LLAMADRS: Evaluating Open-Source LLMs on Real Clinical Interviews—To Reason or Not to Reason?

Gaoussou Youssouf Kebe[1] **Jeffrey M. Girard**[2] **Einat Liebenthal**[3],
Justin Baker[3] **Fernando De la Torre**[1] **Louis-Philippe Morency**[1]
[1]Carnegie Mellon University, School of Computer Science, Pittsburgh, PA, USA
[2]University of Kansas, Department of Psychology, Lawrence, KS, USA
[3]McLean Hospital, Harvard Medical School, Boston, MA, USA
{gyk,ftorre,morency}@cs.cmu.edu
jmgirard@ku.edu
{eliebenthal,jtbaker}@partners.org

Abstract

Large language models (LLMs) excel on many NLP benchmarks, but their behavior on real-world, semi-structured prediction remains underexplored. We present LLAMADRS, a benchmark for structured clinical assessment from dialogue built on the CAMI corpus of psychiatric interviews, comprising 5,804 expert annotations across 541 sessions. We evaluate 25 open-source models (standard and reasoning-augmented; 0.6B–400B parameters) and generate over 400,000 predictions. Our results demonstrate that strong open-source LLMs achieve item-level accuracy with residual error below clinically substantial thresholds. Additionally, an Item-then-Sum (ITS) strategy, assessing symptoms individually through discrete LLM calls before synthesizing final scores, significantly reduces error relative to Direct Total Score (DTS) prediction across most model architectures and scales, despite reasoning models attempting similar decomposition in the reasoning traces of their DTS predictions. In fact, we find that performance gains attributed to "reasoning" depend fundamentally on prompt design: standard models equipped with structured task definitions and examples match reasoning-augmented counterparts. Among the latter, longer reasoning traces correlate with reduced error; while higher model scale does across both architectures. Our results clarify when and why reasoning helps and offer actionable guidance for deploying LLMs in semi-structured clinical assessment.

1 Introduction

Mental health disorders are a leading cause of disability worldwide and a major public health challenge. Even in highly developed nations, a shortage of trained clinicians leaves many individuals without timely care: more than half of the U.S. population lives in designated mental health professional shortage areas, hindering intervention and exacerbating disparities (Nguyen et al., 2025). Depression, characterized by persistent low mood and anhedonia, remains highly prevalent. The Montgomery–Åsberg Depression Rating Scale (MADRS) is a clinician-administered instrument comprising ten symptom domains, each rated on a 0–6 scale (Montgomery and Åsberg, 1979).

Large language models have achieved state-of-the-art performance on diverse natural language tasks, but their alignment with mental health competencies remains underexplored (Na et al., 2025). Emerging analyses reveal that LLMs often rely on surface-level pattern recognition rather than genuine reasoning, leading to brittleness on tasks requiring structured inference (Jin et al., 2025). Reasoning augmentation can improve performance on arithmetic or logical puzzles, but benefits are inconsistent and, in some settings, longer reasoning traces *decrease* accuracy (Jin et al., 2025). In mental health domains, reasoning-driven prompting has been proposed to enhance classification accuracy (Teng et al., 2025), yet improvements are modest and dataset-specific: chain-of-thought and related strategies yield notable gains on some tasks while failing to generalize to others (Patil and Gedhu, 2025). These observations motivate our systematic comparison of reasoning-augmented and standard LLMs on clinical interviews.

Our contributions advance both computational psychiatry and natural language processing:

1. **LLAMADRS**, a benchmark of 5,804 MADRS assessments across 541 real patient–clinician interview sessions from the CAMI corpus.
2. **Clinically reliable accuracy**: Strong LLMs achieve item-level Mean Absolute Error (MAE)

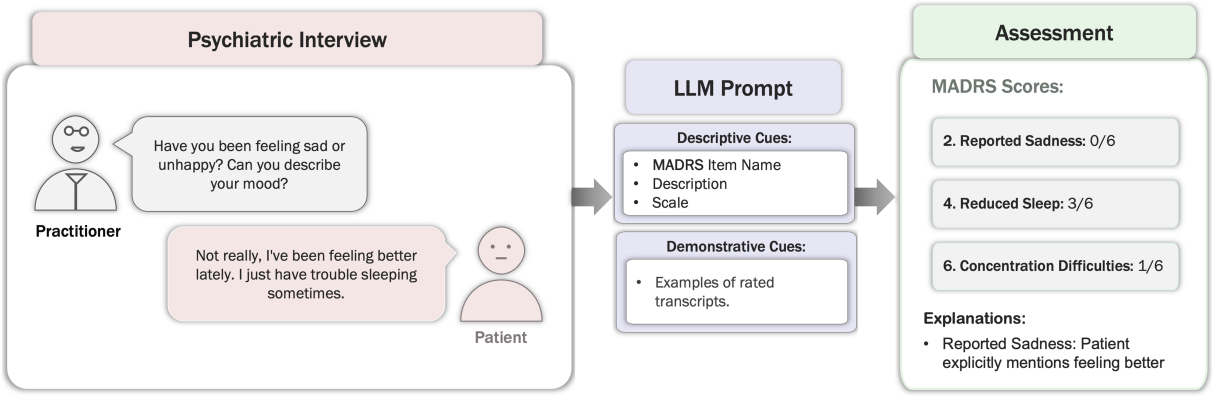

Figure 1: **Overview of the LLAMADRS framework.** Left: a structured clinical interview between a patient and clinician. Right: automated depression assessment using a large language model, including scoring of MADRS items with item-wise explanations.

in the moderate range (0.6–1.2), with total-scale Mean Absolute Error (MAE) in the clinically acceptable range (< 6) under Item-then-Sum (ITS) scoring.

3. **When reasoning helps**: Explicit reasoning helps when scaffolding is sparse, but provides limited benefit under well-structured prompts with clinical descriptive cues, where standard models frequently match or surpass reasoning-augmented variants.
4. **Key performance predictors**: Model scale reliably improves accuracy across both reasoning and non-reasoning architectures, whereas longer reasoning traces appear to benefit reasoning-augmented models.

2 Related Work

2.1 NLP and Mental Health

Early applications of NLP for mental health focused on detecting depression, anxiety, and suicidality in social media data using bag-of-words, topic models, and early transformers (Coppersmith et al., 2014; De Choudhury et al., 2013; Eichstaedt et al., 2018; Shen and Rudzicz, 2017; Ji et al., 2022). With LLMs, research broadened from binary screening to multi-label symptom detection, information extraction, and explanation generation across single posts, user histories, and dialogues (Yang et al., 2023, 2024; Xu et al., 2024; Bao et al., 2024; Raihan et al., 2024; Mohammadi et al., 2024; Schirmer et al., 2024; Skianis et al., 2024). Clinical-facing efforts include multi-turn interview analysis for PTSD and broader symptom delineation/summarization (Tu et al., 2024; So et al., 2024), alongside severity estimation via taxonomy-aligned summaries (e.g., BDI) (Aragón et al., 2024; Wang et al., 2024b). NLP-based evaluations of clinician competencies have also tried to detect gaps in empathy, cultural sensitivity, and ethics, with results underscoring the need for clinically grounded tasks and item-level analyses rather than aggregate labels (Nguyen et al., 2025).

2.2 Reasoning-Augmented LLMs in Mental Health

Debates persist on whether LLMs *reason* or perform sophisticated pattern matching. Evidence shows that explicit chain-of-thought (CoT), self-consistency, and related strategies yield mixed and dataset-specific gains; longer reasoning traces may even degrade accuracy (Jin et al., 2025; Teng et al., 2025; Patil and Gedhu, 2025). In psychotherapy and mental-health NLP, prompting styles (emotion prompting, CoT, role prompting, multi-agent debate) help interpretability or specific subtasks but do not guarantee robust generalization across settings or constructs (Yang et al., 2023; Lim et al., 2024; Singh et al., 2024; Uluslu et al., 2024). These results motivate head-to-head evaluation of reasoning-augmented vs. standard open LLMs under clinically moderate objectives.

2.3 LLMs for Clinical Assessment and Severity Scoring

A stream of work examines LLMs for clinical assessment and severity estimation from interviews or consolidated text. Med-PaLM 2 showed zero-shot alignment with clinician ratings for depression severity (and limited generalization to PTSD) (Galatzer-Levy et al., 2023). On DAIC-WOZ and related corpora, general LLMs exhibit

Table 1: Comparison of LLM-based psychiatric assessment works.

Study	Models	Dataset / N	Population
This work	25 **open-source (R & NR)**	**CAMI /** 541	**Clinician interviews; inpatients**
Tu et al. (2024)	GPT-4, Llama-2	Clinical interviews / 411	Trauma interviews
So et al. (2024)	GPT-4 Turbo; GPT-3.5 (FT)	Psychiatric interviews (KR)	Clinical
Galatzer-Levy et al. (2023)	Med-PaLM 2 (ZS)	Clinical descriptions	Screening
Arcan et al. (2024)	ChatGPT, Llama-2 (ZS)	DAIC-WOZ, etc.	Community/elicited
Aragón et al. (2024)	GPT-3.5/4 (ZS)	Social media (per-user)	General population
Yang et al. (2023)	GPT-3.5/LLaMA	11 social-media sets	General population
Xu et al. (2024)	Inst.-tuned LLM (FT)	Online text (various)	General population
Skianis et al. (2024)	LLMs (ZS)	6 languages (new)	Cross-lingual social media

moderate text-regression performance relative to specialized transformers (Arcan et al., 2024).

Compared to prior work that focuses on social media, synthetic datasets, or aggregate screening (Tu et al., 2024; So et al., 2024; Galatzer-Levy et al., 2023; Arcan et al., 2024; Yang et al., 2023; Xu et al., 2024; Aragón et al., 2024; Skianis et al., 2024), we offer (1) *item-level and total* MADRS scoring over *real* clinician–patient interviews; (2) a broad open-source model comparison spanning architectures, scales, and reasoning styles; and (3) clinically grounded error analyses and prompt ablations disentangling descriptive cues, demonstrative cues, and reasoning augmentation.

3 Dataset: CAMI

The Context-Adaptive Multimodal Informatics (CAMI) dataset (Culhane et al., 2023) comprises authentic patient–clinician dialogues from inpatient psychiatric care settings. All participants provided informed consent under IRB-approved protocols. Each semi-structured session lasted approximately thirty minutes during which trained raters administered the MADRS. The dataset contains 541 interviews from 277 adult patients, totaling 5,804 item-level and total-score MADRS ratings.

MADRS annotations. Each session was rated by a single clinically trained research assistant on the ten MADRS items: (I1) Apparent Sadness, (I2) Reported Sadness, (I3) Inner Tension, (I4) Reduced Sleep, (I5) Reduced Appetite, (I6) Concentration Difficulties, (I7) Lassitude, (I8) Inability to Feel, (I9) Pessimistic Thoughts, and (I10) Suicidal Thoughts. These clinical research assistants were trained to administer and score the MADRS through structured supervision, calibration, and review of recorded semi-structured interviews using standardized rating guidelines. Before independent rating, all raters were required to achieve item-level Cohen's $\kappa \geq 0.80$ on separate calibration data (Gillis et al., 2023). Each item is scored on a seven-point scale from 0 (absent) to 6 (severe), where even anchor points carry specific behavioral descriptions: 0 = no symptoms, 2 = mild, 4 = moderate, 6 = severe. Odd scores (1, 3, 5) denote intermediate severity levels between adjacent anchors (e.g., 1 = between absent and mild; 3 = between mild and moderate; 5 = between moderate and severe). The ten item scores are summed to produce a total score ranging from 0 to 60, interpreted as: absent (0–6), mild (7–19), moderate (20–34), and severe depression (35–60) (Montgomery and Åsberg, 1979). **Figure 2** shows the score distribution across all ten items alongside the total-score severity bands.

Audio processing. Transcriptions were generated using both Whisper (Radford et al., 2023) and Parakeet (Sekoyan et al., 2025) ASR systems, then processed with Qwen 3-32B (Wang et al., 2024a) for diarization to assign utterances to speakers. For item-level prediction, Qwen 3-32B was used to segment interviews by identifying question–response exchanges corresponding to specific MADRS items.

Availability. All experiments were conducted on the original, non-de-identified transcripts under the IRB-approved protocol governing the CAMI dataset. For external access, de-identified versions of the transcripts, clinician ratings, and session-level metadata will be available under controlled access.[1] Access is limited to credentialed researchers who agree to a Data Use Agreement.

[1]For data access requests, please contact Justin Baker, McLean Hospital, Harvard Medical School, Boston, MA, USA, at jtbaker@mgb.org. Code is available at https://github.com/llamadrs/llamadrs.

4 Approach

We evaluate 25 open-source LLMs spanning eight architectural families and parameter counts from 0.6 billion to 400 billion (dense or mixture-of-experts variants).

4.1 Prompting Framework

Each prompt comprises three components: (1) a task description with clinical definitions and severity anchors (descriptive cues); (2) annotated interview excerpts providing rating rationale examples (demonstrative cues); and (3) JSON schema-enforced output requirements. Prompt templates are provided in **Appendix Section A**.

For item-level predictions, we provide the relevant MADRS item definition and severity anchors (as described in **Section 3**), along with the corresponding transcript segment for that item.

For total-score prediction, we provide an overview of all ten MADRS items and refer to the standard total-score interpretation, also defined in **Section 3**, then ask the model to output a single total score. The complete MADRS interview transcript is used for total score prediction.

4.2 Evaluation Protocol

Error metric. We adopt Mean Absolute Error (MAE) as our primary evaluation metric, defined as $\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$, where $\hat{y}_i$ and y_i are predicted and gold-standard ratings, respectively. MAE is preferred over squared-error metrics because it is directly interpretable in the original rating units (i.e., a MAE of 1.0 corresponds to an average misprediction of one scale point) and aligns naturally with clinical minimal-change thresholds (Turkoz et al., 2021).

Seed replication and averaging. For each model–session pair, we generate 3 independent predictions using different random seeds. MAE is computed per seed, then averaged across the three seeds to produce the reported seed-averaged MAE. Cross-seed standard deviations are shown as $\pm$ values throughout Tables 2–3.

Sum-based aggregation. We evaluate two strategies for obtaining total MADRS scores. **Direct total score (DTS)**: the model receives the full interview transcript and predicts the total score (0–60) in a single inference call. **Item-then-sum (ITS)**: the model scores each of the ten items independently via separate inference calls, and the total is computed by mechanically summing the ten item-level predictions, i.e., $\hat{T}_{\text{ItS}} = \sum_{i=1}^{10}\hat{y}_i$, with no model involvement in the summation step.

Clinical error thresholds. We adopt thresholds from Turkoz et al. (2021) to interpret prediction errors: a less than 6-point change on the MADRS total scale (0–60) is "clinically acceptable," while more than 12 points is "clinically substantial." We repurpose these as error evaluation anchors, defining three bands: **clinically acceptable** (MAE < 6 total, < 0.6 per item), **moderate error** (MAE 6–12 total, 0.6–1.2 per item), and **substantial error** (MAE ≥ 12 total, ≥ 1.2 per item). Errors below 6 points are unlikely to affect clinical interpretation, while errors exceeding 12 points may alter treatment decisions.

Handling invalid outputs. Predictions may fail due to (a) unparseable JSON output (e.g., malformed syntax), (b) absent or out-of-range rating fields (e.g., returning a rating of 7 or omitting the score entirely), or (c) generation failures (e.g., incomplete outputs due to context-length truncation). For such cases, we apply a conservative backoff: we assign the maximum absolute error (6 item-level; 60 total), so models are not rewarded for invalid or selectively abstained outputs.

Screening evaluation. For screening, we binarize item scores at ≥ 3 and total scores at ≥ 20, reporting F_1 as the primary metric. The total-score cutoff of ≥ 20 corresponds to the standard MADRS mild–moderate boundary (Montgomery and Åsberg, 1979) and produces a near-balanced split in CAMI (see **Figure 2**): $273/541$ (50.5%) < 20 vs. $268/541$ (49.5%) ≥ 20.

4.3 Prompt Ablation

Building on the prompting framework in **Section 4.1**—(1) clinical task and severity descriptive cues, (2) annotated demonstrative cues, and (3) schema-enforced outputs—we ablate the first two components to assess whether descriptive cues and demonstrative cues differentially benefit LLMs. In parallel, we run a paired, item-level comparison of MADRS performance between the reasoning-augmented and standard variants of **Qwen 3 Next (80B; MoE)** across four configurations: **All** (descriptive cues + demonstrative cues), **No Descriptions** (demonstrative cues only), **No Demonstrations** (descriptive cues only), and **Raw** (no cues).

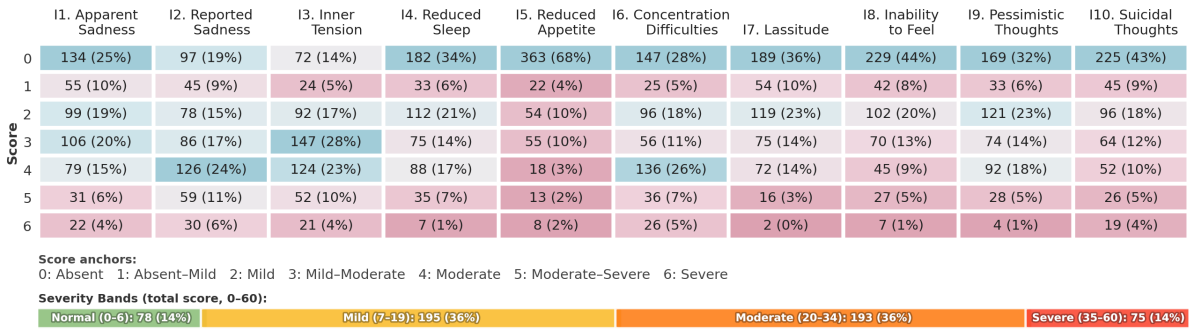

Figure 2: **MADRS score distributions.** Per-item score counts (with percentages) across the 541 sessions and overall severity-band distribution (No depression 0–6, Mild 7–19, Moderate 20–34, Severe 35–60).

4.4 Statistical Analysis of Prediction Errors

We fit two cross-classified linear mixed-effects models (Browne et al., 2001; Goldstein, 2010) for standard and reasoning-augmented LLMs.

Models. For non-reasoning models:

$$\begin{aligned} \mathrm{Y}^{(\mathrm{NR})}_{p,s,m,\tau} = \alpha &+ \beta_1 S^W_{p,s,\tau} + \beta_2 S^B_{p,\tau} + \beta_3 T^W_{p,s} + \beta_4 T^B_p \\ &+ \beta_5\,\mathrm{Par}_m + \beta_6\,\mathrm{Cont}_m + \beta_7\,\mathrm{MOE}_m \\ &+ b_m + u_p + v_{s:p} + w_\tau + \varepsilon_{p,s,m,\tau}. \end{aligned} \tag{1}$$

For reasoning models, we add reasoning-token predictors:

$$\mathrm{Y}^{(\mathrm{R})}_{p,s,m,\tau} = \mathrm{Y}^{(\mathrm{NR})}_{p,s,m,\tau} + \beta_8 R^W_{p,s,\tau} + \beta_9 R^B_{p,\tau}. \tag{2}$$

Variables. $\mathrm{Y}_{p,s,m,\tau}$ is seed-averaged MAE for patient p, session s, model m, task τ (ten MADRS items plus total, with total-score errors rescaled to item range by dividing by 10). Continuous predictors are log-transformed, z-standardized, and decomposed into within- and between-patient components: $S_{p,s,\tau}$ (session–item severity), $T_{p,s}$ (transcript tokens), $R_{p,s,\tau}$ (reasoning tokens; $R = 0$ for non-reasoning). Model covariates are $\log_{10}$ parameters (Par_m), $\log_{10}$ context window (Cont_m), and mixture-of-experts indicator (MOE_m). Random intercepts b_m, u_p, $v_{s:p}$, w_τ account for model, patient, session-within-patient, and item variance; $\varepsilon_{p,s,m,\tau}$ are homoscedastic Gaussian residuals.

5 Results

Our evaluation addresses three central questions: (1) Can LLMs score MADRS from real interviews with clinically reliable accuracy? (2) Under what conditions does reasoning improve performance over standard models? (3) What factors predict model performance on clinical assessment tasks?

5.1 Item-then-Sum (ITS) vs. Direct Total Score (DTS)

Table 2 reveals a striking pattern: ITS aggregation substantially reduces total MAE for most models. Standard Qwen 2.5 (72B) improves from 5.53 (DTS) to 3.80 (ITS, 31% reduction); reasoning-augmented models show even larger improvements, with GPT OSS 120B transitioning from 6.80 to 3.78 (44% reduction) and QwQ (32B) dramatically improving from 21.09 to 4.03 (81% reduction). Most models achieve total MAE in the acceptable band (< 6) after aggregation; however, some low-parameter models (e.g., DeepSeek R1 Llama 3.1 8B with sum MAE of 11.84, and Llama 3.1 8B with 8.30) remain in the moderate or substantial error range, indicating fundamental capacity limitations below certain parameter thresholds.

Critically, reasoning-augmented models' internal aggregation—even when they explicitly sum items in their reasoning traces—remains less accurate than post-hoc item summation. Even when models show step-by-step arithmetic, their DTS predictions remain systematically inferior to mechanically summed item-level outputs, indicating fundamental limits of DTS prediction regardless of reasoning capability.

Output validity further compounds these capacity limitations: **Table 4** shows that reasoning-augmented models produce substantially more invalid total-score outputs than item-level ones (e.g., QwQ 32B averages 34.7 invalid total outputs vs. 2.7–5.0 per item), as the longer transcripts used in DTS prediction leave less generation budget for the reasoning trace, increasing the likelihood that the model fails to produce a score. This asymmetry partly explains the outsized DTS–ITS gap for reasoning models.

Table 2: Total-score evaluation: reasoning vs. non-reasoning. **DTS**: $\hat{T}=f_\theta(x)$; **ItS**: $\hat{T}=\sum \hat{y}_i$. MoE sizes: Active–Total params. **Bold** = best; darker = best/worst. **MAE**: Acceptable (<6), Substantial (≥ 12). **F1** ($T\geq 20$): $\geq$ Q3 (0.86), $<$ Q1 (0.77).

Model (Size)	Arch.	Ctx.	MAE ↓		F1 ↑	
REASONING MODELS			DTS	ITS	DTS	ITS
Qwen 3 Next (3B-80B)	MoE	262k	5.90 ±0.17	4.31 ±0.11	**0.87** ±0.01	0.87 ±0.01
GPT OSS 120B (5B-117B)	MoE	131k	6.80 ±0.19	3.78 ±0.08	0.84 ±0.00	0.86 ±0.01
Qwen 3 (22B-235B)	MoE	262k	7.60 ±0.44	3.68 ±0.02	0.85 ±0.01	0.87 ±0.00
GPT OSS 20B (3B-21B)	MoE	131k	7.86 ±0.18	3.99 ±0.03	0.82 ±0.01	0.85 ±0.02
Magistral Small 2507 (24B)	Dense	40k	8.76 ±0.90	3.81 ±0.02	0.82 ±0.01	0.86 ±0.01
DeepSeek R1 Qwen 2.5 (32B)	Dense	131k	10.03 ±0.78	3.83 ±0.04	0.77 ±0.01	0.86 ±0.01
DeepSeek R1 Llama 3.3 (70B)	Dense	131k	10.14 ±0.79	3.90 ±0.04	0.81 ±0.01	0.87 ±0.01
Qwen 3 (14B)	Dense	131k	10.84 ±3.01	4.12 ±0.04	0.79 ±0.05	0.88 ±0.01
Qwen 3 (32B)	Dense	131k	12.82 ±0.58	4.06 ±0.02	0.75 ±0.00	0.87 ±0.01
Qwen 3 (8B)	Dense	131k	13.56 ±0.60	4.43 ±0.05	0.73 ±0.03	0.86 ±0.00
DeepSeek R1 Llama 3.1 (8B)	Dense	131k	13.78 ±0.21	11.84 ±0.82	0.73 ±0.01	0.71 ±0.01
Qwen 3 (4B)	Dense	131k	15.23 ±0.00	5.46 ±0.11	0.72 ±0.03	0.84 ±0.01
Qwen 3 (1.7B)	Dense	40k	19.05 ±0.29	7.45 ±0.17	0.57 ±0.02	0.78 ±0.01
Qwen 3 (0.6B)	Dense	40k	19.72 ±0.91	9.46 ±0.19	0.58 ±0.01	0.72 ±0.01
QwQ (32B)	Dense	131k	21.09 ±1.06	4.03 ±0.05	0.76 ±0.01	0.87 ±0.00
Qwen 3 (3B-30B)	MoE	262k	50.22 ±2.22	3.98 ±0.11	0.67 ±0.05	0.87 ±0.01
NON-REASONING MODELS			DTS	ITS	DTS	ITS
Qwen 2.5 (72B)	Dense	131k	**5.53** ±0.15	3.80 ±0.03	0.86 ±0.01	0.87 ±0.00
Llama 4 Maverick (17B-400B)	MoE	1m	5.60 ±0.05	4.00 ±0.02	0.84 ±0.00	0.88 ±0.00
Llama 4 Scout (17B-109B)	MoE	10m	5.75 ±0.06	3.70 ±0.00	0.82 ±0.02	0.86 ±0.00
Qwen 2.5 (14B): 1M	Dense	1m	6.09 ±0.20	**3.61** ±0.03	0.82 ±0.01	0.88 ±0.01
Qwen 2.5 (7B): 1M	Dense	1m	6.11 ±0.35	4.71 ±0.27	0.85 ±0.01	0.86 ±0.00
Llama 3.3 (70B)	Dense	131k	7.32 ±0.17	4.57 ±0.07	0.86 ±0.00	0.86 ±0.00
Qwen 3 Next: NR (3B-80B)	MoE	262k	8.88 ±0.24	4.30 ±0.05	0.87 ±0.00	**0.88** ±0.00
Gemma 3 IT (27B)	Dense	131k	12.12 ±0.25	4.76 ±0.15	0.77 ±0.01	0.85 ±0.00
Llama 3.1 (8B)	Dense	131k	13.71 ±1.45	8.30 ±1.22	0.74 ±0.04	0.81 ±0.01

Screening performance (F_1 for total $\geq$ 20) shows considerable variation: top-performing models achieve F_1 scores of 0.86–0.88, while smaller models range from 0.57 to 0.77. For most mid-to-large scale models, differences between DTS and ITS approaches are modest ($\Delta < 0.05$), though some small models show larger improvements with aggregation (e.g., Qwen 3 1.7B: 0.57 $\rightarrow$ 0.78, $\Delta = +0.21$). The distinction between reasoning-augmented and standard models is minimal for screening tasks, with both architectures achieving comparable F_1 scores at similar parameter scales.

5.2 Item-wise Performance

Table 3 presents item-level MAE across all models, with performance categorized into acceptable (<0.6), moderate (0.6–1.2), and substantial (≥ 1.2) error bands. Among standard models, Qwen 2.5 (72B) achieves the lowest average MAE (0.74), with *Reduced Appetite* (0.50) and *Suicidal Thoughts* (0.58) falling in the acceptable band and remaining items in the moderate range (0.65–0.94). Llama 4 Scout attains comparable performance (mean 0.75), again placing *Reduced Appetite* at 0.42 (acceptable error). Reasoning-augmented

Table 3: Item-wise MAE ± std for I1–I10 with mean. Formatted as Table 2.

Model (Size)	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	Mean
REASONING MODELS											
Qwen 3 (22B-235B)	0.95 ± 0.03	0.88 ± 0.03	0.75 ± 0.03	0.93 ± 0.02	0.40 ± 0.02	0.88 ± 0.02	0.75 ± 0.03	0.78 ± 0.01	0.66 ± 0.01	0.64 ± 0.03	0.76 ± 0.01
GPT OSS 120B (5B-117B)	0.98 ± 0.01	0.94 ± 0.01	0.76 ± 0.02	0.87 ± 0.02	**0.39** ± 0.01	0.87 ± 0.02	0.79 ± 0.02	0.76 ± 0.03	0.65 ± 0.02	0.70 ± 0.04	0.77 ± 0.02
Magistral Small 2507 (24B)	1.04 ± 0.05	0.94 ± 0.02	0.80 ± 0.01	**0.81** ± 0.00	0.44 ± 0.02	0.85 ± 0.01	0.75 ± 0.05	0.78 ± 0.01	0.73 ± 0.01	0.60 ± 0.01	0.77 ± 0.01
DeepSeek R1 Qwen 2.5 (32B)	1.05 ± 0.01	0.95 ± 0.02	0.82 ± 0.03	0.91 ± 0.05	0.47 ± 0.01	0.96 ± 0.02	0.74 ± 0.03	0.80 ± 0.01	0.73 ± 0.02	0.67 ± 0.03	0.81 ± 0.01
DeepSeek R1 Llama 3.3 (70B)	1.06 ± 0.01	0.92 ± 0.02	0.78 ± 0.01	0.92 ± 0.04	0.49 ± 0.02	0.89 ± 0.01	0.84 ± 0.01	0.92 ± 0.02	0.74 ± 0.01	0.67 ± 0.02	0.82 ± 0.00
GPT OSS 20B (3B-21B)	1.07 ± 0.01	0.98 ± 0.03	0.85 ± 0.02	0.96 ± 0.02	0.48 ± 0.02	0.99 ± 0.02	0.78 ± 0.02	0.82 ± 0.02	0.68 ± 0.00	0.69 ± 0.03	0.83 ± 0.01
Qwen 3 (32B)	1.06 ± 0.02	0.87 ± 0.02	0.77 ± 0.02	1.00 ± 0.02	0.53 ± 0.02	0.93 ± 0.02	0.81 ± 0.02	0.89 ± 0.02	0.69 ± 0.02	0.82 ± 0.01	0.84 ± 0.00
QwQ (32B)	1.07 ± 0.01	1.00 ± 0.02	0.82 ± 0.03	0.95 ± 0.02	0.56 ± 0.01	1.03 ± 0.04	0.79 ± 0.02	0.88 ± 0.04	0.72 ± 0.01	0.78 ± 0.02	0.86 ± 0.01
Qwen 3 Next (3B-80B)	1.04 ± 0.02	1.04 ± 0.02	0.88 ± 0.01	1.00 ± 0.01	0.53 ± 0.01	0.99 ± 0.02	1.02 ± 0.03	0.98 ± 0.02	0.76 ± 0.03	0.69 ± 0.01	0.89 ± 0.01
Qwen 3 (14B)	1.16 ± 0.03	1.00 ± 0.01	0.91 ± 0.02	1.10 ± 0.04	0.51 ± 0.01	1.06 ± 0.03	0.86 ± 0.01	0.94 ± 0.03	0.71 ± 0.02	0.75 ± 0.03	0.90 ± 0.01
Qwen 3 (8B)	1.18 ± 0.04	1.06 ± 0.00	0.81 ± 0.01	1.23 ± 0.03	0.70 ± 0.04	1.01 ± 0.02	0.88 ± 0.03	1.34 ± 0.03	0.71 ± 0.02	0.79 ± 0.03	0.97 ± 0.01
Qwen 3 (3B-30B)	1.25 ± 0.06	1.08 ± 0.02	1.09 ± 0.04	1.17 ± 0.01	0.62 ± 0.01	1.19 ± 0.04	1.09 ± 0.02	1.07 ± 0.04	0.75 ± 0.02	0.85 ± 0.04	1.02 ± 0.01
Qwen 3 (4B)	1.27 ± 0.02	1.09 ± 0.03	0.95 ± 0.02	1.20 ± 0.02	0.80 ± 0.02	1.27 ± 0.06	1.04 ± 0.01	1.69 ± 0.03	0.81 ± 0.01	0.94 ± 0.01	1.11 ± 0.00
Qwen 3 (1.7B)	1.68 ± 0.01	1.51 ± 0.02	1.16 ± 0.01	1.43 ± 0.02	0.97 ± 0.05	1.28 ± 0.04	1.23 ± 0.06	2.45 ± 0.12	0.93 ± 0.02	0.94 ± 0.04	1.36 ± 0.01
DeepSeek R1 Llama 3.1 (8B)	1.53 ± 0.01	1.41 ± 0.05	0.96 ± 0.06	1.61 ± 0.10	1.65 ± 0.13	1.29 ± 0.06	1.73 ± 0.08	2.20 ± 0.14	1.06 ± 0.06	1.39 ± 0.17	1.48 ± 0.07
Qwen 3 (0.6B)	1.78 ± 0.04	1.47 ± 0.02	1.32 ± 0.03	2.28 ± 0.05	1.72 ± 0.05	1.53 ± 0.01	1.72 ± 0.05	1.84 ± 0.07	1.37 ± 0.03	1.50 ± 0.03	1.65 ± 0.01
NON-REASONING MODELS											
Qwen 2.5 (72B)	0.94 ± 0.02	**0.77** ± 0.01	**0.66** ± 0.02	0.92 ± 0.01	0.50 ± 0.02	**0.80** ± 0.02	**0.71** ± 0.01	0.89 ± 0.01	0.65 ± 0.01	0.58 ± 0.01	**0.74** ± 0.00
Llama 4 Scout (17B-109B)	1.06 ± 0.01	0.89 ± 0.01	0.76 ± 0.01	0.83 ± 0.01	0.42 ± 0.01	0.83 ± 0.01	0.73 ± 0.00	**0.76** ± 0.00	0.66 ± 0.01	0.60 ± 0.01	0.75 ± 0.00
Qwen 2.5 (14B): 1M	**0.89** ± 0.01	0.90 ± 0.03	0.80 ± 0.03	0.89 ± 0.02	0.48 ± 0.01	0.85 ± 0.02	0.78 ± 0.02	0.97 ± 0.02	0.71 ± 0.02	**0.57** ± 0.02	0.78 ± 0.01
Llama 4 Maverick (17B-400B)	0.96 ± 0.00	0.83 ± 0.01	0.75 ± 0.01	0.88 ± 0.01	0.42 ± 0.01	0.91 ± 0.02	0.92 ± 0.01	1.06 ± 0.02	**0.63** ± 0.00	0.59 ± 0.01	0.80 ± 0.00
Llama 3.3 (70B)	1.07 ± 0.01	0.83 ± 0.01	0.83 ± 0.01	0.92 ± 0.00	0.52 ± 0.02	0.82 ± 0.00	0.79 ± 0.01	0.85 ± 0.01	0.78 ± 0.01	0.66 ± 0.01	0.81 ± 0.01
Gemma 3 IT (27B)	1.08 ± 0.00	0.92 ± 0.01	0.99 ± 0.02	0.96 ± 0.01	0.62 ± 0.01	0.83 ± 0.01	0.81 ± 0.03	1.06 ± 0.03	0.82 ± 0.02	0.69 ± 0.01	0.88 ± 0.01
Qwen 3 Next: NR (3B-80B)	1.10 ± 0.01	1.04 ± 0.01	0.94 ± 0.00	0.95 ± 0.01	0.63 ± 0.00	0.88 ± 0.00	0.92 ± 0.02	0.87 ± 0.01	0.83 ± 0.01	0.62 ± 0.01	0.88 ± 0.00
Qwen 2.5 (7B): 1M	1.08 ± 0.03	0.99 ± 0.04	0.78 ± 0.01	0.92 ± 0.01	0.59 ± 0.02	0.93 ± 0.02	1.08 ± 0.04	0.96 ± 0.02	1.00 ± 0.03	0.67 ± 0.04	0.90 ± 0.01
Llama 3.1 (8B)	1.38 ± 0.06	1.15 ± 0.12	1.34 ± 0.07	1.61 ± 0.15	1.19 ± 0.22	1.07 ± 0.07	1.57 ± 0.10	1.46 ± 0.08	1.23 ± 0.08	1.05 ± 0.01	1.30 ± 0.09

models show similar patterns: GPT OSS 120B achieves mean MAE of 0.77, and Qwen 3 (235B) achieves 0.76, with *Reduced Appetite* consistently in the acceptable band (0.39 and 0.40, respectively). Notably, mid-scale models approach large-model performance: Qwen 2.5 (14B; 1M context)

Table 4: Invalid output counts (mean across runs) per MADRS item and total. Lower is better. Formatted as Table 2.

Model (Size)	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	Total
REASONING MODELS											
Magistral Small 2507 (24B)	0	0	0	0	0	0	0	0	0	0	0
GPT OSS 120B (5B-117B)	0	0	0.3	0	0.3	0	0	0	0	0	0.7
DeepSeek R1 Llama 3.3 (70B)	0	0	0.7	0	0	0	0	0	0	0.3	1.0
GPT OSS 20B (3B-21B)	0	0	0	0.3	0	0	0	0.7	0	0	1.0
DeepSeek R1 Llama 3.1 (8B)	0.3	0	0	0.7	0.3	0	0	0	0	0.3	1.7
DeepSeek R1 Qwen 2.5 (32B)	0.3	0.3	0.3	1.0	0	0.7	0	0	0.3	0	3.0
Qwen 3 Next (3B-80B)	0.3	0	1.3	0	0.7	0.3	0	0.7	0	0	3.3
Qwen 3 (32B)	1.7	0	0	0	0.7	0.7	0	1.3	0	0.3	4.7
Qwen 3 (8B)	0.7	1.0	1.0	1.0	3.0	0.3	1.0	0.7	0.3	1.0	10.0
Qwen 3 (22B-235B)	2.7	2.3	0.7	1.0	1.3	0	4.0	1.0	0.3	3.7	17.0
Qwen 3 (14B)	2.3	2.3	1.7	2.0	2.0	1.7	2.3	1.7	0	1.3	17.3
Qwen 3 (4B)	1.7	1.3	3.7	3.3	4.3	5.0	2.3	4.0	0.7	3.7	30.0
QwQ (32B)	2.7	3.7	4.0	4.7	4.7	3.0	4.0	3.0	0	5.0	34.7
Qwen 3 (0.6B)	7.7	7.0	9.7	9.0	5.7	9.3	9.3	9.3	0.7	8.7	76.3
Qwen 3 (1.7B)	8.3	5.3	10.7	8.3	12.0	11.0	8.0	8.7	1.3	6.3	80.0
Qwen 3 (3B-30B)	20.3	22.3	22.3	9.0	18.7	17.3	24.3	21.7	0	19.7	175.7
NON-REASONING MODELS											
All 9 non-reasoning models produced 0 invalid outputs.											

achieves mean MAE of 0.78 with acceptable error on physiological symptom items.

In contrast, models with $\leq$ 8B parameters exhibit multiple cells in the substantial band ($\geq$ 1.2), with Llama 3.1 (8B) and Qwen 3 (0.6B–1.7B) frequently exceeding MAE of 1.3 across several items. Comparing architectures, standard models demonstrate superior average item-level performance (mean MAE 0.87) compared to reasoning-augmented models (mean MAE 0.99), suggesting that explicit reasoning traces do not uniformly improve accuracy at this task granularity.

Part of the accuracy gap between architectures is attributable to output validity: as **Table 4** shows, all nine non-reasoning models produced zero invalid outputs, whereas reasoning-augmented models—particularly at smaller scales—frequently exhaust their generation budget on reasoning traces before emitting a parseable prediction (e.g., Qwen 3 0.6B: 76.3; 1.7B: 80.0; 3B–30B MoE: 175.7 invalid outputs on the total-score task).

Item-specific patterns reveal consistent difficulty hierarchies: *Reduced Appetite* (Item 5) achieves acceptable error (< 0.6) across all models regardless of architecture or scale, likely reflecting its concrete behavioral nature. Conversely, smaller reasoning-augmented models struggle particularly with *Lassitude* (Item 7) and *Inability to Feel* (Item 8), with MAE reaching 2.45 for the smallest variant (Qwen 3 1.7B) on Item 8, indicating that affective-behavioral items pose unique challenges for low-capacity reasoning systems. The persistent difficulty of *Apparent Sadness* (Item 1) across both architectures—despite interview questions specifically probing what others observe about the patient's appearance—suggests an inherent limitation of text-only assessment: this item fundamentally requires visual and paralinguistic observations that necessitate multimodal integration.

5.3 Prompt Ablation: When Does Reasoning Help?

Using **Qwen 3 Next (80B; MoE)** with and without reasoning augmentation, we ablated descriptive cues and demonstrative cues across four configurations. **Figure 3** presents paired comparisons (reasoning-augmented in blue, standard in red), with Δ = Standard $-$ Reasoning labeled above each pair.

With **full scaffolding** (descriptive + demonstrative cues), the standard model slightly outperforms: Standard=0.88, Reasoning=0.89, $\Delta = -0.03$. Removing descriptive cues (**No Descriptions**) equalizes performance at $\approx$ 0.98 for both. When only descriptive cues remain (**No Demonstrations**), reasoning-augmented models gain a small edge:

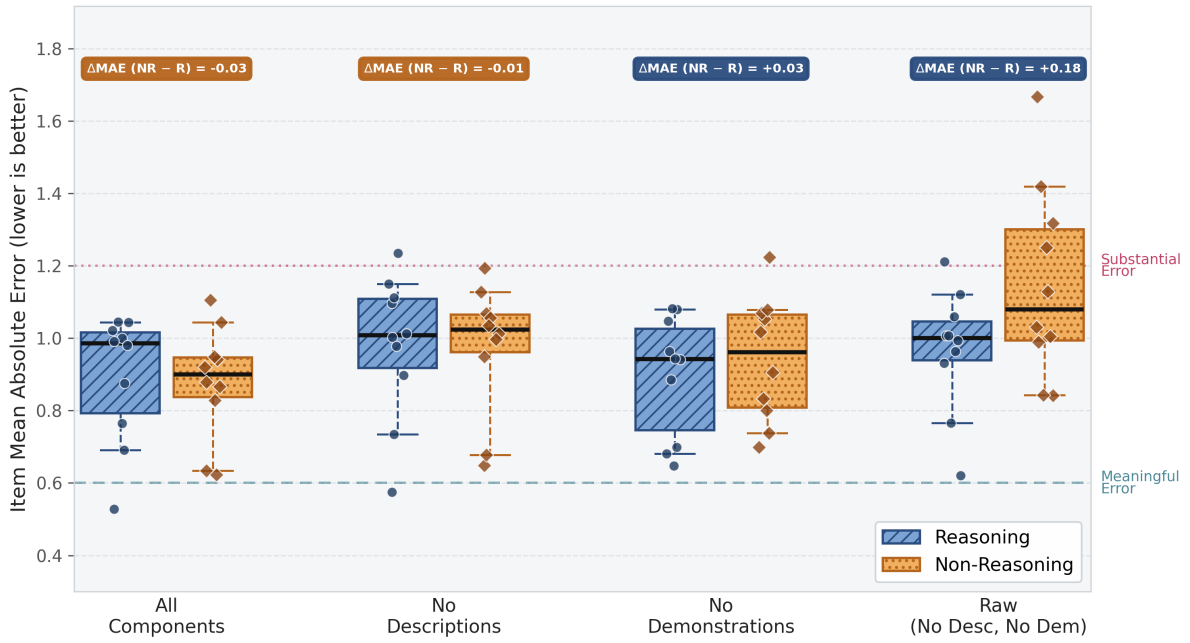

Figure 3: **Prompt ablation: descriptive & demonstrative cues, reasoning vs. no-reasoning.** Mean Item MAE for Reasoning-Augmented (R) vs. Standard (NR) models across four prompt configurations. Values above each bar pair show Δ = Standard − Reasoning; positive values indicate a reasoning advantage.

Reasoning=0.90, Standard=0.94, $\Delta = +0.03$. The advantage of reasoning becomes pronounced in the **Raw** condition (no cues): Reasoning=0.97, Standard=1.15, $\Delta = +0.18$, with reasoning-augmented models also showing narrower variance. This pattern reveals that explicit reasoning primarily benefits performance when prompt scaffolding is sparse—when models lack structured clinical definitions and examples, reasoning helps navigate the ambiguity. Conversely, with strong descriptive cues, standard models can match or even exceed reasoning-augmented ones.

5.4 Mixed-Effects Analysis: Disentangling Model and Task Factors

Figure 4 summarizes separate task-level, seed-averaged mixed-effects fits from **Section 4.4**: one fit for reasoning-augmented models (N=92,864 observations; sessions within patients n=541, patients n=277, tasks n=11, models n=16) and one fit for non-reasoning models (N=52,236 observations; sessions within patients n=541, patients n=277, tasks n=11, models n=9).

Model characteristics. Larger models yield lower error in both strata (log-parameters, z-standardized): reasoning-augmented $\hat{\beta} = -0.638$ ($t = -4.84$), standard $\hat{\beta} = -0.206$ ($t = -2.49$). Context length shows divergent effects: positive for reasoning-augmented ($\hat{\beta} = +0.966$, $t = 3.51$) but negligible for standard models ($\hat{\beta} = -0.093$, $t = -1.92$). Mixture-of-Experts architectures correlate with higher error in both strata (reasoning-augmented: $\hat{\beta} = +0.525$, $t = 2.02$; standard: $\hat{\beta} = +0.308$, $t = 1.73$).

Clinical complexity. Higher severity predicts higher error both *between* patients (reasoning-augmented: $\hat{\beta} = +0.119$, $t = 21.4$; standard: $\hat{\beta} = +0.131$, $t = 20.1$) and *within* patients (reasoning-augmented: $\hat{\beta} = +0.137$, $t = 33.7$; standard: $\hat{\beta} = +0.078$, $t = 15.3$). Transcript length shows small effects: between-patient verbosity is slightly harder (reasoning-augmented: $\hat{\beta} = +0.035$, $t = 2.46$; standard: $\hat{\beta} = +0.105$, $t = 6.64$).

Reasoning trace patterns. For reasoning-augmented models only, reasoning length exhibits a bidirectional pattern: models/sessions with *typically* longer traces show higher error (between-patient/item $\hat{\beta} = +0.225$, $t = 43.5$), but producing *more* reasoning than usual for a given patient–item correlates with lower error (within-session $\hat{\beta} = -0.929$, $t = -64.9$). This suggests adaptive reasoning—deploying extra computation when needed—is beneficial, while

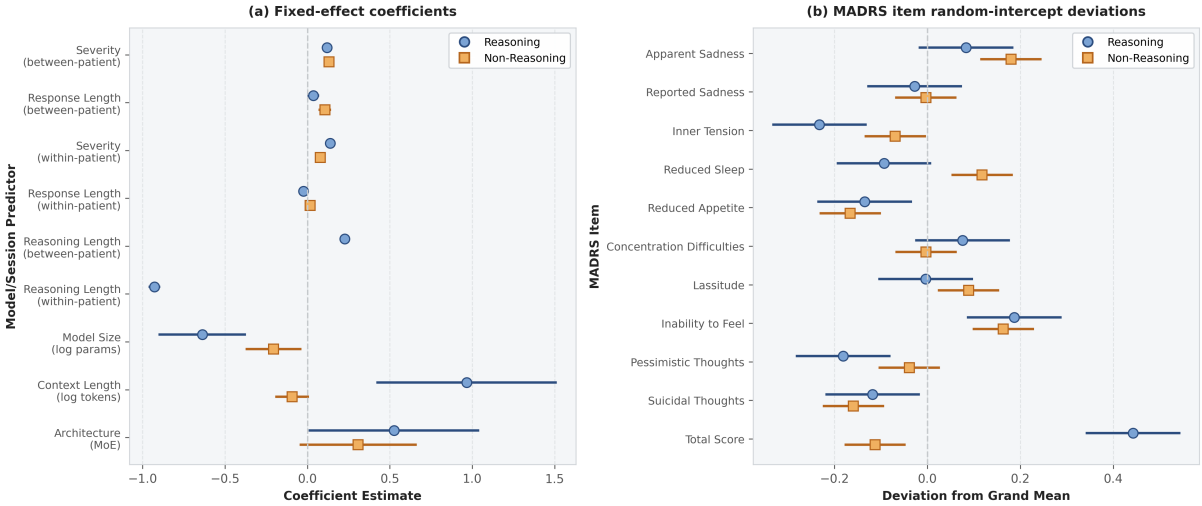

Figure 4: **Mixed-effects analysis (task-level, seed-averaged).** Blue = reasoning; red = standard. (a) Fixed-effect estimates (points: coefficients; bars: 95% CI). Longer *within-patient* reasoning is linked to lower error, while patients/models with *typically* longer reasoning show higher error. Larger models reduce error; MoE tends to increase it. (b) Item random-intercept deviations (BLUPs) highlight item-specific difficulty (e.g., *Inability to Feel* >0; *Inner Tension* <0).

uniform verbosity correlates with harder cases rather than improved accuracy.

Item-level heterogeneity. Random-intercept BLUPs (**Figure 4**b) reveal a consistent difficulty hierarchy across architectures. Both reasoning-augmented and standard models find *Inability to Feel* (I8; deviations $+0.19$ and $+0.16$, respectively) and *Apparent Sadness* (I1; $+0.08$ and $+0.18$) the hardest items, while *Inner Tension* (I3; -0.23), *Reduced Appetite* (I5; -0.13 and -0.17), and *Suicidal Thoughts* (I10; -0.12 and -0.16) are consistently easiest. The broad alignment of these rankings across the two strata indicates that item difficulty is driven primarily by construct properties—concrete behavioral items are easier than subjective affective ones—rather than by architectural differences. The elevated difficulty of *Apparent Sadness* is particularly instructive: despite interview questions explicitly asking patients what others comment on about their appearance, this item retains a fundamental observational component—clinical MADRS administration relies heavily on the clinician's direct visual assessment of facial expression, posture, and nonverbal presentation, modalities entirely absent from text transcripts, suggesting that certain depression symptoms would benefit from multimodal integration of visual and acoustic signals. One notable divergence concerns the rescaled total-score task: reasoning-augmented models show the largest positive deviation ($+0.44$), reflecting the difficulty of DTS scoring discussed in **Section 5.1**, whereas standard models show a negative deviation (-0.11), suggesting that DTS prediction is comparatively less penalized in the non-reasoning stratum.

6 Discussion

Our comprehensive evaluation of 25 state-of-the-art open-source LLMs on the LLAMADRS benchmark reveals nuanced patterns about when reasoning augmentation benefits clinical assessment. We find that strong open LLMs achieve clinically moderate item-wise accuracy on MADRS from real interviews, using established interpretability bands for error magnitude (Montgomery and Åsberg, 1979; Turkoz et al., 2021). Building on this baseline, the ITS procedure—score items, then sum—substantially reduces total MAE for most models, with the majority achieving acceptable error (< 6); notably, even when reasoning-augmented models "self-aggregate" in their traces, their DTS predictions remain inferior to summing item estimates. However, this benefit does not extend uniformly to all models: smaller models with limited capacity ($\leq$ 8B parameters) may remain in the moderate error band even after aggregation, highlighting the importance of adequate model scale for clinical reliability.

The relationship between reasoning augmentation and performance proves context-dependent. Under anchored, schema-constrained prompts that encode clinical descriptive cues, standard variants frequently match or surpass their reasoning-augmented counterparts. Our ablation study clarifies this pattern: explicit reasoning helps primarily when scaffolding is sparse, aligning with recent reports that reasoning gains are task- and prompt-dependent. With full clinical scaffolding (descriptive and demonstrative cues), standard models achieve $\Delta = -0.03$ advantage; without any cues, reasoning-augmented models gain $\Delta = +0.18$ advantage with reduced variance.

A mixed-effects analysis further contextualizes these results: larger parameter counts predict lower error, context-window length has near-zero effect, and—after adjusting for case mix—both MoE architecture and uniform reasoning augmentation correlate with higher error. Critically, longer *within-session* reasoning traces correlate with lower error ($\hat{\beta} = -0.929$), while models that *typically* produce longer traces show higher error ($\hat{\beta} = +0.225$). This bidirectional pattern suggests that adaptive reasoning—deploying extra computation selectively—is beneficial, whereas indiscriminate verbosity indicates difficulty rather than capability.

Clinical implications. The consistent achievement of acceptable-to-moderate error bands across strong models suggests potential clinical utility, particularly for the ITS approach (item-level scoring followed by summation). The systematic difficulty pattern across items—with subjective/affective items showing higher error than behavioral items—highlights fundamental challenges in text-based assessment of internal states. This pattern persists regardless of reasoning augmentation, suggesting inherent limitations in inferring mood states from dialogue alone.

Methodological insights. Our findings challenge assumptions about the uniform superiority of reasoning augmentation. The effectiveness of descriptive cues—clinical anchors and severity definitions—as the highest-value prompt component suggests that for well-structured clinical tasks with clear rubrics, careful prompt engineering can substitute for explicit reasoning. This has practical implications for deployment: standard models with good prompts may be more efficient than reasoning-augmented models that require additional computation without commensurate gains.

7 Limitations

Our analysis is text-only; nonverbal cues and prosody—important for affective items—are not modeled. The CAMI dataset represents inpatient psychiatric settings, potentially limiting generalizability to outpatient or community samples. Future work should explore multimodal integration, cross-population validation, and methods for uncertainty quantification appropriate for clinical deployment.

8 Conclusion

We introduced LLAMADRS, a benchmark for evaluating LLM performance on clinical depression assessment using real patient–clinician interviews with $5{,}804$ clinician-rated MADRS annotations. Evaluating 25 models reveals that strong open LLMs achieve clinically acceptable item-level accuracy, while ITS scoring reduces total MAE by 30–70% compared to DTS prediction, exposing fundamental limits of end-to-end DTS regression.

Reasoning augmentation benefits are context-dependent: standard models often match reasoning-augmented variants under well-structured prompts ($\Delta = -0.03$), but reasoning provides clear gains with sparse scaffolding ($\Delta = +0.18$). Mixed-effects analyses identify model scale and reasoning length as key predictors, with reasoning tokens showing the largest effect ($\beta = -0.929$). These findings challenge assumptions about reasoning augmentation and establish a foundation for automated yet interpretable clinical assessment.

Acknowledgments

This material is based upon work partially supported by National Institutes of Health awards R01MH125740, R01MH132225, U01MH136535, and R21MH130767. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors, and no official endorsement should be inferred.

References

Mario Aragón, Javier Parapar, and David E. Losada. 2024. Delving into the depths: Evaluating depression severity through BDI-biased summaries. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 12–22, St. Julians, Malta. Association for Computational Linguistics.

Mihael Arcan, David-Paul Niland, and Fionn Delahunty. 2024. An assessment on comprehending mental health through large language models. ArXiv preprint arXiv:2401.04592.

Eliseo Bao, Anxo Pérez, and Javier Parapar. 2024. Explainable depression symptom detection in social media. *Health Information Science and Systems*, 12(1):47.

William J. Browne, Harvey Goldstein, and Jon Rasbash. 2001. Multiple membership multiple classification (mmmc) models. *Statistical Modelling*, 1(2):103–124.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Brian Culhane, Alexander Yip, Lisette Liebson, Robert Patterson, Hamid Rahimi-Eichi, Jeffrey Girard, Louis-Philippe Morency, Einat Liebenthal, and Justin T Baker. 2023. Toward expert systems in mental health assessment: A framework for context-adaptive multimodal informatics (cami). Poster presented at Harvard Psychiatry Research Day.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM 2013)*, pages 128–137, Cambridge, Massachusetts, USA. AAAI Press.

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

Isaac R. Galatzer-Levy, Daniel McDuff, Vivek Natarajan, Alan Karthikesalingam, and Matteo Malgaroli. 2023. The capability of large language models to measure psychiatric functioning. ArXiv preprint arXiv:2308.01834.

B. Gillis, Justin T. Baker, and Einat Liebenthal. 2023. Assessing interrater reliability of clinical ratings in monthly interviews of subjects with mood and psychotic disorders. Poster presented at Harvard Psychiatry Research Day.

Harvey Goldstein. 2010. *Multilevel Statistical Models*, 4 edition. Wiley, Chichester.

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.

Haibo Jin, Peiyan Zhang, Man Luo, and Haohan Wang. 2025. Reasoning can hurt the inductive abilities of large language models. In *Advances in Neural Information Processing Systems (NeurIPS 2025)*, volume 38. To appear.

Sehee Lim, Yejin Kim, Chi-Hyun Choi, Jy-yong Sohn, and Byung-Hoon Kim. 2024. ERD: A framework for improving LLM reasoning for cognitive distortion classification. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 292–300, Mexico City, Mexico. Association for Computational Linguistics.

Seyedali Mohammadi, Edward Raff, Jinendra Malekar, Vedant Palit, Francis Ferraro, and Manas Gaur. 2024. WellDunn: On the robustness and explainability of language models and large language models in identifying wellness dimensions. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 364–388, Miami, Florida, US. Association for Computational Linguistics.

Stuart A. Montgomery and Marie Åsberg. 1979. A new depression scale designed to be sensitive to change. *British Journal of Psychiatry*, 134(4):382–389.

Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. A survey of large language models in psychotherapy: Current landscape and future directions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.

Viet Cuong Nguyen, Mohammad Taher, Dongwan Hong, Vinicius Konkolics Possobom, Vibha Thirunellayi Gopalakrishnan, Ekta Raj, Zihang Li, Heather J. Soled, Michael L. Birnbaum, Srijan Kumar, and Munmun De Choudhury. 2025. Do large language models align with core mental health counseling competencies? In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7488–7511, Albuquerque, New Mexico. Association for Computational Linguistics.

Avinash Patil and Amardeep Kour Gedhu. 2025. Cognitive-mental-llm: Evaluating reasoning in large language models for mental health prediction via online text. ArXiv preprint arXiv:2503.10095.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Nishat Raihan, Sadiya Sayara Chowdhury Puspo, Shafkat Farabi, Ana-Maria Bucur, Tharindu Ranasinghe, and Marcos Zampieri. 2024. MentalHelp: A multi-task dataset for mental health in social media. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11196–11203, Torino, Italia. ELRA and ICCL.

Miriam Schirmer, Tobias Leemann, Gjergji Kasneci, Jürgen Pfeffer, and David Jurgens. 2024. The language of trauma: Modeling traumatic event descriptions across domains with explainable AI. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13224–13242, Miami, Florida, USA. Association for Computational Linguistics.

Monica Sekoyan, Nithin Rao Koluguri, Nune Tadevosyan, Piotr Zelasko, Travis Bartley, Nikolay Karpov, Jagadeesh Balam, and Boris Ginsburg. 2025. Canary-1b-v2 & parakeet-tdt-0.6b-v3: Efficient and high-performance models for multilingual asr and ast. *arXiv preprint arXiv:2509.14128*.

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, Vancouver, BC, Canada. Association for Computational Linguistics.

Loitongbam Gyanendro Singh, Junyu Mao, Rudra Mutalik, and Stuart E. Middleton. 2024. Extracting and summarizing evidence of suicidal ideation in social media contents using large language models. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 218–226, St. Julians, Malta. Association for Computational Linguistics.

Konstantinos Skianis, John Pavlopoulos, and A. Seza Doğruöz. 2024. Severity prediction in mental health: Llm-based creation, analysis, evaluation of a novel multilingual dataset. ArXiv preprint arXiv:2409.17397.

Jae-hee So, Joonhwan Chang, Eunji Kim, Junho Na, JiYeon Choi, Jy-yong Sohn, Byung-Hoon Kim, and Sang Hui Chu. 2024. Aligning large language models for enhancing psychiatric interviews through symptom delineation and summarization: Pilot study. *JMIR Formative Research*, 8:e58418.

Shiyu Teng, Jiaqing Liu, Rahul Kumar Jain, Shurong Chai, Ruibo Hou, Tomoko Tateyama, Lanfen Lin, and Yen-Wei Chen. 2025. Enhancing depression detection with chain-of-thought prompting: From emotion to reasoning using large language models. ArXiv preprint arXiv:2502.05879.

Sichang Tu, Abigail Powers, Natalie Merrill, Negar Fani, Sierra Carter, Stephen Doogan, and Jinho D. Choi. 2024. Automating PTSD diagnostics in clinical interviews: Leveraging large language models for trauma assessments. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 644–663, Kyoto, Japan. Association for Computational Linguistics.

Ibrahim Turkoz, Larry Alphs, Jaskaran Singh, Carol Jamieson, Ella Daly, May Shawi, John J. Sheehan, Madhukar H. Trivedi, and A. John Rush. 2021. Clinically meaningful changes on depressive symptom measures and patient-reported outcomes in patients with treatment-resistant depression. *Acta Psychiatrica Scandinavica*, 143(3):253–263.

Ahmet Yavuz Uluslu, Andrianos Michail, and Simon Clematide. 2024. Utilizing large language models to identify evidence of suicidality risk through analysis of emotionally charged posts. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 264–269, St. Julians, Malta. Association for Computational Linguistics.

Quan Wang, Hannah Muckenhirn, Kevin Svanfeldt, Ernest Pusateri, Yaqian Saraf, and Wei-Ning Hsu. 2024a. Diarizationlm: Speaker diarization post-processing with large language models. In *Proceedings of Interspeech*.

Yuxi Wang, Diana Inkpen, and Prasadith Kirinde Gamaarachchige. 2024b. Explainable depression detection using large language models on social media data. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 108–126, St. Julians, Malta. 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. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–27.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, and Sophia Ananiadou. 2024. MentaLLaMA: Interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024 (WWW '24)*, pages 4489–4500, Singapore. ACM.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077, Singapore. Association for Computational Linguistics.

A Prompt Templates

Each prompt follows the three-component structure from §4.1: (1) descriptive cues, (2) demonstrative cues, and (3) JSON output schema. Ablation configurations (§4.3) selectively remove (1) and/or (2). Full transcripts are provided in the actual prompts; excerpts below are abbreviated for space.

A.1 DTS Prompt

DTS: Total Score (0–60)

Task: Analyze a diarized transcript of a psychiatric session where the Montgomery-Åsberg Depression Rating Scale (MADRS) questionnaire is being administered. Predict the total MADRS score (0–60) that the practitioner would likely give based on the patient's responses and the conversation. Remember that the rating is for the last week, not based on the patient's history or general condition.

[Descriptive Cues]

The MADRS consists of 10 items, each scored from 0–6:
1. Apparent Sadness, 2. Reported Sadness, 3. Inner Tension, 4. Reduced Sleep, 5. Reduced Appetite, 6. Concentration Difficulties, 7. Lassitude, 8. Inability to Feel, 9. Pessimistic Thoughts, 10. Suicidal Thoughts

Severity ranges: 0–6 Normal; 7–19 Mild; 20–34 Moderate; 35–60 Severe.

Instructions:
1. Read the provided interview transcript.
2. Predict the total MADRS score (0–60) based on the overall conversation.
3. Provide a brief explanation (2–3 sentences) for your predicted score.

Output Schema: {rating: 0–60, explanation: 2–3 sentences}

[Demonstrative Cues] – 3 examples at different severity levels

Example 1 (No Depression):
Transcript: <full transcript>
Rating: 4
Explanation: [2–3 sentences]

Example 2 (Moderate):
Transcript: <full transcript>
Rating: 25
Explanation: [2–3 sentences]

Example 3 (Severe):
Transcript: <full transcript>
Rating: 51
Explanation: [2–3 sentences]

A.2 Item-Level Prediction Prompt

Each of the ten MADRS items uses the same template with item-specific definitions, questions, anchors, and seven demonstrative examples (one per anchor 0–6). Below we illustrate with *Reported Sadness* (Item 2).

Item Prompt: Reported Sadness (0–6)

Task: Analyze a diarized transcript of a psychiatric session where the MADRS 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. Remember that the rating is for the last week, not based on the patient's history or general condition. Focus on the questions that are directly related to the MADRS item and the patient's answers.

MADRS Item: Reported Sadness

[Descriptive Cues]

Description: Representing 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. Rate according to intensity, duration and the extent to which the mood is reported to be influenced by events.

Rating Scale:
0 – Occasional sadness in keeping with the circumstances.
1 – Between 0 and 2.
2 – Sad or low but brightens up without difficulty.
3 – Between 2 and 4.
4 – Pervasive feelings of sadness or gloominess. The mood is still influenced by external circumstances.
5 – Between 4 and 6.
6 – Continuous or unvarying sadness, misery or despondency.

Instructions:
1. Read the provided interview transcript.
2. Rate the interviewee's reported sadness on a scale of 0–6.
3. Provide a brief explanation (2–3 sentences) for your rating.
4. List 2–3 key utterances from the conversation supporting your assessment.

Output Schema: {rating: 0–6, explanation: 2–3 sentences, key_utterances: [line numbers], most_relevant_question: [...]}

[Demonstrative Cues] – 7 examples (0–6)

Transcript: <item-segmented transcript>
Rating: 0
Explanation: [2–3 sentences]
Key Utterances: [line nos.]
Most Relevant Question: [...]

...

Transcript: <item-segmented transcript>
Rating: 6
Explanation: [2–3 sentences]
Key Utterances: [line nos.]
Most Relevant Question: [...]