire Completion for Automatic Psychiatric Assessment. *arXiv preprint*. ArXiv:2406.06636.

Farig Sadeque, Ted Pedersen, Thamar Solorio, Prasha Shrestha, Nicolas Rey-Villamizar, and Steven Bethard. 2016. Why do they leave: Modeling participation in online depression forums. In *Proceedings of the fourth international workshop on natural language processing for social media*, pages 14–19.

Shoffan Saifullah, Yuli Fauziah, and Agus Sasmito Aribowo. 2021. Comparison of machine learning for sentiment analysis in detecting anxiety based on social media data. *arXiv preprint arXiv:2101.06353*.

Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting anxiety through reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, pages 58–65.

Michael Mesfin Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of suicide ideation in social media forums using deep learning. *Algorithms*, 13(1):7.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

John Torous, Sandra Bucci, Imogen H. Bell, Lars V. Kessing, Maria Faurholt-Jepsen, Pauline Whelan, Andre F. Carvalho, Matcheri Keshavan, Jake Linardon, and Joseph Firth. 2021. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry*, 20(3):318–335.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint*. ArXiv:2307.09288.

Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. 2015. Recognizing depression from twitter activity. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 3187–3196.

Mengxi Xiao, Qianqian Xie, Ziyan Kuang, Zhicheng Liu, Kailai Yang, Min Peng, Weiguang Han, and Jimin Huang. 2024. HealMe: Harnessing cognitive reframing in large language models for psychotherapy. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1707–1725, Bangkok, Thailand. Association for Computational Linguistics.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(1).

Amisha Yadav, Ilvira Alimova, Gaurav Bajaj, and Rajesh Pathak. 2020. Identifying depression in the era of social media: A sentiment analysis approach. *International Journal of Interactive Multimedia and Artificial Intelligence*, 6(2):113–118.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, and Sophia Ananiadou. 2023. On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis. *arXiv preprint arXiv:2304.03347*.

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

RC Young, JT Biggs, VE Ziegler, and DA Meyer. 1978. A rating scale for mania: reliability, validity and sensitivity. *The British Journal of Psychiatry*, 133(5):429–435.

A Methodological Details

B Computational Details

All experiments were conducted using NVIDIA A100 GPUs with 80GB memory.

C Model Details

C.1 Architecture and Parameters

The Qwen 2.5 architecture implements a decoder-only transformer with approximately 72 billion parameters (72B). We utilized the Qwen-72B-Instruct variant, which has undergone instruction-tuning to enhance its performance on natural language instructions and multi-turn conversations.

C.2 Implementation and Inference

The model was accessed and deployed through the Hugging Face Transformers library (version 4.36.2) and the VLLM inference framework, which enables efficient serving through automatic quantization and optimization techniques. We maintained consistency across our experimental framework by utilizing the same library stack for all baseline models in our comparative analysis.

C.3 Generation Parameters

For all inference tasks, we employed a standardized configuration with the following hyperparameters:

- Temperature (τ) = 0.6, controlling the randomness in the output distribution
- Top-p (nucleus sampling) = 0.9, limiting the cumulative probability mass for token selection
- Maximum sequence length = 500 tokens
- 4-bit quantization

This configuration was selected to balance output quality and diversity while maintaining reproducibility across experimental conditions. The relatively conservative temperature setting of 0.6 helps maintain coherent outputs while allowing for sufficient variability in generated responses.

MADRS Assessment Prompt Structure

Task Description:

Analyze a diarized transcript of a psychiatric session where the Montgomery-Åsberg Depression Rating Scale (MADRS) questionnaire is being administered. Predict the rating (0-6) that the practitioner would likely give for the specified MADRS item based on the patient's responses and the conversation.

MADRS Item Components:

- **Item Name:** Reported Sadness
- **Description:** Represents reports of depressed mood, regardless of whether it is reflected in appearance or not. Includes low spirits, despondency or the feeling of being beyond help and without hope.
- **Key Questions:**
  - In the past week, have you been feeling sad or unhappy?
  - Does the feeling lift at all if something good happens?
  - How much of each day? How many days this week?

Rating Scale (0-6):

- **0:** Occasional sadness in keeping with circumstances
- **2:** Sad or low but brightens up without difficulty
- **4:** Pervasive feelings of sadness or gloominess
- **6:** Continuous or unvarying sadness, misery

– (Odd numbers represent intermediate states)

Required Output Format:

```
Rating: [0-6]
Explanation: [2-3 sentences]
Key Utterances: [relevant lines]
Most Relevant Question: [from transcript]
```

Figure 4: **Structured Prompt for MADRS Assessment.** The prompt provides comprehensive guidance for analyzing psychiatric interview transcripts and assigning depression severity ratings. It includes the core components: task description, item definition, standardized questions, rating scale definitions, and required output format. This structure ensures consistent assessment across different raters and maintains compatibility with clinical standards.

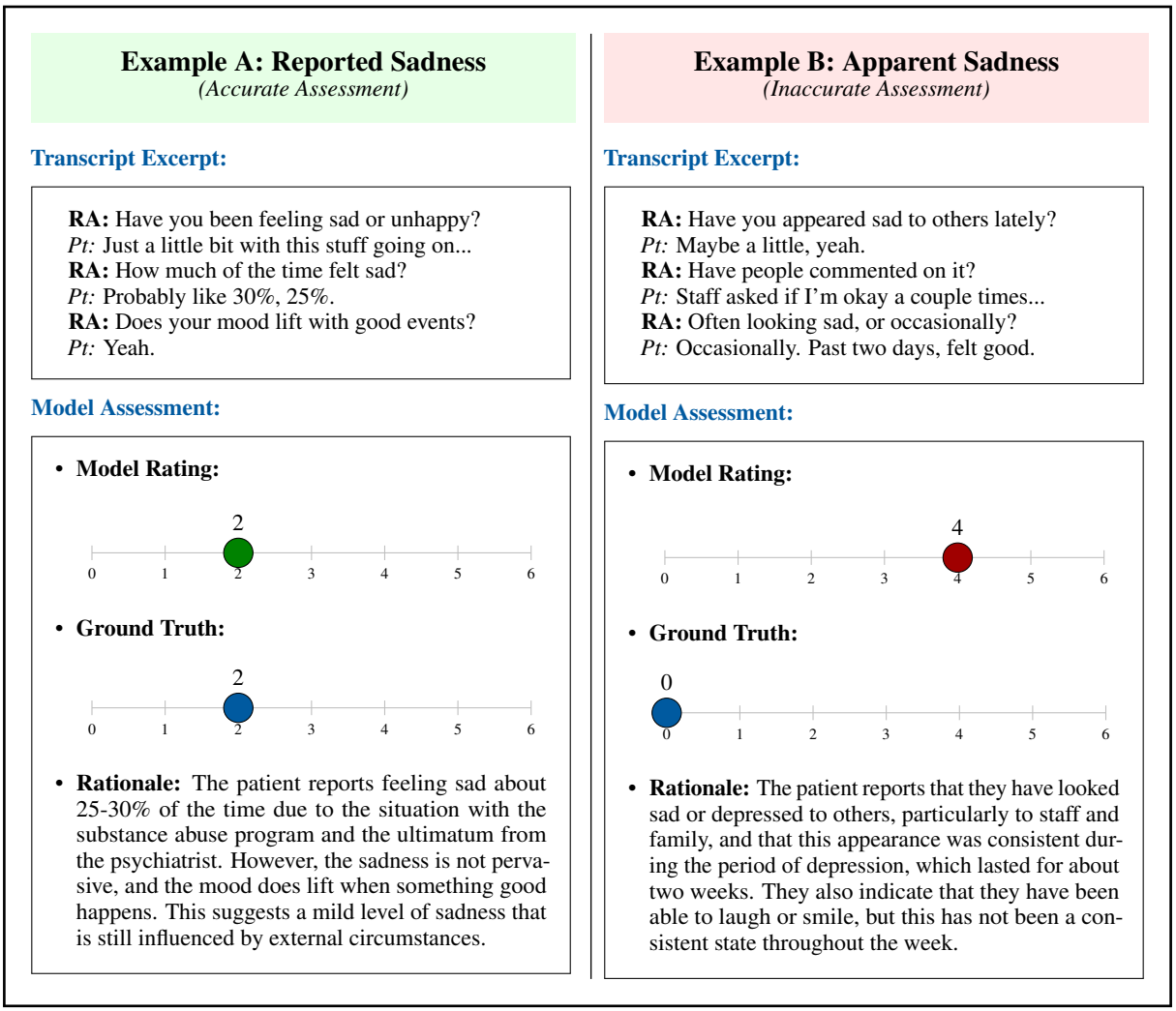


Figure 5: **Representative Examples of LLAMADRS Assessment Performance.** Comparison of two cases demonstrating the model's varying capability in MADRS item scoring. Example A shows accurate interpretation of reported sadness, matching the ground truth score of 2/6. Example B reveals a significant deviation from ground truth (4/6 vs 0/6), highlighting challenges in interpreting qualitative responses and temporal context for apparent sadness assessment.