# DEPRESSIONBENCH: A MULTIMODAL BENCHMARK FOR DEPRESSION RECOGNITION AND REASONING ANALYSIS WITH LARGE LANGUAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Depression is a common and serious mental health condition, but its signs can be difficult to recognize in what people say, write, and do across everyday contexts. Large language models (LLMs) could help identify patterns in the language and behavior of people experiencing depression and explain the basis for their assessments. However, it remains unclear how reliably LLMs recognize signs of depression from text, audio, and video collected in different everyday contexts. We introduce DepressionBench, a curated large cohort datasets consists of 34,973 participants and a unified benchmark for depression recognition and analysis of model-reported reasoning. It covers text, audio, vision, and audiovisual inputs across four source scenarios and three languages. To assess depression-recognition performance and whether post-training improves models' ability to identify signs of depression across modalities and everyday contexts, we evaluate twelve open-weight models under zero-shot and few-shot prompting, supervised fine-tuning, and reinforcement learning. Also, to examine whether the models' explanations are consistent with the evidence in the inputs and to identify where their predictions fail, we assess the quality of their reported reasoning and analyze common prediction errors. DepressionBench provides a foundation for future research on multimodal depression recognition, including methods that generalize across everyday contexts. Our project page is available at: https://anonymous.4open.science/r/DepreesionBenchmark-C669

## 1 INTRODUCTION

Depression, also known as depressive disorder, affects an estimated 332 million people worldwide Rong et al. (2025). It can disrupt education, work, relationships, and daily life and, in severe cases, increase the risk of self-harm. Recognizing depression early is therefore important, yet assessment still depends largely on trained clinicians and structured interviews, which are not readily available in many communities. The way depression is expressed also differs from one person and situation to another. Its signs can appear not only in what people say or write, but also in how they speak and behave Sobin C (1997). Changes in language, tone of voice, facial expression, and everyday activity may all be informative, but each provides only a partial view and may be difficult to interpret in isolation.

Large language models (LLMs) offer a possible way to examine language at scale, while multimodal LLMs (MLLMs) can also consider speech and visual behavior. Unlike conventional prediction models, they can report the observations on which a prediction is based. This ability could make an assessment easier to review and help reveal when a model has relied on weak or irrelevant evidence. It does not, however, guarantee that the explanation is faithful to the input or that the prediction is clinically meaningful. *It remains unclear whether these models can recognize signs of depression reliably across different forms of evidence and real-world settings, or whether their explanations are supported by the information they receive.*

Existing studies provide only a fragmented answer. They differ in the information they examine, the populations and languages they cover, the labels they use, and the way they assess performance Chen et al. (2026). Their results are therefore difficult to compare, and success on one dataset may not



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

carry over to another. A social-media post, a clinical interview, a daily vlog, and a passive-sensing record each contain different kinds of evidence. A difference in performance may reflect the model, the source material, the language, the definition of depression, or simply the method of evaluation. Most studies also emphasize the final prediction. Measures such as F1 or mean absolute error are essential, but they cannot show whether a model identified relevant signs, interpreted them correctly, or reached its conclusion from an unsupported assumption. A*s such, a unified benchmark is needed both to compare models fairly and to examine the reasoning behind their predictions for depression detection.*

Privacy and access place further constraints on the models that can be used for this purpose. Records related to mental health often contain sensitive personal information, and sending them to a closed online service may be unacceptable in clinical or personal settings. Open-weight models that can run locally offer a possible alternative: they allow the data to remain on the user's device, can operate without a continuous network connection, and can be examined independently. We therefore focus on open-weight models of up to 30B parameters, a range chosen with local and mobile use in mind, rather than relying on proprietary cloud models.

we present DEPRESSIONBENCH, a benchmark for studying how well language models recognize depression and explain their predictions. We curated a large cohort datasets containing text, audio, video, or a combination of audio and video of 34,973 participants. Together, they represent clinical interviews, daily vlogs, passive sensing, and social media in Chinese, English, and German. Every model is assessed using the same task definitions, data preparation, and measures of performance. We examine twelve open-weight models without examples and with examples in the prompt, after supervised fine-tuning, and reinforcement learning. The benchmark includes both classification and severity estimation, allowing us to ask not only whether a model recognizes depression, but also whether it can recover clinically meaningful differences in severity. We then examine the explanations reported by the models, asking whether they are supported by the available evidence and why incorrect predictions arise.

Our contributions are threefold:

- **A unified multimodal benchmark.** We curated a large-scale depression-recognition benchmark consisting of six public datasets that cover text, audio, vision, and audiovisual information across four settings and three languages. It assesses both classification and severity estimation, reporting F1 and accuracy for classification and MAE, RMSE, and clinical-band accuracy for severity.
- **A broad comparison of current approaches.** We compare twelve open-weight models under zero-shot and few-shot prompting, supervised fine-tuning, and reinforcement learning. Current models perform better than the majority-class baseline only for social-media text. Supervised fine-tuning improves classification but not severity estimation, while reinforcement learning brings no consistent further improvement.
- **An analysis of model explanations.** We introduce an analyst agent that assesses explanations along four dimensions and identifies five types of failure and four underlying causes in incorrect predictions. This analysis shows whether a model recognizes relevant symptoms, connects them to its judgment, and uses evidence specific to the available form of data. Three clinical experts evaluated a random sample of 50 assessments in a blinded review.

## 2 RELATED WORK

**LLMs for Depression Recognition.** Large language models have recently been used to recognize depression from both written and spoken language. Wang et al. Wang et al. (2024) study depression-related symptoms in social-media posts and use LLMs to provide an explanation alongside each prediction. Work on speech has sought to preserve information that is lost in a transcript. Zhang et al. Zhang et al. (2024) represent acoustic landmarks in a form that can be incorporated into an LLM, whereas Li et al. Li et al. (2025) combine speech representations with knowledge from psychology. Together, these studies show that LLMs can draw on more than lexical content alone. However, each study examines only one source of data, form of evidence, or model design. It therefore remains difficult to determine whether an apparent improvement arises from the model, the dataset, or the information made available to the model. DEPRESSIONBENCH instead compares text, audio, vision,

Table 1: The six datasets included in DEPRESSIONBENCH. "CLS" denotes classification using a binary depression label, and "REG" denotes severity estimation using a clinical-scale score. Duration gives the total length of the recordings; "—" indicates that recording duration is not applicable.

| Dataset | Modality | Scenario | Tasks | Language | Size | Duration (min) |
|---|---|---|---|---|---|---|
| WU3D | Text | Social media | CLS | Chinese | 32,570 | — |
| GLOBEM | Text | Passive sensing | CLS, REG | English | 705 | — |
| EATD | Audio | Clinical interview | CLS, REG | Chinese | 162 | 135.6 |
| EDAIC | Audio | Clinical interview | CLS, REG | English | 275 | 4,282 |
| D-Vlog | Vision | Daily vlog | CLS | English | 961 | 960 |
| AVEC2014 | Audiovisual | Clinical interview | CLS, REG | German | 300 | 274 |

and audiovisual information across several real-world settings, languages, model families, and ways of using or training the models.

**Clinical LLM Benchmarks.** The growing use of LLMs in medical domain has led to broader benchmarks of clinical knowledge and decision-making. MultiMedQA brings together medical question-answering tasks and supplements answer accuracy with human judgments of factuality, reasoning, and possible harm Singhal et al. (2023). ClinicBench examines a wider range of clinical tasks, including the understanding of clinical information, text generation, and decisions about patient care Liu et al. (2024). CliMedBench extends this line of work to Chinese medical knowledge and clinical scenarios Ouyang et al. (2024). These benchmarks have been important for showing that strong performance on general language tasks does not necessarily imply reliable performance in medicine. Their evidence, however, is mainly presented as written questions, records, or clinical cases. Depression recognition poses a different problem: the relevant signs may be distributed across a person's words, voice, appearance, and everyday behavior, and may vary with language and setting. DEPRESSIONBENCH complements existing clinical benchmarks by asking whether models can recognize these behavioral signs, estimate their severity, and retain their performance when the form of evidence changes.

**Clinical Reasoning Evaluation.** Accuracy alone cannot reveal whether a medical model reached the correct answer for the right reason. Recent benchmarks therefore examine the reasoning that models report. MedR-Bench uses an automated evaluator to assess reasoning about clinical cases in terms of efficiency, factuality, and completeness Qiu et al. (2025). MedThink-Bench uses explanations written by experts together with an LLM-based evaluator to provide a more detailed assessment of medical reasoning Zhou et al. (2026). Both approaches move beyond the final answer, but they mainly concern reasoning from textual medical questions or clinical cases. In depression recognition, an explanation must also reflect the kind of evidence that was available. A model listening to speech, for example, should not reason as though it had received only a transcript, and any claim about tone, pauses, or facial behavior should be supported by the recording. DEPRESSIONBENCH therefore examines clinical or acoustic grounding, faithfulness to the evidence, depth of symptom-based reasoning, and understanding of the clinical scale. It also separates the assessment of explanation quality from the analysis of incorrect predictions, making it possible to identify both how an explanation fails and why the model reached the wrong conclusion.

## 3 DEPRESSIONBENCH

DEPRESSIONBENCH, as shown in Figure 1, consists of three parts: curate a large-scale depression datasets, measures how well open-weight language models recognize depression, and examines the reasoning reported with their predictions. More detailed description of the workflow and all candidate datasets are given in Appendix A and Appendix B, respectively.

### 3.1 CURATED DATASET

**Coverage.** The selected datasets jointly cover four forms of input, four settings, and three languages, making it possible to compare performance along each of these dimensions. We use raw recordings whenever they are available and use meaningful features when a dataset was released only in a processed form. AVEC2014 is particularly useful because it provides both audio and video, and its

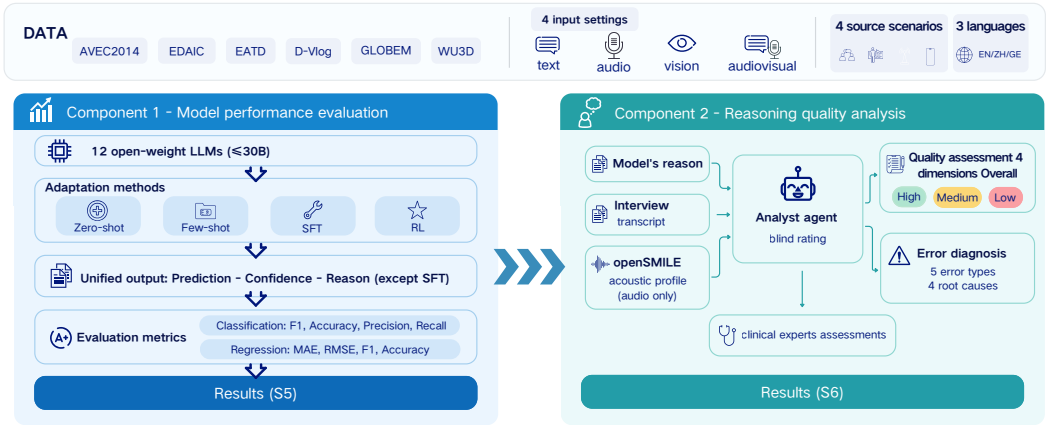


Figure 1: Overview of DEPRESSIONBENCH. The benchmark begins with a curated collection of six public depression datasets covering four forms of input, four settings, and three languages. The left side measures recognition performance: twelve open-weight LLMs of up to 30B parameters are evaluated with zero-shot and few-shot prompting, SFT, and RL. Performance is measured for both classification and severity estimation. The right side examines the reasoning reported by the models. An analyst agent rates each explanation along four dimensions without seeing the reference label and then identifies the failures behind clinically incorrect predictions.

speech can also be transcribed. We can therefore present the same participants to models as text, audio, vision, or audiovisual information, reducing the influence of differences between datasets when modalities are compared. Across the full collection, the data come from clinical interviews, daily vlogs, passive sensing, and social media in Chinese, English, and German. The Chinese and English subsets contain comparable numbers of samples.

**Label quality.** We include only datasets labeled with a clinically validated instrument, the BDI-II, PHQ-8, or SDS, or with an explicit diagnostic annotation, and exclude datasets that use indirect indicators of depression as labels. We retain several instruments because no self-report scale performs equally well in every population. A quantitative review of 53 studies involving 35,676 participants found that the internal consistency and criterion validity of commonly used depression scales varied across populations Zhao et al. (2025). For example, lower internal consistency was reported for the Patient Health Questionnaire and QIDS-SR in older adults, the Beck Depression Inventory in women with infertility, and the SDS in patients with liver cirrhosis and women living in rural areas. Restricting the benchmark to one scale could therefore make properties of that scale appear to be properties of the models. The three instruments also provide continuous scores and published thresholds. We use the thresholds for each dataset to map predicted scores into two to five clinical bands, allowing severity predictions to be assessed in relation to the categories used to interpret the scale.

**Feasibility.** Each dataset must be publicly accessible, large enough for evaluation, and accompanied by labels of sufficient quality. We exclude datasets that are not designed for depression recognition, contain too few samples, or have labels that cannot support a reliable comparison. Table 1 summarizes the form of input, setting, task, language, sample size, and recording duration of the six selected datasets.

**Text.** The text portion of the curated collection contains WU3D and GLOBEM, covering both language written by participants and textual descriptions of passive behavior. WU3D contains 2.19 million Chinese social-media posts written by 32,570 users Wang et al. (2020). Its binary labels are derived from self-reported diagnostic markers, which may make the two classes easier to separate than labels obtained through independent assessment. GLOBEM contains passive measurements from the smartphones and smartwatches of 705 students in four cohorts studied between 2018 and 2021 Xu et al. (2023). The measurements describe screen use, location, call records, sleep, and related patterns, and are accompanied by BDI-II scores. Because these data are numerical time

Table 2: Models and dataset assignments. A check mark indicates that a model is included under the corresponding condition. ZS and FS denote zero-shot and few-shot prompting, SFT denotes supervised fine-tuning, and RL denotes reinforcement learning. Dataset abbreviations are AV, AVEC2014; ET, EATD; ED, EDAIC; GB, GLOBEM; WU, WU3D; and DV, D-Vlog.

| Input setting | Model | Params | ZS/FS | SFT | RL | Datasets (CLS) | Datasets (REG) |
|---|---|---|---|---|---|---|---|
| Text-only | Qwen3-4B | 4B | ✓ | ✓ | | AV, ET, ED, GB, WU | AV, ET, ED, GB |
| Text-only | GPT-OSS-20B | 20B | ✓ | | | AV, ET, ED, GB, WU | AV, ET, ED, GB |
| Text-only | Qwen3-1.7B | 1.7B | ✓ | ✓ | ✓ | AV, ET, ED, GB, WU | AV, ET, ED, GB |
| Audio-only | Ultravox-0.5-1B | 1B | ✓ | ✓ | | AV, DV, ET, ED | AV, ET, ED |
| Audio-only | Ultravox-0.6-8B | 8B | ✓ | | | AV, DV, ET, ED | AV, ET, ED |
| Audio-only | Flamingo-Next (AF-Next) | ∼7B | ✓ | ✓ | ✓ | AV, DV, ET, ED | AV, ET, ED |
| Vision-only | Qwen3.5-4B | 4B | ✓ | ✓ | | AV, DV, ED | AV, ED |
| Vision-only | Qwen3.5-9B | 9B | ✓ | | | AV, DV, ED | AV, ED |
| Vision-only | MedGemma-4B | 4B | ✓ | | | AV, DV, ED | AV, ED |
| Vision-only | Qwen3.5-2B | 2B | | ✓ | ✓ | AV, DV, ED | AV, ED |
| Audiovisual | Gemma4-E4B | 8B | ✓ | ✓ | | AV, DV, ED | AV, ED |
| Audiovisual | Gemma4-E2B | 5.1B | ✓ | ✓ | ✓ | AV, DV, ED | AV, ED |
| Audiovisual | Phi-4-multimodal-instruct | ∼5.6B | ✓ | | | AV, DV, ED | AV, ED |

series rather than natural language, we describe the measured features in text before presenting them to a language model.

**Audio.** The audio portion contains EATD and EDAIC and allows us to assess whether models can interpret vocal information that is unavailable in a transcript. EATD contains recordings of semi-structured interviews with 162 Chinese-speaking participants Shen et al. (2022). Depression is defined by an SDS score of at least 53, giving 30 participants with depression and 132 controls. EDAIC, an extended version of DAIC-WoZ, contains 275 clinical interviews in English, each accompanied by a PHQ-8 score Ringeval et al. (2019) Gratch et al. (2014). Models receive the recordings without transcripts so that their predictions must rely on speech rather than lexical content.

**Vision.** The vision portion contains D-Vlog, which represents visible facial behavior without spoken content. The dataset consists of 961 daily vlogs with binary depression labels Yoon et al. (2022). The raw recordings are not included in the released dataset. We therefore use the provided 136-dimensional facial representation, derived from 68 dlib landmarks, together with the released OpenSMILE acoustic features where the model setting permits them.

**Audiovisual.** The audiovisual portion contains AVEC2014, which combines speech and visible behavior from the same recording. The dataset provides 300 recordings from 84 German-speaking participants completing the Freeform and Northwind tasks Valstar et al. (2014). Each recording includes audio and video and is labeled with a BDI-II score between 0 and 63. The speech can also be transcribed, allowing the same participants to be assessed from text, audio, vision, and combined audio and video. Differences among these four results can therefore be studied without changing the participants or the task.

### 3.2 Model Selection

We consider only open-weight models and select them with local use. This is because Appendix D show that there is no relationship between accuracy and LLM size. In addition, Mental-health records contain sensitive personal information, and privacy concerns may discourage people from seeking help Bell et al. (2011). A model that runs locally can keep these records on the user's device and remain available without an internet connection. This differs from many existing benchmarks that depend on proprietary models accessed through cloud services Zhang et al. (2026); Li et al. (2026). We require the weights to be publicly available so that the results can be reproduced without access to a private service, and we set an upper limit of 30B parameters to retain the possibility of deployment on devices with limited memory. For each form of input, we include models from several families and at several sizes, all designed to process the information they receive.

The models receive one of four forms of input. Text models receive social-media posts, interview transcripts, or textual descriptions of sensor measurements. Audio models receive the recording and task instructions but no transcript. Vision models receive video frames without the audio track, and

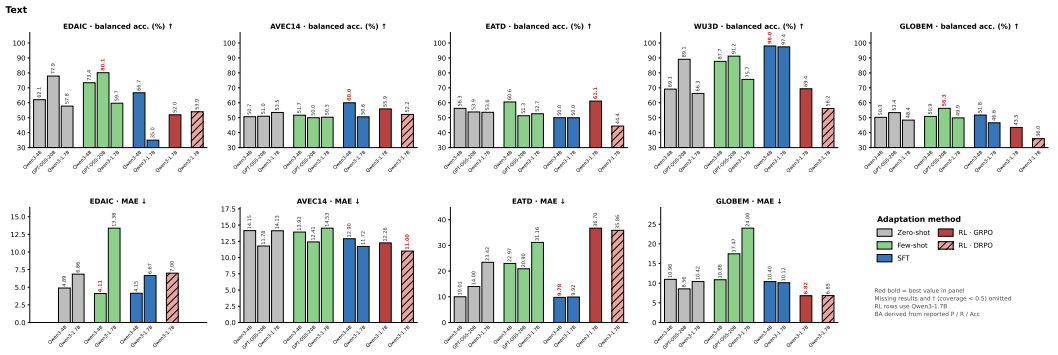


Figure 2: Results for models given text. Classification panels report balanced accuracy, and severity-estimation panels report MAE. Red values mark the best result in each panel. Missing results and results with coverage below 0.5 are omitted.

audiovisual models receive both parts of the recording. We use the raw recording whenever it is available; when it has not been released or cannot be processed by a model, we use the features provided with the dataset (§4.1). The model pool includes several sizes and families for each form of input, as well as a medically fine-tuned model in the vision group. Table 2 lists all 13 models and the datasets on which each is assessed.

Twelve models are assessed with zero-shot and few-shot prompting. Because model training requires substantially more computation than prompting, we fine-tune a smaller group. For supervised fine-tuning, we choose one larger and one smaller model for each form of input, giving eight models in total. This pairing shows whether any improvement is shared across model sizes. For reinforcement learning, we retain one smaller model from each pair. Using the same four models for supervised fine-tuning and reinforcement learning allows the two forms of training to be compared directly and tests whether additional training can improve a small model without increasing its size. Detailed experimental setup can be found at Appendix C.

## 4 Results

We compare zero-shot prompting, few-shot prompting, SFT, and RL across text, audio, vision, and audiovisual information. Most datasets are imbalanced. A model that always predicts the majority class can therefore reach high accuracy while failing to identify participants with depression. This occurs on EDAIC and EATD, where the majority-class accuracy is approximately 76% and 80%, respectively. We therefore report balanced accuracy, which gives equal weight to recall for each class. For zero-shot prompting, few-shot prompting, and RL, balanced accuracy is calculated from the confusion counts. For SFT, binary macro recall gives the same value. Accuracy is also reported when it helps reveal majority-class prediction. For severity estimation, we report MAE. MAE is interpreted within each dataset because the clinical scales have different ranges. Figures 2 to 5 show the results for each form of input. Detailed experimental results can be found at Appendix C.1.

### 4.1 Text

Figure 2 shows the results for models given text. Performance differs sharply across datasets. Under zero-shot prompting, GPT-OSS-20B reaches 77.9% balanced accuracy on EDAIC and 89.1% on WU3D. Its corresponding EDAIC accuracy is 80.9%. Few-shot prompting raises balanced accuracy to 80.1% on EDAIC and 91.2% on WU3D. Qwen3-4B shows a similar change. Its balanced accuracy rises from 62.1% to 73.4% on EDAIC and from 69.1% to 87.7% on WU3D.

The same gains do not appear on AVEC2014, EATD, or GLOBEM. Across zero-shot and few-shot prompting, balanced accuracy remains between 48% and 61%. High accuracy can also conceal majority-class prediction. On EATD, few-shot GPT-OSS-20B reaches 81.9% accuracy but only 6.5% F1.

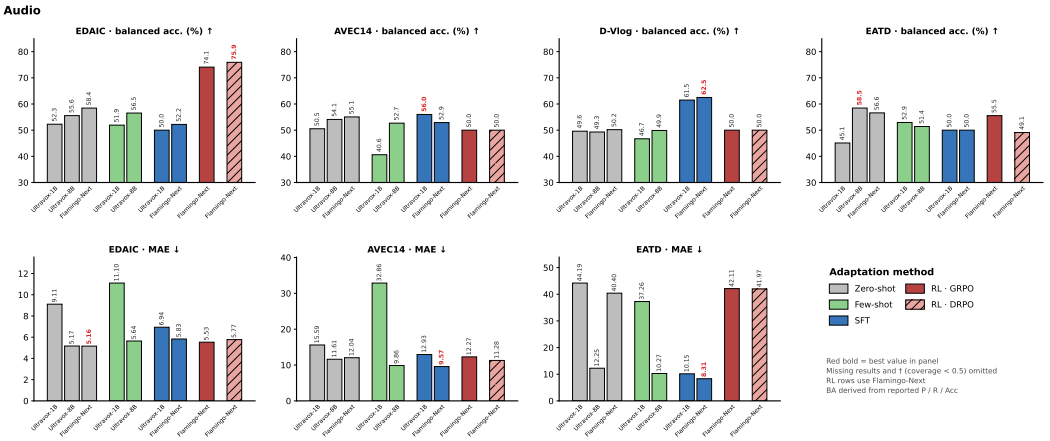


Figure 3: Results for models given audio. Classification panels report balanced accuracy, and severity-estimation panels report MAE. Red values mark the best result in each panel.

SFT gives the strongest text result on WU3D, where both Qwen3 models exceed 97% balanced accuracy. Qwen3-4B also reaches 60% balanced accuracy on AVEC2014, the only text result clearly above chance on this dataset. SFT does not improve every dataset. Both models predict only the majority class on EATD, giving 50% balanced accuracy despite 80% accuracy. On EDAIC, balanced accuracy for Qwen3-1.7B falls to 35%. For severity estimation, Qwen3-4B obtains an EDAIC MAE of 4.11 with few-shot prompting and 4.15 after SFT. SFT also gives the lowest text MAE on EATD at 9.78.

RL gives mixed results for Qwen3-1.7B. On GLOBEM, MAE falls from 10.12 after SFT to 6.82 with GRPO and 6.85 with DRPO. DRPO also gives the lowest text MAE on AVEC2014 at 11.00. Other results decline. EATD MAE rises from 9.92 after SFT to 36.70 with GRPO and 35.86 with DRPO. On EDAIC, both RL methods predict almost every participant as depressed, with 100% recall and accuracy between 17.2% and 23.3%. WU3D balanced accuracy falls from 97.4% after SFT to 69.4% with GRPO and 56.2% with DRPO.

Overall, models given text improve consistently only on EDAIC and WU3D.

### 4.2 Audio

Figure 3 shows the results for models given audio. Under zero-shot prompting, these models remain close to chance on every dataset, with balanced accuracy between 45% and 59%. Some accuracy values again reflect predictions dominated by one class. Flamingo-Next reaches 61.2% accuracy on D-Vlog with 99.3% recall, but its balanced accuracy is 50.2%.

Few-shot prompting is evaluated for the two Ultravox models and does not produce consistent classification gains. On EATD, the balanced accuracy of Ultravox-1B rises from 45.1% to 52.9%, while its accuracy falls to 25%. On AVEC2014, its MAE rises from 15.59 to 32.86. Ultravox-8B improves in severity estimation. Its MAE falls from 11.61 to 9.86 on AVEC2014 and from 12.25 to 10.27 on EATD.

SFT improves audio classification on D-Vlog. Balanced accuracy reaches 61.5% for Ultravox-1B and 62.5% for Flamingo-Next, compared with approximately 50% under zero-shot prompting. Flamingo-Next also obtains the lowest EATD MAE in the study at 8.31 and an AVEC2014 MAE of 9.57. On EDAIC and EATD classification, however, balanced accuracy remains between 50% and 52.2%.

RL improves Flamingo-Next on EDAIC. Balanced accuracy reaches 74.1% with GRPO and 75.9% with DRPO, compared with 58.4% under zero-shot prompting and 52.2% after SFT. The corresponding accuracy values are 78.8% and 81.8%. This result does not extend to the other datasets. On AVEC2014 and D-Vlog, neither RL method predicts a positive case, giving 0% F1 and 50%

balanced accuracy. EATD MAE also rises from 8.31 after SFT to 42.11 with GRPO and 41.97 with DRPO.

Overall, after SFT, Ultravox-1B and Flamingo-Next improve on D-Vlog, and Flamingo-Next obtains the lowest EATD MAE. With RL, Flamingo-Next improves only on EDAIC classification.

### 4.3 VISION

Figure 4 shows the results for models given vision. Under zero-shot prompting, these models do not distinguish reliably between participants with and without depression. Balanced accuracy ranges from 42.5% to 56.1%. MedGemma-4B reaches 75.5% accuracy on EDAIC, but its F1 is only 0.9% because it predicts almost every participant as non-depressed.

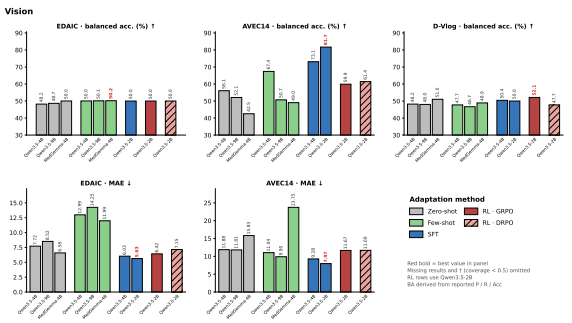


Figure 4: Results for models given vision. Classification panels report balanced accuracy, and severity-estimation panels report MAE. Red values mark the best result in each panel. Missing results and results with coverage below 0.5 are omitted.

Few-shot prompting gives one clear improvement. On AVEC2014, balanced accuracy for Qwen3.5-4B rises from 56.1% to 67.4%. Severity estimation on EDAIC becomes worse for all three models. Zero-shot MAE ranges from 6.58 to 8.52, while few-shot MAE ranges from 11.99 to 14.25.

After SFT, Qwen3.5-2B obtains the highest AVEC2014 scores. It reaches 81.7% balanced accuracy, 83.3% accuracy, 82.4% F1, and an MAE of 7.97. Qwen3.5-4B reaches 73.1% balanced accuracy and an MAE of 9.28. On EDAIC, Qwen3.5-2B obtains an MAE of 5.63, the lowest among models given vision. On D-Vlog, balanced accuracy remains between 50% and 50.4%. On EDAIC, Qwen3.5-2B reaches 50% balanced accuracy. Balanced accuracy cannot be calculated for Qwen3.5-4B because its precision and recall are unavailable. Qwen3.5-2B was not evaluated with zero-shot or few-shot prompting, so its SFT result cannot be compared with its own prompting result.

RL reduces the AVEC2014 balanced accuracy of Qwen3.5-2B from 81.7% after SFT to 59.9% with GRPO and 61.4% with DRPO. EDAIC MAE also rises from 5.63 to 6.42 and 7.15. On EDAIC classification, both RL methods predict only the majority class. Their F1 is 0%, despite an accuracy of approximately 78%.

Overall, after SFT, Qwen3.5-2B and Qwen3.5-4B reach 81.7% and 73.1% balanced accuracy on AVEC2014. Few-shot prompting and RL do not consistently improve models given vision.

### 4.4 AUDIOVISUAL

Figure 5 shows the results for models given audio and video. Gemma-4-E4B gives the strongest zero-shot result in this setting. It reaches 58% balanced accuracy on AVEC2014 and an MAE of 9.90, which is the lowest zero-shot AVEC2014 MAE among the four forms of input. Its EDAIC accuracy is 73.2%, but balanced accuracy is only 54%.

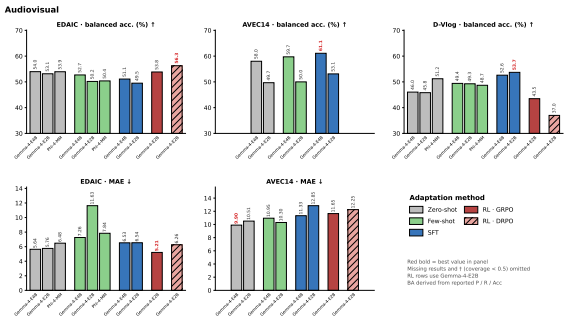


Figure 5: Results for models given audio and video. Classification panels report balanced accuracy, and severity-estimation panels report MAE. Red values mark the best result in each panel.

Table 3: Quality ratings and diagnoses of model-reported reasoning. Overall quality is reported over all responses. Failure types and root causes are reported over clinically wrong predictions only. Quality is rated without access to the reference label.

| Overall quality | | Failure type | | | Root cause | | |
|---|---|---|---|---|---|---|---|
| Level | (%) | Rank | Type | (%) | Rank | Cause | (%) |
| High | 5.9 | 1 | Symptom recognition failure | 48.4 | 1 | Reasoning process defect | 45.9 |
| Medium | 50.1 | 2 | Severity misjudgment | 19 | 2 | Missing clinical knowledge | 31.1 |
| Low | 44 | 3 | Context misunderstanding | 11.9 | 3 | Modality underutilization | 12.1 |
| | | 4 | Missing modality evidence | 11.5 | 4 | Scale misunderstanding | 10.9 |
| | | 5 | Scale mapping bias | 9.1 | | | |

Few-shot prompting raises the AVEC2014 balanced accuracy of Gemma-4-E4B to 59.7%, but does not improve the other results consistently. On EDAIC, the accuracy of Gemma-4-E2B falls to 30.9%, while MAE rises from 5.76 to 11.63.

SFT produces more balanced predictions, but performance remains modest. Gemma-4-E4B reaches 61.1% balanced accuracy on AVEC2014. Across EDAIC and D-Vlog, balanced accuracy for the two models remains between 49.5% and 53.7%. Severity estimation is also weaker than under zero-shot prompting. For Gemma-4-E4B, EDAIC MAE rises from 5.64 to 6.53, while AVEC2014 MAE rises from 9.90 to 11.33.

RL gives the lowest audiovisual EDAIC MAE, with GRPO reaching 5.21. DRPO raises EDAIC balanced accuracy to 56.3%. On AVEC2014, valid classification outputs cover fewer than half of the samples. On D-Vlog, balanced accuracy falls to 43.5% with GRPO and 37% with DRPO.

Overall, models given both audio and video do not consistently outperform models given a single input type.

## 5 Analysis of Reported Reasoning

We develop an analyst agent to determine whether each model-reported explanation is supported by the input. Appendix E gives the full design, review protocol and classic example to demonstrate the problem. For audio, the analyst also receives acoustic measurements from openSMILE. It first rates explanation quality without seeing the reference label, then diagnoses errors in clinically wrong predictions. We analyze zero-shot, few-shot, and RL explanations from AVEC2014, EATD, and GLOBEM. SFT is excluded because it returns no `reason`. Three clinical experts reviewed 50 random assessments in a blinded review and found no systematic hallucination. Table 3 shows that only 5.9% of explanations were rated High, while 44% were rated Low. Correctness rose from 40.9% for Low ratings to 66.2% for High ratings, but the two measures did not always agree. Symptom recognition accounted for 48.4% of failures. The leading causes were defects in reasoning at 45.9% and missing clinical knowledge at 31.1%. Models given audio rarely used acoustic evidence. None received sufficient grounding, about eight in ten used no acoustic evidence, and under 1% cited a measured quantity. Poor explanations were also more common in the most severe cases. Future work should improve symptom recognition and the use of acoustic evidence. This analysis is limited to model-reported explanations and the datasets and conditions examined here.

## 6 Conclusion

We introduced DepressionBench, a unified benchmark for depression recognition and reasoning analysis with open-weight language models. Models given text improve consistently only on EDAIC and WU3D; audio and vision improve mainly after SFT on isolated datasets; audiovisual input adds no consistent gain. Reported reasoning is weak: only 5.9% of explanations are rated high quality, and audio explanations almost never cite acoustic evidence. DepressionBench provides a reproducible foundation for models that recognize and reason about depression.

### AI USE STATEMENT

In this work, we used generative AI tools for one task that requires disclosure. An LLM-based analyst agent 5 supported the qualitative and thematic analysis of model-reported reasoning and the interpretation of its results. All research ideas, hypotheses, experimental design and method implementation were by us. We did not use generative AI tools for any other task that requires disclosure: Generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, interpret results. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

We release the source code. Full hyperparameters, implementation details, random seeds and computational resources are in Appendix C to ensure replicability.

Our project page is available at: https://anonymous.4open.science/r/DepreesionBenchmark-C669.

### ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

## REFERENCES

Shaykhah A. Almaghrabi, Scott R. Clark, and Mathias Baumert. Bio-acoustic features of depression: A review. *Biomedical Signal Processing and Control*, 85:105020, 2023. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2023.105020. URL https://www.sciencedirect.com/science/article/pii/S1746809423004536.

Robert A. Bell, Peter Franks, Paul R. Duberstein, Ronald M. Epstein, Mitchell D. Feldman, Erik Fernandez y Garcia, and Richard L. Kravitz. Suffering in silence: Reasons for not disclosing depression in primary care. *The Annals of Family Medicine*, 9(5):439–446, 2011. ISSN 1544-1709. doi: 10.1370/afm.1277. URL https://www.annfammed.org/content/9/5/439.

Kai Bian, Xucheng Guo, Bin Chen, Lingyan Ruan, Yiran Shen, Ting Dang, and Hong Jia. Pocket-dentist: On-device dental image understanding via efficient multimodal large language models, 2026. URL https://arxiv.org/abs/2605.29299.

Jiuyi Chen, Mingkui Tan, Haifeng Lu, Qiuna Xu, Zhihua Wang, Runhao Zeng, and Xiping Hu. Towards stable cross-domain depression recognition under missing modalities. *Pattern Recognition*, 177:113367, 2026. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2026.113367. URL https://www.sciencedirect.com/science/article/pii/S0031320326003328.

Wei Dai, Peilin Chen, Chanakya Ekbote, and Paul Pu Liang. Qoq-med: Building multimodal clinical foundation models with domain-aware grpo training, 2025. URL https://arxiv.org/abs/2506.00711.

DeepSeek-AI, :, Anyi Xu, B. Li, Bangcai Lin, Bing Xue, BingCheng Xian, Bingzheng Xu, Bochao Wu, Bowei Zhang, Boyi Deng, C. C. Yu, Chao Jin, Chaofan Lin, Chen Dong, Chenbing Wang, Chenfan Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chengyuan Zhang, Chenhao Xu, Chenqi Zhao, Chenze Shao, Chuhao Wang, Chuqi Zhang, Damai Dai, Dejian Yang, Deli Chen, Di Huang, Di Wu, Donghao Li, Erhang Li, Eric Fu, F. Zhou, Fangwei Zhou, Fangyun Lin, Fangzhou Yuan, Feiyu Xia, Fucong Dai, Guangbo Hao, Guanglin Li, Guanting Chen, Guoai Cao, Guofan Fan, Guolai Meng, Guowei Li, Haichuan Zhang, Haiyang Ma, Haiyang Shen, Han Li, Han Yu, Han Zhang, Hangyuan Deng, Hanwei Xu, Hanxiang Xu, Hanxun Zhong, Hao Guo, Hao Jiang, Hao Li, Hao Qin, Haodong Wen, Haofen Liang, Haofeng Huang, Haohua Liu, Haoling

Zhang, Haoming Luo, Haoran Yang, Haotian Xu, Haotian Yuan, Haoting Huang, Haowen Luo, Haoyang Cai, Haoyu Chen, Haozhe Ji, Hengran Zhang, Hengrui Wang, Hengxu Wu, Honghui Ding, Hongxuan Tang, Huadong Wang, Huanqi Cao, Huazuo Gao, Hui Qu, Hui Zeng, J. Yang, J. H. Jin, J. H. Zhang, J. X. Zou, Jia Yu, Jiahui Zhou, Jiajun Chen, Jialiang Huang, Jialin Zhao, Jiamin Tang, Jian Zhou, Jianan Tong, Jianwen Li, Jiaqi Zhu, Jiarui Wang, Jiasheng Ye, Jiashi Li, Jiaxin Xu, Jiaying Ding, Jibai Lu, Jiewen Hu, Jin Yan, Jincheng Zhai, Jingchang Chen, Jingcheng Hu, Jingli Zhou, Jingsheng Xu, Jingting Xiang, Jingyan Yun, Jingyang Yuan, Jingyuan Cheng, Jinhua Zhu, Jinpeng Wang, Jinyi Chen, Jinyi Hu, Jiping Yu, Jueliang Guo, Junbo Pei, Junbo Sun, Junguang Jiang, Junjie Qiu, Junkang Zhou, Junqi Liu, Junren Li, Junxian Li, Junxiao Song, Junyi Guo, Kai Dong, Kaifeng Chen, Kaige Gao, Kang Guan, Kangdong Yuan, Ke Hong, Ke Xu, Kefan Zhao, Kexin Ji, Kexin Zhang, Kexing Zhou, Kuai Yu, Lan Zhang, Lean Wang, Lecong Zhang, Lei Wang, Letian Gao, Liang Zhao, Liansheng Xu, Lihua Guo, Lingxiao Luo, Lingyue Fu, Litao Deng, Litong Wang, Liyue Zhang, Longhao Chen, Lu Chen, Luotian Huang, Luyao Ma, Luyao Wang, M. S. Di, Max Mei, Menghao Ye, Miao Cui, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingjing Zhang, Mingqi Wei, Mingshu Chen, Mingxing Liu, Mingxu Zhou, Mingyu Xu, Mingyu Yang, Mingze Wang, Muyang Chen, Ni Shentu, Ning Wang, Niufang Ning, Panpan Huang, Peixin Cong, Peiyi Wang, Peiyuan Xin, Pengfei Ren, Pengfei Yan, Pengle Zhang, Qi Kang, Qi Tang, Qiancheng Wang, Qiang Li, Qihao Zhu, Qingyang Li, Qinyu Chen, Qiushi Du, Qizhou Guo, Rongxian Xu, Rui Ding, Rui Hu, Rui Tian, Rui Yu, Ruidong Zhu, Ruifan Xu, Ruihan Yang, Ruihang Xia, Ruijie Lu, Ruilin Geng, Ruipeng Hong, Ruiqi Ge, Ruisong Zhang, Ruize Sun, Ruizhe Pan, Runji Wang, Runqian Chen, Runxin Xu, Ruohong Tian, Ruomeng Shen, Ruoyu Zhang, Ryan X., S. H. Liu, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaofei Cai, Shaoheng Nie, Shaoyuan Chen, Shengding Hu, Shengkai Lin, Shengwen Ran, Shengyu Liu, Shengyuan Jia, Shi Bai, Shi Feng, Shicheng Xu, Shichun Liu, Shiqiang Hu, Shirong Ma, Shiyu Wang, Shiyuan Feng, Shufan Gong, Shuhan Lin, Shuiping Yu, Shunfeng Zhou, Shuo Yang, Shuomeng Wang, Shuting Guo, Shuting Pan, Shuying Yu, Sinuo Cao, Siyi Lin, Sizhe Chen, Songyang Chen, Songyang Zhou, Tao Ni, Tao Yun, Tian Jin, Tian Pei, Tian Ye, Tianle Lin, Tianran Ji, Tianyi Cui, Tianyuan Yue, Tingting Yu, Tongrui Xiong, Wangding Zeng, Wei Liu, Wei Zhang, Weibin Xu, Weihao Zeng, Weilin Zhao, Wen Liu, Wenfeng Liang, Wenjie Pang, Wenjing Luo, Wenjing Yao, Wenjun Gao, Wenkai Shao, Wenkai Yang, Wenli Zhang, Wenlu Wang, Wenlve Huang, Wenqian Yan, Wentao Zhang, Xi Gao, Xiang He, Xiang Li, Xiangli Li, Xiangwen Wang, Xiangying Zhang, Xiankui Wei, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojian Qu, Xiaokang Chen, Xiaokang Zhang, Xiaotao Nie, Xiaoyao Zou, Xiaoyuan Li, Xicheng Guo, Xieting Chu, Xin Cheng, Xin Liu, Xin Xie, Xinbo Xu, Xingchao Liu, Xingchen Liu, Xingkai Yu, Xingyou Li, Xintong Yao, Xinyang Chen, Xinyong Jiang, Xinyu Yang, Xinyu Yang, Xu Chen, Xuanyu Wang, Xubei Zhong, Xuecheng Su, Xuejie Liu, Xuheng Lin, Xujie Fan, Xuncheng Zhao, Xuwei Fu, Y. C. Yan, Y. H. Jiang, Y. T. Wu, Y. W. M., Y. Z. Wang, Yafei Gao, Yang Yang, Yang Zhang, Yanru Ma, Yanwen Huang, Yao Li, Yao Li, Yao Meng, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yaoyang Ye, Yehang Yin, Yexinrui Wu, Yi Qian, Yi Tao, Yi Yu, Yichao Zhang, Yichen Jiang, Yicheng Wang, Yifan Ding, Yifan Shi, Yifeng Peng, Yifeng Zhai, Yijia Wu, Yiliang Xiong, Yilun Wang, Ying He, Ying Zhou, Yingjia Luo, Yinmin Zhong, Yiping Wang, Yisong Wang, Yixiang Zhang, Yixiao Chen, Yixuan Tan, Yixuan Wei, Yiyang Ma, Yiyao Yang, Yiyuan Liu, Yizai Cai, Yizhen Wei, Yizhi Wang, Yonglun Yang, Yongqi Zhuo, Yongqiang Guo, Yongtong Wu, Yu Wu, Yu Zhang, Yuan Bian, Yuan Cheng, Yuan Ou, Yuan Sun, Yuanfan Xu, Yuanhang Sun, Yuanhao Li, Yuchen Liu, Yuchen Yao, Yudong Han, Yuduan Wang, Yuhan Wu, Yuhao Meng, Yuheng Zou, YuKun Li, Yunchuan Wang, Yunfan Xiao, Yunfan Xiong, Yupeng Chen, Yuqian Cao, Yuqian Wang, Yuqing Chen, Yushun Zhang, Yutong Lin, Yuwei Xiao, Yuxian Gu, Yuxiang Chen, Yuxiang Huang, Yuxiang Luo, Yuxiang You, Yuxin Chen, Yuxin Xiang, Yuxuan Liu, Yuxuan Zhou, Yuyang Zhou, Yuzhe Guo, Yuzhen Huang, Yuzhuo Bai, Z. Y. Z., Zanlin Ni, Zehao Wang, Zehua Zhao, Zehui Ren, Zejun Zhao, Zhangli Sha, Zhanying Wang, Zhaochen Zhang, Zhaoshuai Du, Zhe Fu, Zhean Xu, Zhenda Xie, Zheng Liu, Zhengyan Zhang, Zhenhua Dong, Zhewen Hao, Zhibang Wang, Zhibin Gou, Zhicheng Ma, Zhihao Li, Zhihong Shao, Zhihuan Huang, Zhijie Li, Zhirui Lu, Zhixian Huang, Zhixuan Chen, Zhixuan Chen, Zhixuan Pan, Zhiyu Wu, Zhizhou Ren, Zhu He, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihui Gu, Zijia Zhu, Zili Zhang, Zilin Li, Zilong Hou, Zilong Lyu, Ziqiao Wang, Ziwei Xie, Ziya Zhang, Ziyi Gao, Zizheng Pan, Zonglin Li, Zongqing Yao, Zui Chen, Zuofan Wu, Chenchen Ling, Chengyu Hou, Chong Chen, D. Li, Di Qi, Dongjie Ji, Fang Wei, Fanyi Xia, Fei Xie, Feiyi Tan, Hailong Guo, Haiyan Zhai, Hui Zhou, Huihui Tan, Huijie Li, Jia Luo, Jia Song, Jialu Cai, Jian Liang, Jiangting Zhou, Jiaqi Gao, Jiayi Shao, Jie Chen, Jieyu Yang, Jin Chen, Jingde Zhang, Jingzi Zhou, Jinqian Wang,

Jinyang Liu, JinZhao Sun, Junhua Ling, Junmin Zheng, Kaicheng Yang, Ke Xu, Le Su, Leyi Xia, Liangfeng Ding, Lin Zhuo, Linwang Ma, Linyan Zhu, Liyu Cai, Luqi Yao, M. K. Zhang, Meng Li, Miao Lin, Miaojun Wang, Min Zhang, Mingming Li, Mingming Wang, Mingze Yin, Minmin Han, Nan Cao, Ning Wang, Ningxin Ma, Panpan Wang, Peihan Lin, Peng Sun, Peng Zhang, Qian Ying, Qiang Xiang, Qiao Wang, Qingmiao Mao, Qiwei Jiang, Rongli Jin, Ruyi Chen, Sha Tao, Shangmian Sun, Shaoqing Wu, Shichao Zou, Si Lei, Tianyang Zhang, Tianyu Sun, Tingting Yin, W. L. Xiao, Wei An, Wei Li, Wei Wang, Weiwei Lin, Wenqing Hou, X. Lin, Xiangfei Meng, Xianzhu Huang, Xiao Peng, Xiaoqian Li, Xiaoting Zhang, Xiaowen Sun, Xiaoxiang Wang, Xiaoyu Ye, Xinrou Zhang, Xinyu Zhang, Xue Cao, Xueyin Chen, Yanan Zhou, Yanhong Xu, Yao Xia, Yao Xu, Yi Shao, Yihong Zhang, Yiling Ma, Ying Tang, Yining Lou, Yiru Chen, Yishi Piao, Yixuan Chen, Yong Xiong, Yuchen Xuan, Yuehan Yang, Yuer Xu, Yukun Zha, Yunxian Ma, Yuping Lin, Yuting Yan, Yutong Xie, Yuwen Sheng, Yuxuan Zhu, Zekai Zhang, Zhe Ju, Zhenzhen Lin, Zheren Gao, Zheyang Sun, Zhigang Yan, Zhongyu Wu, Zi Wang, Zihua Qu, Ziling Yan, and Ziyi Wan. Deepseek-v4.1-flash: Pushing the limits of kv cache compression, 2026. URL https://arxiv.org/abs/2609.19969.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. The distress analysis interview corpus of human and computer interviews. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 3123–3128, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL https://aclanthology.org/L14-1421/.

Loukas Ilias and Dimitris Askounis. A cross-attention layer coupled with multimodal fusion methods for recognizing depression from spontaneous speech. In *Proc. Interspeech 2024*, pp. 912–916, 2024. doi: 10.21437/Interspeech.2024-188.

Yahan Li, Jifan Yao, John Bosco S. Bunyi, Adam C. Frank, Angel Hsing-Chi Hwang, and Ruishan Liu. Counselbench: A large-scale expert evaluation and adversarial benchmarking of large language models in mental health question answering, 2026. URL https://arxiv.org/abs/2506.08584.

Yupei Li, Shuaijie Shao, Manuel Milling, and Björn W. Schuller. Large language models for depression recognition in spoken language integrating psychological knowledge. *Frontiers in Computer Science*, 7:1629725, 2025. doi: 10.3389/fcomp.2025.1629725.

Fenglin Liu, Zheng Li, Hongjian Zhou, Qingyu Yin, Jingfeng Yang, Xianfeng Tang, Chen Luo, Ming Zeng, Haoming Jiang, Yifan Gao, Priyanka Nigam, Sreyashi Nag, Bing Yin, Yining Hua, Xuan Zhou, Omid Rohanian, Anshul Thakur, Lei Clifton, and David A. Clifton. Large language models are poor clinical decision-makers: A comprehensive benchmark. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13696–13710. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.759.

Zetian Ouyang, Yishuai Qiu, Linlin Wang, Gerard De Melo, Ya Zhang, Yanfeng Wang, and Liang He. CliMedBench: A large-scale chinese benchmark for evaluating medical large language models in clinical scenarios. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8428–8438. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.480.

Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Yanjie Fan, Weike Zhao, Zhuoxia Chen, Hongfei Gu, Chuanjin Peng, Ya Zhang, Yanfeng Wang, and Weidi Xie. Quantifying the reasoning abilities of LLMs on clinical cases. *Nature Communications*, 16:9799, 2025. doi: 10.1038/s41467-025-64769-1.

Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, Siyang Song, Shuo Liu, Ziping Zhao, Adria Mallol-Ragolta, Zhao Ren, Mohammad Soleymani, and Maja Pantic. Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and

cross-cultural affect recognition. AVEC '19, pp. 3–12, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369138. doi: 10.1145/3347320.3357688. URL https://doi.org/10.1145/3347320.3357688.

Jian Rong, Xueqin Wang, Pan Cheng, Dan Li, and Dahai Zhao. Global, regional and national burden of depressive disorders and attributable risk factors, from 1990 to 2021: results from the 2021 global burden of disease study. *The British Journal of Psychiatry*, 227(4):688–697, 2025. doi: 10.1192/bjp.2024.266.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Ying Shen, Huiyu Yang, and Lin Lin. Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model, 2022. URL https://arxiv.org/abs/2202.08210.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023. doi: 10.1038/s41586-023-06291-2.

Sackeim HA Sobin C. Psychomotor symptoms of depression. *American Journal of Psychiatry*, 154(1):4–17, 1997. doi: 10.1176/ajp.154.1.4. URL https://psychiatryonline.org/doi/abs/10.1176/ajp.154.1.4.

Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, AVEC '14, pp. 3–10, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450331197. doi: 10.1145/2661806.2661807. URL https://doi.org/10.1145/2661806.2661807.

Yiding Wang, Zhenyi Wang, Chenghao Li, Yilin Zhang, and Haizhou Wang. A multitask deep learning approach for user depression detection on sina weibo. *CoRR*, abs/2008.11708, 2020. URL https://arxiv.org/abs/2008.11708.

Yuxi Wang, Diana Inkpen, and Prasadith Kirinde Gamaarachchige. 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)*, pp. 108–126. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.clpsych-1.8.

Yihong Wu, Liheng Ma, Lei Ding, Muzhi Li, Xinyu Wang, Kejia Chen, Zhan Su, Zhanguang Zhang, Chenyang Huang, Yingxue Zhang, Mark Coates, and Jian-Yun Nie. It takes two: Your grpo is secretly dpo, 2026. URL https://arxiv.org/abs/2510.00977.

Xuhai Xu, Han Zhang, Yasaman Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Kuehn, Mike Merrill, Paula Nurius, Shwetak Patel, Tim Althoff, Margaret E. Morris, Eve Riskin, Jennifer Mankoff, and Anind K. Dey. Globem dataset: Multi-year datasets for longitudinal human behavior modeling generalization, 2023. URL https://arxiv.org/abs/2211.02733.

Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. D-vlog: Multimodal vlog dataset for depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 12226–12234, 2022. doi: 10.1609/aaai.v36i11.21483. URL https://ojs.aaai.org/index.php/AAAI/article/view/21483.

Fan Zhang, Zebang Cheng, Chong Deng, Haoxuan Li, Zheng Lian, Qian Chen, Huadai Liu, Wen Wang, Yi-Fan Zhang, Renrui Zhang, Ziyu Guo, Zhihong Zhu, Hao Wu, Haixin Wang, Yefeng

Zheng, Xiaojiang Peng, Xian Wu, Kun Wang, Xiangang Li, Jieping Ye, and Pheng-Ann Heng. Mme-emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models, 2026. URL https://arxiv.org/abs/2508.09210.

Xiangyu Zhang, Hexin Liu, Kaishuai Xu, Qiquan Zhang, Daijiao Liu, Beena Ahmed, and Julien Epps. When LLMs meets acoustic landmarks: An efficient approach to integrate speech into large language models for depression detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 146–158. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.8.

Yan-Jie Zhao, Ping-Ping Gui, Jin-Jie Xu, Tong Guo, Jing Li, Jing Wang, Gang Wang, and Linghui Meng. Exploring the differences in psychometric properties of commonly used self-rating depression scales across various populations in china: A quantitative systematic review. *Asian Journal of Psychiatry*, 111:104635, 2025. ISSN 1876-2018. doi: https://doi.org/10.1016/j.ajp.2025.104635. URL https://www.sciencedirect.com/science/article/pii/S1876201825002783.

Shuang Zhou, Wenya Xie, Jiaxi Li, Zaifu Zhan, Meijia Song, Han Yang, Cheyenna Espinoza, Lindsay Welton, Xinnie Mai, Yanwei Jin, Zidu Xu, Yuen-Hei Chung, Yiyun Xing, Meng-Han Tsai, Emma Schaffer, Yucheng Shi, Ninghao Liu, Zirui Liu, and Rui Zhang. Automating expert-level medical reasoning evaluation of large language models. *npj Digital Medicine*, 9(1):34, 2026. doi: 10.1038/s41746-025-02208-7.

# A Overview of DepressionBench

DepressionBench consists of three parts: curating a large-scale collection of depression datasets, measuring how well open-weight language models recognize depression, and examining the reasoning reported with their predictions.

- **Dataset collection.** We curate six public depression datasets that cover text, audio, vision, and audiovisual information from four settings and three languages. Their labels are derived from established depression scales rather than indirect indicators.
- **Recognition performance.** We evaluate twelve open-weight models of up to 30B parameters using zero-shot and few-shot prompting, eight after supervised fine-tuning, and four after reinforcement learning. Classification is measured by F1 and accuracy; severity estimation is measured by MAE, RMSE, and clinical-band accuracy.
- **Reported reasoning.** An analyst agent examines the explanation provided with each prediction. It rates the explanation along four dimensions without seeing the reference label, then identifies the type and underlying cause of each clinically incorrect prediction.

**Study design.** Figure 1 shows how the three parts of the benchmark are connected. We first select six public datasets that together represent clinical interviews, daily vlogs, passive sensing, and social media in Chinese, English, and German. Their labels are derived from established depression scales rather than indirect indicators. We then select twelve open-weight models of up to 30B parameters with local deployment in mind. Each model receives the same task instructions and returns an explanation, a confidence score, and a prediction; long audio and video recordings are divided into segments before assessment. The data preparation, task definitions, and measures of performance remain the same across zero-shot prompting, few-shot prompting, supervised fine-tuning, and reinforcement learning. This allows us to compare results by dataset, form of input, and training condition in the same settings.

We examine the reported explanations separately from prediction performance. The analyst agent receives the explanation produced by the model and, for audio, also receives a transcript and an acoustic profile calculated from the recording. Its assessment proceeds in two stages. First, without access to the reference label, it rates whether the explanation uses the available evidence faithfully and connects relevant symptoms to the prediction. It then sees the reference label and identifies how and why clinically incorrect predictions failed. Keeping these stages separate reduces the chance that knowledge of correctness will influence the initial assessment of explanation quality.

Table 4: Publicly available depression datasets considered in this work. "—" indicates that the corresponding information is not recorded. Modality refers to the source signal of the dataset.

| Dataset | Modality | Language | Tasks (Labels) |
|---|---|---|---|
| AVEC2014 | Video | German | Reg (BDI-II) |
| DAIC-WoZ | Video | English | Cls, Reg (PHQ-8) |
| E-DAIC | Video | English | Cls, Reg (PHQ-8) |
| CMDC | Video | Chinese | Cls (PHQ-9, HAMD-17) |
| D-Vlog | Video | English | Cls (binary) |
| LMVD | Video | Chinese, English | Cls (binary) |
| MPDD | Video | Chinese | Cls (PHQ-9, HAMD-24) |
| IDQWLT | Video | Chinese | Cls |
| EATD | Audio | Chinese | Cls (SDS) |
| PDCH | Audio | Chinese | Cls (HAMD) |
| ReMAP | Audio | German | Cls, Reg |
| EWRE | Audio | Chinese | Cls, Reg |
| MODMA | Audio + sensors | Chinese | Cls (PHQ-9) |
| StudentSADD | Audio + sensors | English | Cls (PHQ-9) |
| ChAMP | Audio + sensors | English | Cls, Reg |
| GLOBEM | Sensor | — | Cls, Reg (BDI-II, PHQ-4) |
| Brighten | Sensor | — | Cls, Reg (PHQ-9, SDS) |
| Mental Health of Adolescents during COVID-19 | Sensor | — | Cls, Reg |
| ECSMP | Sensor | — | Cls, Reg |
| OBF-Psychiatric | Sensor | — | Reg, Cls |
| BLDC/MLDC | Social media | English, Arabic | Cls (binary; 3-class) |
| DATD | Social media | English | Cls (binary) |
| WU3D | Social media | Chinese | Cls (binary) |
| Post-COVID-dataset | Social media | — | Cls (binary) |
| Student Depression Dataset | Text | — | Cls (binary) |
| BANDA | Medical | German | Reg, Cls |

## B Depression Datasets

We surveyed publicly available datasets for depression recognition to determine which could support a unified benchmark. We retained datasets that are publicly accessible, relevant to depression recognition, of sufficient scale, and annotated with a validated clinical scale or an explicit diagnostic label. Table 4 lists the 26 datasets that met these criteria, grouped by source modality. Nine further datasets were excluded because they are not publicly available, and three because they were unsuitable for our task: these are unrelated to depression, lack a depression label, or are not released at a usable scale. From this pool we selected the six datasets described in 3.1, which jointly cover the four input settings and the four source scenarios.

## C Experimental Setup

All models are assessed through the same chat interface. We use a temperature of 0.01, `top_p` of 0.01, a random seed of 42, and a maximum response length of 1,024 tokens. Internal thinking mode is disabled. Instead, the prompt asks each model to return an explicit explanation in a structured `reason` field, followed by its confidence and prediction. The prompts have the same structure across models, with only the changes needed to present text, audio, vision, or audiovisual information.

Interviews and other recordings are often too long to be presented in a single request. We therefore divide recorded media into 60-second segments for zero-shot assessment and 20-second segments for few-shot prompting, supervised fine-tuning, and reinforcement learning. Text is presented in full without segmentation. Few-shot prompts contain two examples drawn from opposite ends of the label or severity range of each dataset. The same examples are used for every model assessed on that dataset.

For supervised fine-tuning, we use 4-bit QLoRA with rank 16 and alpha 32 (Bian et al., 2026) on four H100 GPUs. Fine-tuned models return a single integer rather than the JSON object used for zero-shot and few-shot prompting. For reinforcement learning, we apply both GRPO (Shao et al., 2024) and DRPO (Dai et al., 2025) to one model for each form of input. The reward combines prediction accuracy, compliance with the required answer format, and response length. Datasets with the same form of input and task are combined during training to make better use of their limited sample sizes. We use a group size of four, following the balance between performance and computational cost reported in Wu et al. (2026).

For classification, we report micro-averaged F1, precision, recall, and accuracy. For severity estimation, we report MAE and RMSE. We also map predicted scores to the clinical bands defined for each dataset and report macro-averaged F1 and band accuracy in Appendix C.1.

### C.1 Extra Results

Table 5: Zero-shot results, text-only models. Classification reports F1, accuracy, precision and recall (%); severity estimation reports MAE and RMSE (lower is better) and the F1 (0–1) and accuracy (%) over predicted clinical bands (Band F1, Band Acc). Bold and underlined values retain the original ranking annotations (best and second best accuracy or MAE across all modalities in this regime). † denotes coverage below 0.5; 'N/A' denotes a missing result. MAE is not comparable across datasets.

*Classification*

| Model | EDAIC | | AVEC14 | | EATD | | WU3D | | GLOBEM | |
|---|---|---|---|---|---|---|---|---|---|---|
| | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy |
| Qwen3-4B | 45.4 | 45.6 | 8.6 | 50.7 | 29.0 | 72.8 | 60.0 | <u>58.2</u> | 57.0 | 41.7 |
| GPT-OSS-20B | 68.2 | **80.9** | 3.9 | 51.2 | 16.7 | **81.5** | 83.1 | **86.5** | 28.4 | **59.8** |
| Qwen3-1.7B | 42.3 | 37.2 | 18.1 | 54.7 | 26.1 | 68.5 | 57.7 | 54.2 | 53.4 | <u>42.0</u> |

*Classification (precision and recall)*

| Model | EDAIC | | AVEC14 | | EATD | | WU3D | | GLOBEM | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | Precision | Recall | Precision | Recall | Precision | Recall | Precision | Recall |
| Qwen3-4B | 30.0 | 93.9 | 58.3 | 4.7 | 28.1 | 30.0 | 43.1 | 99.1 | 40.6 | 95.2 |
| GPT-OSS-20B | 65.9 | 70.7 | 100.0 | 2.0 | 50.0 | 10.0 | 72.5 | 97.2 | 51.0 | 19.7 |
| Qwen3-1.7B | 27.0 | 96.9 | 75.0 | 10.3 | 23.1 | 30.0 | 40.7 | 98.9 | 39.6 | 82.2 |

*Severity estimation*

| Model | EDAIC | | | | AVEC14 | | | | EATD | | | | GLOBEM | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc |
| Qwen3-4B | **4.89** | 5.92 | 0.333 | 37.1 | 14.15 | 18.25 | 0.183 | 47.3 | **10.01** | 12.83 | 0.523 | 77.2 | 10.98 | 12.85 | 0.130 | 16.7 |
| GPT-OSS-20B | 3.42† | 4.93† | 0.367† | 60.6† | 11.78 | 15.56 | 0.191 | 51.7 | 14.00 | 17.72 | 0.480 | 81.5 | **8.56** | 11.31 | 0.233 | 55.6 |
| Qwen3-1.7B | 6.86 | 8.13 | 0.214 | 22.1 | 14.13 | 18.04 | 0.295 | 38.3 | 23.42 | 26.46 | 0.447 | 80.9 | <u>10.42</u> | 13.02 | 0.210 | 29.1 |

Table 6: Zero-shot results, audio-only models. Classification reports F1, accuracy, precision and recall (%); severity estimation reports MAE and RMSE (lower is better) and the F1 (0–1) and accuracy (%) over predicted clinical bands (Band F1, Band Acc). Bold and underlined values retain the original ranking annotations (best and second best accuracy or MAE across all modalities in this regime). † denotes coverage below 0.5; 'N/A' denotes a missing result. MAE is not comparable across datasets.

*Classification*

| Model | EDAIC | | AVEC14 | | D-Vlog | | EATD | |
|---|---|---|---|---|---|---|---|---|
| | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy |
| Ultravox-1B | 37.9 | 44.4 | 55.8 | 50.7 | 74.9 | <u>60.5</u> | 16.7 | 60.3 |
| Ultravox-8B | 31.9 | 68.6 | 41.2 | 54.2 | 57.4 | 50.4 | 34.0 | 61.7 |
| Flamingo-Next | 39.1 | 62.6 | 45.4 | <u>56.1</u> | 75.8 | **61.2** | 29.5 | <u>73.3</u> |

*Classification (precision and recall)*

| Model | EDAIC | | AVEC14 | | D-Vlog | | EATD | |
|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | Precision | Recall | Precision | Recall | Precision | Recall |
| Ultravox-1B | 26.3 | 68.0 | 51.2 | 61.3 | 61.7 | 95.4 | 14.7 | 19.2 |
| Ultravox-8B | 34.0 | 30.0 | 57.1 | 32.2 | 61.1 | 54.0 | 25.0 | 53.3 |
| Flamingo-Next | 31.9 | 50.5 | 54.2 | 39.1 | 61.3 | 99.3 | 29.0 | 30.0 |

*Severity estimation*

| Model | EDAIC | | | | AVEC14 | | | | EATD | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc |
| Ultravox-1B | 9.11 | 10.53 | 0.112 | 16.8 | 15.59 | 18.44 | 0.144 | 17.2 | 44.19 | 45.78 | 0.437 | 77.6 |
| Ultravox-8B | 5.17 | 6.55 | 0.229 | 34.2 | 11.61 | 14.67 | 0.222 | 37.0 | <u>12.25</u> | 16.01 | 0.502 | 73.8 |
| Flamingo-Next | <u>5.16</u> | 6.56 | 0.225 | 34.7 | 12.04 | 15.54 | 0.198 | 50.3 | 40.40 | 42.09 | 0.445 | 80.2 |

Table 7: Zero-shot results, vision-only models. Classification reports F1, accuracy, precision and recall (%); severity estimation reports MAE and RMSE (lower is better) and the F1 (0–1) and accuracy (%) over predicted clinical bands (Band F1, Band Acc). Bold and underlined values retain the original ranking annotations (best and second best accuracy or MAE across all modalities in this regime). † denotes coverage below 0.5; 'N/A' denotes a missing result. MAE is not comparable across datasets.

*Classification*

| Model | EDAIC | | AVEC14 | | D-Vlog | |
|---|---|---|---|---|---|---|
| | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy |
| Qwen3.5-4B | 34.3 | 40.6 | 68.0 | 56.0 | 60.9 | 51.3 |
| Qwen3.5-9B | 32.5 | 47.0 | 67.0 | 52.0 | 64.9 | 53.2 |
| MedGemma-4B | 0.9 | <u>75.5</u> | 51.1 | 41.7 | 71.9 | 59.1 |

*Classification (precision and recall)*

| Model | EDAIC | | AVEC14 | | D-Vlog | |
|---|---|---|---|---|---|---|
| | Precision | Recall | Precision | Recall | Precision | Recall |
| Qwen3.5-4B | 23.5 | 63.1 | 53.4 | 93.3 | 60.4 | 61.4 |
| Qwen3.5-9B | 23.6 | 51.9 | 51.0 | 97.3 | 60.5 | 70.0 |
| MedGemma-4B | 27.3 | 0.4 | 42.8 | 63.6 | 62.9 | 83.8 |

*Severity estimation*

| Model | EDAIC | | | | AVEC14 | | | |
|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc |
| Qwen3.5-4B | 7.72 | 9.23 | 0.141 | 20.5 | 11.88 | 14.44 | 0.260 | 29.7 |
| Qwen3.5-9B | 8.52 | 9.75 | 0.103 | 15.2 | 11.81 | 13.83 | 0.148 | 17.7 |
| MedGemma-4B | 6.58 | 8.54 | 0.189 | 37.3 | 15.83 | 18.54 | 0.130 | 15.5 |

Table 8: Zero-shot results, audiovisual models. Classification reports F1, accuracy, precision and recall (%); severity estimation reports MAE and RMSE (lower is better) and the F1 (0–1) and accuracy (%) over predicted clinical bands (Band F1, Band Acc). Bold and underlined values retain the original ranking annotations (best and second best accuracy or MAE across all modalities in this regime). † denotes coverage below 0.5; 'N/A' denotes a missing result. MAE is not comparable across datasets.

*Classification*

| Model | EDAIC | | AVEC14 | | D-Vlog | |
|---|---|---|---|---|---|---|
| | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy |
| Gemma-4-E4B | 23.2 | 73.2 | 45.2 | **58.0** | 69.8 | 55.0 |
| Gemma-4-E2B | 27.0 | 67.9 | 2.6 | 49.8 | 37.1 | 41.6 |
| Phi-4-MM | 29.0 | 67.7 | N/A | N/A | 71.0 | 58.5 |

*Classification (precision and recall)*

| Model | EDAIC | | AVEC14 | | D-Vlog | |
|---|---|---|---|---|---|---|
| | Precision | Recall | Precision | Recall | Precision | Recall |
| Gemma-4-E4B | 37.8 | 16.7 | 65.0 | 34.7 | 59.6 | 84.2 |
| Gemma-4-E2B | 30.5 | 24.2 | 40.0 | 1.3 | 55.4 | 27.9 |
| Phi-4-MM | 31.5 | 26.9 | N/A | N/A | 62.5 | 82.1 |

*Severity estimation*

| Model | EDAIC | | | | AVEC14 | | | |
|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc |
| Gemma-4-E4B | 5.64 | 7.10 | 0.207 | 31.0 | **9.90** | 11.75 | 0.224 | 26.7 |
| Gemma-4-E2B | 5.76 | 7.06 | 0.209 | 28.8 | <u>10.51</u> | 13.62 | 0.226 | 48.3 |
| Phi-4-MM | 6.48 | 8.62 | 0.218 | 33.7 | N/A | N/A | N/A | N/A |

Table 9: Few-shot results, text-only models. Classification reports F1, accuracy, precision and recall (%); severity estimation reports MAE and RMSE (lower is better) and the F1 (0–1) and accuracy (%) over predicted clinical bands (Band F1, Band Acc). Bold and underlined values retain the original ranking annotations (best and second best accuracy or MAE across all modalities in this regime). † denotes coverage below 0.5; 'N/A' denotes a missing result. MAE is not comparable across datasets.

*Classification*

| Model | EDAIC | | AVEC14 | | EATD | | WU3D | | GLOBEM | |
|---|---|---|---|---|---|---|---|---|---|---|
| | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy |
| Qwen3-4B | 55.2 | 66.7 | 10.0 | 51.7 | 35.6 | 70.6 | 80.4 | <u>85.5</u> | 46.9 | <u>50.2</u> |
| GPT-OSS-20B | 65.5 | **86.2** | 0.0 | 49.8 | 6.5 | **81.9** | 84.8 | **89.3** | 47.2 | **58.2** |
| Qwen3-1.7B | 44.1 | 40.5 | 1.4 | 51.9 | 12.1 | <u>81.3</u> | 65.6 | 67.6 | 57.0 | 40.8 |

*Classification (precision and recall)*

| Model | EDAIC | | AVEC14 | | EATD | | WU3D | | GLOBEM | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | Precision | Recall | Precision | Recall | Precision | Recall | Precision | Recall |
| Qwen3-4B | 40.6 | 86.2 | 72.7 | 5.4 | 29.5 | 44.8 | 70.3 | 93.7 | 41.2 | 54.5 |
| GPT-OSS-20B | 61.3 | 70.4 | 0.0 | 0.0 | 50.0 | 3.4 | 75.8 | 96.2 | 48.2 | 46.3 |
| Qwen3-1.7B | 28.5 | 96.9 | 100.0 | 0.7 | 50.0 | 6.9 | 49.4 | 97.6 | 40.3 | 97.0 |

*Severity estimation*

| Model | EDAIC | | | | AVEC14 | | | | EATD | | | | GLOBEM | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc |
| Qwen3-4B | **4.11** | 5.73 | 0.371 | 49.8 | 13.93 | 18.23 | 0.215 | 52.0 | 22.97 | 26.21 | 0.506 | 74.4 | **10.88** | 14.02 | 0.264 | 42.3 |
| GPT-OSS-20B | 3.32† | 5.01† | 0.277† | 60.9† | 12.41 | 16.10 | 0.201 | 51.7 | <u>20.90</u> | 24.84 | 0.498 | 79.6 | <u>17.47</u> | 22.19 | 0.234 | 32.0 |
| Qwen3-1.7B | 13.38 | 14.81 | 0.059 | 6.7 | 14.53 | 18.66 | 0.223 | 50.7 | 31.16 | 34.83 | 0.417 | 43.1 | 24.00 | 27.05 | 0.109 | 12.5 |

Table 10: Few-shot results, audio-only models. Classification reports F1, accuracy, precision and recall (%); severity estimation reports MAE and RMSE (lower is better) and the F1 (0–1) and accuracy (%) over predicted clinical bands (Band F1, Band Acc). Bold and underlined values retain the original ranking annotations (best and second best accuracy or MAE across all modalities in this regime). † denotes coverage below 0.5; 'N/A' denotes a missing result. MAE is not comparable across datasets.

*Classification*

| Model | EDAIC | | AVEC14 | | D-Vlog | | EATD | |
|---|---|---|---|---|---|---|---|---|
| | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy |
| Ultravox-1B | 39.1 | 36.6 | 47.0 | 41.1 | 65.3 | 52.7 | 30.4 | 25.0 |
| Ultravox-8B | 33.9 | 68.6 | 18.5 | 52.7 | 58.9 | 51.4 | 26.8 | 55.6 |

*Classification (precision and recall)*

| Model | EDAIC | | AVEC14 | | D-Vlog | | EATD | |
|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | Precision | Recall | Precision | Recall | Precision | Recall |
| Ultravox-1B | 25.6 | 82.2 | 44.7 | 49.6 | 59.9 | 71.7 | 18.1 | 95.5 |
| Ultravox-8B | 34.8 | 33.0 | 66.7 | 10.7 | 61.9 | 56.2 | 19.1 | 44.8 |

*Severity estimation*

| Model | EDAIC | | | | AVEC14 | | | | EATD | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc |
| Ultravox-1B | 11.10 | 12.75 | 0.114 | 14.2 | 32.86 | 36.43 | 0.091 | 16.6 | 37.26 | 41.38 | 0.258 | 26.0 |
| Ultravox-8B | <u>5.64</u> | 7.31 | 0.227 | 35.0 | **9.86** | 12.11 | 0.200 | 52.0 | **10.27** | 13.31 | 0.545 | 75.6 |

Table 11: Few-shot results, vision-only models. Classification reports F1, accuracy, precision and recall (%); severity estimation reports MAE and RMSE (lower is better) and the F1 (0–1) and accuracy (%) over predicted clinical bands (Band F1, Band Acc). Bold and underlined values retain the original ranking annotations (best and second best accuracy or MAE across all modalities in this regime). † denotes coverage below 0.5; 'N/A' denotes a missing result. MAE is not comparable across datasets.

*Classification*

| Model | EDAIC | | AVEC14 | | D-Vlog | |
|---|---|---|---|---|---|---|
| | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy |
| Qwen3.5-4B | 39.3 | 24.7 | 70.3 | **67.4** | 64.8 | 53.0 |
| Qwen3.5-9B | 39.4 | 24.9 | 29.7 | 50.7 | 43.0 | 43.7 |
| MedGemma-4B | 2.6 | <u>75.1</u> | 33.1 | 50.7 | 49.2 | 47.1 |

*Classification (precision and recall)*

| Model | EDAIC | | AVEC14 | | D-Vlog | |
|---|---|---|---|---|---|---|
| | Precision | Recall | Precision | Recall | Precision | Recall |
| Qwen3.5-4B | 24.5 | 99.7 | 64.6 | 77.2 | 60.4 | 69.9 |
| Qwen3.5-9B | 24.5 | 99.5 | 51.7 | 20.8 | 57.7 | 34.3 |
| MedGemma-4B | 31.0 | 1.3 | 44.6 | 26.3 | 60.7 | 41.4 |

*Severity estimation*

| Model | EDAIC | | | | AVEC14 | | | |
|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc |
| Qwen3.5-4B | 12.99 | 14.48 | 0.081 | 7.7 | 11.04 | 13.39 | 0.324 | 37.9 |
| Qwen3.5-9B | 14.25 | 15.54 | 0.045 | 5.0 | <u>9.90</u> | 12.08 | 0.293 | 31.2 |
| MedGemma-4B | 11.99 | 13.76 | 0.104 | 12.6 | 23.75 | 27.39 | 0.133 | 16.3 |

Table 12: Few-shot results, audiovisual models. Classification reports F1, accuracy, precision and recall (%); severity estimation reports MAE and RMSE (lower is better) and the F1 (0–1) and accuracy (%) over predicted clinical bands (Band F1, Band Acc). Bold and underlined values retain the original ranking annotations (best and second best accuracy or MAE across all modalities in this regime). † denotes coverage below 0.5; 'N/A' denotes a missing result. MAE is not comparable across datasets.

*Classification*

| Model | EDAIC | | AVEC14 | | D-Vlog | |
|---|---|---|---|---|---|---|
| | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy |
| Gemma-4-E4B | 31.3 | 61.0 | 45.9 | <u>59.7</u> | 75.3 | **60.7** |
| Gemma-4-E2B | 38.4 | 30.9 | 2.6 | 50.0 | 15.8 | 39.7 |
| Phi-4-MM | 37.3 | 35.2 | N/A | N/A | 66.2 | <u>54.3</u> |

*Classification (precision and recall)*

| Model | EDAIC | | AVEC14 | | D-Vlog | |
|---|---|---|---|---|---|---|
| | Precision | Recall | Precision | Recall | Precision | Recall |
| Gemma-4-E4B | 27.4 | 36.5 | 69.9 | 34.2 | 61.6 | 96.9 |
| Gemma-4-E2B | 24.6 | 88.0 | 50.0 | 1.3 | 58.3 | 9.1 |
| Phi-4-MM | 24.3 | 79.7 | N/A | N/A | 61.0 | 72.4 |

*Severity estimation*

| Model | EDAIC | | | | AVEC14 | | | |
|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc |
| Gemma-4-E4B | 7.26 | 9.30 | 0.205 | 32.4 | 10.95 | 12.99 | 0.197 | 20.5 |
| Gemma-4-E2B | 11.63 | 13.17 | 0.083 | 12.2 | 10.30 | 12.83 | 0.226 | 42.6 |
| Phi-4-MM | 7.84 | 9.51 | 0.166 | 20.4 | N/A | N/A | N/A | N/A |

Table 13: SFT (final epoch) results, text-only models. Classification reports F1, accuracy, precision and recall (%); severity estimation reports MAE and RMSE (lower is better) and the F1 (0–1) and accuracy (%) over predicted clinical bands (Band F1, Band Acc). Bold and underlined values retain the original ranking annotations (best and second best accuracy or MAE across all modalities in this regime). † denotes coverage below 0.5; 'N/A' denotes a missing result. MAE is not comparable across datasets.

*Classification*

| Model | EDAIC | | AVEC14 | | EATD | | WU3D | | GLOBEM | |
|---|---|---|---|---|---|---|---|---|---|---|
| | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy |
| Qwen3-4B | 68.7 | **83.6** | 59.6 | 60.0 | 44.4 | <u>80.0</u> | 98.2 | **98.5** | 51.7 | **52.9** |
| Qwen3-1.7B | 36.8 | 50.9 | 48.7 | 55.0 | 44.4 | 80.0 | 96.9 | <u>97.3</u> | 41.1 | <u>50.7</u> |

*Classification (precision and recall)*

| Model | EDAIC | | AVEC14 | | EATD | | WU3D | | GLOBEM | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | Precision | Recall | Precision | Recall | Precision | Recall | Precision | Recall |
| Qwen3-4B | 72.3 | 66.7 | 59.7 | 60.0 | 40.0 | 50.0 | 98.4 | 98.0 | 51.8 | 51.8 |
| Qwen3-1.7B | 40.1 | 35.0 | 50.8 | 50.6 | 40.0 | 50.0 | 96.3 | 97.4 | 43.6 | 46.6 |

*Severity estimation*

| Model | EDAIC | | | | AVEC14 | | | | EATD | | | | GLOBEM | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc |
| Qwen3-4B | **4.15** | 5.68 | 0.256 | 54.5 | 12.90 | 15.73 | 0.259 | 35.0 | <u>9.78</u> | 12.14 | 0.444 | 80.0 | <u>10.40</u> | 13.01 | 0.247 | 40.4 |
| Qwen3-1.7B | 6.67 | 8.29 | 0.221 | 36.4 | 11.72 | 14.50 | 0.264 | 45.0 | 9.92 | 12.57 | 0.549 | 72.5 | **10.12** | 13.48 | 0.211 | 46.3 |

Table 14: SFT (final epoch) results, audio-only models. Classification reports F1, accuracy, precision and recall (%); severity estimation reports MAE and RMSE (lower is better) and the F1 (0–1) and accuracy (%) over predicted clinical bands (Band F1, Band Acc). Bold and underlined values retain the original ranking annotations (best and second best accuracy or MAE across all modalities in this regime). † denotes coverage below 0.5; 'N/A' denotes a missing result. MAE is not comparable across datasets.

*Classification*

| Model | EDAIC | | AVEC14 | | D-Vlog | | EATD | |
|---|---|---|---|---|---|---|---|---|
| | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy |
| Ultravox-1B | 43.1 | 75.7 | 48.8 | 63.3 | 61.7 | <u>65.5</u> | 45.0 | **81.8** |
| Flamingo-Next | 48.2 | <u>76.2</u> | 35.5 | 45.0 | 62.6 | **66.9** | 44.4 | 80.0 |

*Classification (precision and recall)*

| Model | EDAIC | | AVEC14 | | D-Vlog | | EATD | |
|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | Precision | Recall | Precision | Recall | Precision | Recall |
| Ultravox-1B | 37.8 | 50.0 | 80.7 | 56.0 | 63.4 | 61.5 | 40.9 | 50.0 |
| Flamingo-Next | 69.8 | 52.2 | 71.6 | 52.9 | 65.1 | 62.5 | 40.0 | 50.0 |

*Severity estimation*

| Model | EDAIC | | | | AVEC14 | | | | EATD | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc |
| Ultravox-1B | 6.94 | 9.26 | 0.134 | 50.1 | 12.93 | 17.35 | 0.229 | 60.0 | 10.15 | 12.06 | 0.450 | 81.8 |
| Flamingo-Next | 5.83 | 7.22 | 0.183 | 32.2 | 9.57 | 11.79 | 0.310 | 56.7 | **8.31** | 11.82 | 0.429 | 75.0 |

Table 15: SFT (final epoch) results, vision-only models. Classification reports F1, accuracy, precision and recall (%); severity estimation reports MAE and RMSE (lower is better) and the F1 (0–1) and accuracy (%) over predicted clinical bands (Band F1, Band Acc). Bold and underlined values retain the original ranking annotations (best and second best accuracy or MAE across all modalities in this regime). † denotes coverage below 0.5; 'N/A' denotes a missing result. MAE is not comparable across datasets.

*Classification*

| Model | EDAIC | | AVEC14 | | D-Vlog | |
|---|---|---|---|---|---|---|
| | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy |
| Qwen3.5-4B | 41.2 | 50.2 | 73.8 | <u>76.7</u> | 42.1 | 60.1 |
| Qwen3.5-2B | 43.1 | 75.7 | 82.4 | **83.3** | 37.8 | 60.9 |

*Classification (precision and recall)*

| Model | EDAIC | | AVEC14 | | D-Vlog | |
|---|---|---|---|---|---|---|
| | Precision | Recall | Precision | Recall | Precision | Recall |
| Qwen3.5-4B | N/A | N/A | 80.0 | 73.1 | 51.9 | 50.4 |
| Qwen3.5-2B | 37.8 | 50.0 | 83.9 | 81.7 | 30.4 | 50.0 |

*Severity estimation*

| Model | EDAIC | | | | AVEC14 | | | |
|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc |
| Qwen3.5-4B | 6.03 | 8.29 | 0.172 | 51.3 | <u>9.28</u> | 12.05 | 0.244 | 55.0 |
| Qwen3.5-2B | <u>5.63</u> | 7.65 | 0.223 | 44.5 | **7.97** | 10.81 | 0.433 | 65.0 |

Table 16: SFT (final epoch) results, audiovisual models. Classification reports F1, accuracy, precision and recall (%); severity estimation reports MAE and RMSE (lower is better) and the F1 (0–1) and accuracy (%) over predicted clinical bands (Band F1, Band Acc). Bold and underlined values retain the original ranking annotations (best and second best accuracy or MAE across all modalities in this regime). † denotes coverage below 0.5; 'N/A' denotes a missing result. MAE is not comparable across datasets.

*Classification*

| Model | EDAIC | | AVEC14 | | D-Vlog | |
|---|---|---|---|---|---|---|
| | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy |
| Gemma-4-E4B | 51.1 | 64.9 | 60.0 | 60.0 | 52.6 | 54.6 |
| Gemma-4-E2B | 49.5 | 63.0 | 52.9 | 53.3 | 53.6 | 55.1 |

*Classification (precision and recall)*

| Model | EDAIC | | AVEC14 | | D-Vlog | |
|---|---|---|---|---|---|---|
| | Precision | Recall | Precision | Recall | Precision | Recall |
| Gemma-4-E4B | 51.2 | 51.1 | 60.9 | 61.1 | 52.6 | 52.6 |
| Gemma-4-E2B | 49.5 | 49.5 | 53.1 | 53.1 | 53.6 | 53.7 |

*Severity estimation*

| Model | EDAIC | | | | AVEC14 | | | |
|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc |
| Gemma-4-E4B | 6.53 | 8.46 | 0.200 | 37.4 | 11.33 | 15.58 | 0.301 | 48.3 |
| Gemma-4-E2B | 6.54 | 8.40 | 0.213 | 35.7 | 12.85 | 17.22 | 0.340 | 45.0 |

Table 17: RL results, text-only models. Classification reports F1, accuracy, precision and recall (%); severity estimation reports MAE and RMSE (lower is better) and the F1 (0–1) and accuracy (%) over predicted clinical bands (Band F1, Band Acc). Bold and underlined values retain the original ranking annotations (best and second best accuracy or MAE across all modalities in this regime). † denotes coverage below 0.5; 'N/A' denotes a missing result. MAE is not comparable across datasets.

*Classification*

| Model | EDAIC | | AVEC14 | | EATD | | WU3D | | GLOBEM | |
|---|---|---|---|---|---|---|---|---|---|---|
| | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy |
| Qwen3-1.7B (GRPO) | 25.0 | 17.2 | 21.1 | <u>53.1</u> | 36.4 | 57.6 | 68.6 | **63.3** | 48.8 | **36.4** |
| Qwen3-1.7B (DRPO) | 30.3 | 23.3 | 19.0 | 48.5 | 20.0 | 51.5 | 63.2 | <u>50.0</u> | 42.1 | <u>31.2</u> |

*Classification (precision and recall)*

| Model | EDAIC | | AVEC14 | | EATD | | WU3D | | GLOBEM | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | Precision | Recall | Precision | Recall | Precision | Recall | Precision | Recall |
| Qwen3-1.7B (GRPO) | 14.3 | 100.0 | 100.0 | 11.8 | 25.0 | 66.7 | 52.2 | 100.0 | 35.7 | 76.9 |
| Qwen3-1.7B (DRPO) | 17.9 | 100.0 | 66.7 | 11.1 | 14.3 | 33.3 | 46.2 | 100.0 | 32.0 | 61.5 |

*Severity estimation*

| Model | EDAIC | | | | AVEC14 | | | | EATD | | | | GLOBEM | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc |
| Qwen3-1.7B (GRPO) | 7.91† | 8.58† | 0.062† | 18.2† | 12.26 | 17.11 | 0.290 | 45.2 | <u>36.70</u> | 38.56 | 0.431 | 75.8 | **6.82** | 9.13 | 0.108 | 24.2 |
| Qwen3-1.7B (DRPO) | 7.00 | 7.74 | 0.060 | 17.6 | **11.00** | 16.54 | 0.297 | 46.7 | **35.86** | 37.57 | 0.431 | 75.8 | <u>6.85</u> | 8.45 | 0.083 | 18.2 |

Table 18: RL results, audio-only models. Classification reports F1, accuracy, precision and recall (%); severity estimation reports MAE and RMSE (lower is better) and the F1 (0–1) and accuracy (%) over predicted clinical bands (Band F1, Band Acc). Bold and underlined values retain the original ranking annotations (best and second best accuracy or MAE across all modalities in this regime). † denotes coverage below 0.5; 'N/A' denotes a missing result. MAE is not comparable across datasets.

*Classification*

| Model | EDAIC | | AVEC14 | | D-Vlog | | EATD | |
|---|---|---|---|---|---|---|---|---|
| | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy |
| Flamingo-Next (GRPO) | 53.3 | <u>78.8</u> | 0.0 | 36.4 | 0.0 | **48.5** | 28.6 | **69.7** |
| Flamingo-Next (DRPO) | 57.1 | **81.8** | 0.0 | 36.4 | 0.0 | <u>48.5</u> | 16.7 | <u>69.7</u> |

*Classification (precision and recall)*

| Model | EDAIC | | AVEC14 | | D-Vlog | | EATD | |
|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | Precision | Recall | Precision | Recall | Precision | Recall |
| Flamingo-Next (GRPO) | 44.4 | 66.7 | 0.0 | 0.0 | 0.0 | 0.0 | 25.0 | 33.3 |
| Flamingo-Next (DRPO) | 50.0 | 66.7 | 0.0 | 0.0 | 0.0 | 0.0 | 16.7 | 16.7 |

*Severity estimation*

| Model | EDAIC | | | | AVEC14 | | | | EATD | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc |
| Flamingo-Next (GRPO) | <u>5.53</u> | 7.38 | 0.200 | 36.4 | 12.27 | 15.10 | 0.095 | 15.6 | 42.11 | 43.63 | 0.431 | 75.8 |
| Flamingo-Next (DRPO) | 5.77 | 7.56 | 0.152 | 27.3 | <u>11.28</u> | 15.83 | 0.184 | 33.3 | 41.97 | 43.43 | 0.431 | 75.8 |

Table 19: RL (final epoch) results, vision-only models. Classification reports F1, accuracy, precision and recall (%); severity estimation reports MAE and RMSE (lower is better) and the F1 (0–1) and accuracy (%) over predicted clinical bands (Band F1, Band Acc). Bold and underlined values retain the original ranking annotations (best and second best accuracy or MAE across all modalities in this regime). † denotes coverage below 0.5; 'N/A' denotes a missing result. MAE is not comparable across datasets.

*Classification*

| Model | EDAIC | | AVEC14 | | D-Vlog | |
|---|---|---|---|---|---|---|
| | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy |
| Qwen3.5-2B (GRPO) | 0.0 | 78.3 | 56.3 | 52.5 | 28.6 | 47.9 |
| Qwen3.5-2B (DRPO) | 0.0 | 78.0 | 58.5 | **53.4** | 17.1 | 40.8 |

*Classification (precision and recall)*

| Model | EDAIC | | AVEC14 | | D-Vlog | |
|---|---|---|---|---|---|---|
| | Precision | Recall | Precision | Recall | Precision | Recall |
| Qwen3.5-2B (GRPO) | 0.0 | 0.0 | 41.9 | 85.7 | 62.5 | 18.5 |
| Qwen3.5-2B (DRPO) | 0.0 | 0.0 | 43.2 | 90.5 | 50.0 | 10.3 |

*Severity estimation*

| Model | EDAIC | | | | AVEC14 | | | |
|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc |
| Qwen3.5-2B (GRPO) | 6.42 | 8.90 | 0.136 | 51.3 | 11.67 | 14.66 | 0.219 | 22.0 |
| Qwen3.5-2B (DRPO) | 7.15 | 9.48 | 0.124 | 43.6 | 11.69 | 14.86 | 0.199 | 23.7 |

Table 20: RL (final epoch) results, audiovisual models. Classification reports F1, accuracy, precision and recall (%); severity estimation reports MAE and RMSE (lower is better) and the F1 (0–1) and accuracy (%) over predicted clinical bands (Band F1, Band Acc). Bold and underlined values retain the original ranking annotations (best and second best accuracy or MAE across all modalities in this regime). † denotes coverage below 0.5; 'N/A' denotes a missing result. MAE is not comparable across datasets.

*Classification*

| Model | EDAIC | | AVEC14 | | D-Vlog | |
|---|---|---|---|---|---|---|
| | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy |
| Gemma-4-E2B (GRPO) | 38.3 | 34.1 | 0.0† | 20.0† | 57.5 | 46.6 |
| Gemma-4-E2B (DRPO) | 41.9 | 39.0 | 40.0† | 25.0† | 50.7 | 39.7 |

*Classification (precision and recall)*

| Model | EDAIC | | AVEC14 | | D-Vlog | |
|---|---|---|---|---|---|---|
| | Precision | Recall | Precision | Recall | Precision | Recall |
| Gemma-4-E2B (GRPO) | 24.3 | 90.0 | 0.0† | 0.0† | 53.8 | 61.8 |
| Gemma-4-E2B (DRPO) | 27.3 | 90.0 | 100.0† | 25.0† | 48.6 | 52.9 |

*Severity estimation*

| Model | EDAIC | | | | AVEC14 | | | |
|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | Band F1 | Band Acc | MAE | RMSE | Band F1 | Band Acc |
| Gemma-4-E2B (GRPO) | **5.21** | 7.05 | 0.242 | 46.2 | 11.65 | 16.00 | 0.211 | 51.7 |
| Gemma-4-E2B (DRPO) | 6.26 | 8.61 | 0.156 | 41.0 | 12.25 | 16.37 | 0.219 | 48.3 |

Table 21: Zero-shot text-only performance of open-weight models, ordered by parameter count. Classification is measured by F1 (higher is better) and regression by MAE (lower is better). EDAIC and EATD are reported under both tasks. "—" indicates that the model was not evaluated on that dataset.

| Model | Params | AVEC14 MAE↓ | EDAIC MAE↓ | EDAIC F1↑ | EATD MAE↓ | EATD F1↑ |
|---|---|---|---|---|---|---|
| Qwen3.6-27B | 27B | 10.76 | 3.88 | 0.58 | 13.25 | 0.23 |
| Qwen3-32B | 32B | 9.60 | — | — | — | — |
| Qwen3-Omni-32B | 32B | 10.71 | 6.80 | 0.52 | 13.92 | 0.29 |
| Llama-4-Scout | 109B | 11.20 | 3.83 | 0.62 | 9.75 | 0.06 |
| Qwen3-235B | 235B | 10.39 | — | — | — | — |
| GLM-5.1 | 754B | 10.16 | 4.35 | 0.56 | 11.55 | 0.32 |

### C.2 Across Modalities

SFT provides the most consistent gains. It gives the best result on WU3D with text, AVEC2014 with vision, D-Vlog with audio, and EATD severity estimation with audio. These gains do not extend to every dataset. On EDAIC and EATD classification, SFT often produces majority-class predictions.

RL is less consistent. It gives the best audio classification result on EDAIC and the lowest MAE on GLOBEM and audiovisual EDAIC. It also produces single-class predictions in several settings and substantially worsens EATD severity estimation. GRPO and DRPO show similar behavior.

The model and input combination with the highest score differs across datasets. GPT-OSS-20B given text obtains the highest EDAIC classification result. After SFT, Qwen3-4B given text obtains the highest WU3D result, Qwen3.5-2B given vision obtains the highest AVEC2014 result, and Flamingo-Next given audio obtains the highest D-Vlog result. Accuracy alone overstates several results because many models reproduce the majority-class distribution. Balanced accuracy shows that most zero-shot and few-shot results remain close to chance outside EDAIC and WU3D.

In sum, SFT improves the largest number of model-dataset comparisons, but no model and input combination performs best on every dataset.

## D Effect of Model Scale

To test whether larger models recognise depression more accurately, we compare open-weight models of increasing size under the same zero-shot, text-only setting. Table 21 reports F1 for classification and MAE for regression on AVEC2014, EDAIC and EATD. Performance does not increase monotonically with parameter count. The largest model, with 754B parameters, obtains the best score on only one of the five metrics, and the 109B model obtains the lowest EATD F1 (0.06). Regression does not improve with scale either: AVEC2014 MAE is lowest for a 32B model, and it rises from 9.60 for the 32B model to 10.39 for the 235B model.

## E Analyst Agent

### E.1 Design Motivation

Previous experiments 4 D show that giving the audio or vision models more modality information does not improve recognition performance, and in some settings it degrades it. The bottleneck may therefore lie less in model capacity than in how models treat the available evidence: they may ignore it, or use it incorrectly. Judging why a model reaches its prediction requires an external tool that assesses the quality of its reasoning.

We design that tool from the literature Almaghrabi et al. (2023). Audio-based depression recognition follows two representation routes: end-to-end representations learned from raw speech, and handcrafted acoustic descriptors. The former include self-supervised speech models Ilias & Askounis (2024); the latter include fundamental-frequency mean and range, jitter, shimmer, speech rate, pause ratio and loudness extracted with openSMILE Chen et al. (2026). These descriptors can be named and quantified, so a model's verbal account can be checked against them item by item.

Because they are also widely used as acoustic markers of depression, they provide a verifiable reference for whether audio reasoning is grounded in evidence. The analyst therefore sets two criteria: the acoustic dimension against openSMILE markers, and the clinical dimension against the symptoms of the scale each dataset uses. Valid reasoning must also cite evidence that is present, follow a symptom-to-severity-to-score chain, and read the clinical bands correctly. The analysis thus answers two questions, how good the reasoning is and why it fails E.2.

Visual cues have been studied far more than acoustic ones, and whether an LLM actually uses acoustic evidence remains largely unexplored. We therefore focus on the acoustic dimension and keep text as a contrast. To check that the analyst's judgments are trustworthy, we randomly sample 50 of its assessments. Three speech-language rehabilitation and clinical experts review them blind to the model and the condition. They judged the analyses clinically sound and found no systematic hallucination.

### E.2 Agent Framework

We build an analyst agent based on DeepSeek-AI et al. (2026) that acts as both a clinical psychologist and a diagnostic analyst. Each sample provides a task context $t$ (the transcript or sensor-derived text, or a textual form of the recording), a ground-truth score $y$, a predicted score $\hat{y}$, the model's `reason` $r$, the input setting $m \in \{\text{TEXT-ONLY}, \text{AUDIO-ONLY}\}$, and, for audio samples, an objective acoustic profile $a$. The profile is computed with openSMILE — for example fundamental frequency mean and range, jitter, shimmer, speech rate, pause ratio, and loudness — rather than generated by another language model, so that the acoustic claims in $r$ are checked against measured values.

The analyst works in two passes, in this order. The quality pass is blind to the label: it reads only $t$, $r$, $m$, and $a$, and returns

$$\mathcal{A}_q = \text{Analyst}_q(t, r, m, a) = \big(q_1,\ \ldots,\ q_4,\ q_{\text{overall}}\big), \tag{1}$$

where $q_1, \ldots, q_4$ are the four dimension scores and $q_{\text{overall}} \in \{\text{High}, \text{Medium}, \text{Low}\}$ summarizes them. Because $y$ and $\hat{y}$ are withheld, a rating cannot be inferred from whether the prediction happens to be correct. The error pass runs only when the prediction is clinically wrong, and then additionally reads $y$, $\hat{y}$, and the quality ratings, returning

$$e = \text{Analyst}_e(t, r, m, a, y, \hat{y}, \mathcal{A}_q) = (l,\ c,\ k), \tag{2}$$

where $l$ localizes the failure point, $c$ is the failure type, and $k$ the underlying factor. The full assessment is thus $\mathcal{A} = (\mathcal{A}_q, e)$, with $e$ empty for clinically correct predictions. We use DeepSeek 4.1 Flash, a later-generation model distinct from every model under evaluation.

The quality pass rates four dimensions, each on a four-level ordinal scale. *Clinical Grounding* (for text) and *Acoustic Grounding* (for audio) ask whether the reasoning draws on clinically meaningful evidence — the symptom criteria of the dataset's diagnostic scale for text, and the acoustic markers of depression for audio — rather than on surface emotion words. *Evidence Fidelity* asks whether the evidence cited in $r$ faithfully reflects the transcript or the acoustic profile, rather than being distorted or invented. *Diagnostic Depth* asks whether the reasoning moves through a symptom-to-severity-to-score chain instead of jumping from a vague impression to a number. *Scale Comprehension* asks whether the model reads the rating scale, its range, and its clinical bands correctly. These four ratings are combined into an overall judgement of High, Medium, or Low.

The error pass runs only when the prediction is clinically wrong — for classification, when the predicted label differs from the ground truth, and for regression, when the predicted score falls into a different severity band under the conversion — and proceeds in three steps that fill $l$, $c$, and $k$. First, it localizes the failure: it traces the reasoning and quotes the step that first departs from a correct clinical assessment, and explains what is wrong at that step. Second, it classifies the failure as one of five types: Symptom Recognition Failure, Severity Misjudgment, Scale Mapping Bias, Context Misunderstanding, or Missing Modality Evidence. Third, it identifies the underlying factor, chosen from missing clinical knowledge, modality underutilization, a defect in the reasoning process, or a misunderstanding of the scale. The type describes the observable break in the reasoning whereas the factor describes the capability gap behind it. Definitions follow the rubric conventions of prior work.

We then aggregate the per-sample assessments and focus on the problems that recur most often, so that the analysis points to systematic rather than incidental weaknesses. We apply the analyst to

the zero-shot, few-shot, and RL reasoning (SFT is excluded because it returns only a score and no `reason`) on AVEC2014 (its Text-only and Audio-only settings), EATD, and GLOBEM, and report the results 3.

### E.3 Analysis Results

The analyst was applied to the zero-shot, few-shot and reinforcement-learning predictions and evaluated each in two passes E.2. The first rated the quality of the reasoning, blind to the label; the second diagnosed why clinically wrong predictions failed.

**Accuracy is a poor proxy for reasoning quality.** Each response received an overall rating of High, Medium or Low. Reasoning was weak overall, and its relation to correctness was loose. The probability of a correct prediction rose only from 40.9% for Low-quality reasoning to 66.2% for High-quality reasoning, compared with 47.6% overall. Beyond this gradient, the two signals diverged in both directions: many correct predictions were rated Low, and some high-rated responses were still wrong. A fluent and clinically plausible justification can therefore accompany a wrong answer, and reasoning quality should be reported separately from accuracy.

**Failures center on symptom recognition.** The leading root cause was a defect in the reasoning process itself, and the full error taxonomy appears in Table 3. Symptom Recognition Failure was the most common type by a wide margin (48.4%). Models often failed to identify symptoms in the evidence and often could not connect them into a symptom-to-severity-to-score chain. On EATD, for example, a model dismissed irritability and interpersonal conflict as mild distress while anchoring on the absence of classic sadness. The chain thus broke at its first step, the recognition of symptoms.

**Audio models barely use acoustic evidence.** Audio is the one modality whose evidence cannot be recovered from text, and it is where the failure to use available evidence is sharpest. By the analyst's rating, no audio response grounded its reasoning in acoustic evidence, and about eight in ten used none at all. An independent keyword scan agrees: only about one in ten responses mentioned an acoustic feature at all, and fewer than 1% drew on a measured quantity such as jitter or pause ratio. The three audio models failed in two different ways. The Ultravox models barely mentioned acoustic features and reasoned from the transcript, or not at all. FlamingoNext was the exception: within its own samples it cited acoustic evidence 54% of the time. These citations were shallow, however, often contradicted the measured profile, and were still not credited as sufficient grounding. Mentioning acoustic evidence is therefore not the same as using it. This pattern is consistent with the benchmark-level result that added modalities do not improve recognition, although it does not establish that the unused channel causes it.

**Reasoning is weakest for the most severe cases.** Among the regression samples, the share of poor-quality reasoning was stable across the lower severity bands but rose in the most severe band, the cases in which sound reasoning matters most. Even there the failure took the same form, a break at symptom recognition.

### E.4 Review Protocol

To check that the ratings are trustworthy, we sample 50 assessments at random. Three experts in speech-language rehabilitation and clinical psychology review them blind to the model and the condition. They judged the analyses clinically sound and found no systematic hallucination.

### E.5 Case Examples

We give representative examples, grouped by the input setting in which they arose.

**Audio-only.** Audio is the setting whose evidence a transcript cannot recover, and there the break is clearest. In one AVEC2014 sample with ground truth 29, an Ultravox-8B model returns 14. It reasons that "the subject's narrative about enjoying a family meal and being well-fed as a child suggests a positive emotional state." The model guesses affect from content and ignores the measured profile. That profile points to psychomotor retardation: a pause ratio of 0.73, a voiced ratio of 0.27, an HNR of 6.35 dB and a jitter of 0.045.

A Flamingo-Next run on two Freeform samples with ground truth 34 and 27 returns 9 and 7. On the first it invents "some background noise and slight distortion" although the audio-quality flag is normal, and it misses the real pause ratio of 0.65. On the second it reasons from semantics alone. It calls the absence of any mention of sleep, appetite or concentration problems evidence that "there is no evidence of clinically significant depression," while a pause ratio of 0.62 and an HNR of 8.3 dB go unused.

The same model can also over-read. On an EATD sample with no depression it turns a specific "I don't want to listen to him speak" into "a sense of hopelessness and resignation" and predicts depression. In every case the acoustic channel is either ignored or read against the recording.

**Text-only.** Text-only models fail on symptoms that are easy to overlook. On an EATD sample with an SDS index of 58.75 the model dismisses a conflict with a cafeteria worker. It reasons that "the negative segment suggests mild distress, but without more detailed emotional expressions or symptoms, the score cannot be accurately estimated," and returns 35. Irritability and anger are core SDS symptoms, so the conflict should have raised the score rather than been dropped.

A second EATD sample shows the same pattern from the opposite anchor. With ground truth 66.25 the model notes "interpersonal conflicts" but concludes that "there is no explicit mention of depressive symptoms such as persistent sadness, loss of interest, or fatigue," and returns 45. Anchoring on the absence of classic sadness, it treats hostility and friction as background rather than as symptoms.

In an AVEC2014 sample with ground truth 23, the model reads "vivid, nostalgic recollections of childhood joy and friendship" as showing "no signs of significant depressive symptoms" and returns 2. BDI-II scores the past two weeks, and positive reminiscence coexists with depression. Recalling a happy childhood is therefore not evidence against current symptoms.

A final EATD sample shows a subtler cost. With ground truth 53.75 the model performs the SDS conversion correctly, discounts a broken friendship as not severe, and returns 38, just below the clinical line.

**Passive sensing.** GLOBEM provides only behavioural proxies, and the model treats them as if they covered the disorder. In a sample with ground truth 15 the reasoning is that "the combination of high energy expenditure and high-quality, uninterrupted sleep strongly suggests a healthy psychological state with no signs of the lethargy, insomnia, or anhedonia associated with depression." The estimate is 0. Core BDI-II symptoms such as low mood, guilt and suicidal ideation leave no trace in step counts or sleep, so normal behaviour cannot exclude mild depression.

**Reinforcement learning.** Reinforcement learning produces a failure the other settings do not show: the reasoning collapses onto the scale. On EATD regression samples with ground truth 41.25, 53.75 and 76.25, the model returns 10, 10 and 12. It asserts in turn that "SDS is a 10-point scale, 10 is max," that the SDS score is the "sum of negative items," and that SDS stands for "Depression Severity Disorder" with one point per item. These are not prompt artefacts. They are fabricated facts about a scale the model does not know, and they replace the symptom-to-score step entirely.

Across the examples the reasoning breaks at different steps of the clinical chain. All share one pattern: the reasoning does not connect the available evidence to the clinical judgement.

## F Discussion and Limitations

We present DepressionBench, a unified benchmark for depression recognition with open-weight LLMs and a companion agent that examines the reasoning behind their predictions. Across four input settings, six datasets and four training conditions, current models do not recognize depression reliably, and the reasoning they produce does not make their predictions trustworthy.

Two gaps stand out. First, models succeed only where the evidence is textual and social: on social-media text they approach the ceiling, but on clinical interviews their audio and vision models remain near the majority-class baseline. This pattern suggests that they rely on the language in the input rather than on the acoustic and visual evidence that defines these settings. Second, training does not close the gap. Supervised fine-tuning reliably improves classification, mainly on text, but it does not

transfer to audio or vision and does not make severity regression work, so its gains remain narrow rather than general. Reinforcement learning does not improve on this supervised starting point and in several settings performs worse; a possible reason is that its reward, which scores answer accuracy, format and length, can be satisfied without reasoning about symptoms, the step at which the supervised models already fail.

The reasoning analysis shows why accuracy alone is a poor guide. Reasoning quality and correctness are only loosely related, so a correct prediction does not imply a sound rationale, and the two should be reported separately. Wrong predictions fail at the first step of the clinical chain: models do not identify the relevant symptoms and cannot connect them into a symptom-to-severity judgment. The failure is sharpest for audio, where the evidence cannot be recovered from a transcript. Most responses ignore the acoustic signal altogether, while a minority refer to it but misinterpret what they hear, citing features that contradict the recording. Neither pattern connects the available evidence to the clinical judgment.

These results define what the benchmark can and cannot measure. All labels come from self-report scales rather than clinical assessment, so DepressionBench measures recognition performance, not diagnostic validity, and its scores should not be read as clinical evidence. What it offers is a reproducible way to measure where current models fall short and to compare attempts to improve them.

The results also point to a concrete next step. Current models underuse modality-specific evidence and lack the clinical knowledge to reason from symptoms to severity. We therefore intend to use DepressionBench to train an edge-side model, deployable on mobile devices, that makes effective use of multimodal features and integrates depression-related knowledge to reason correctly.

The main limitation is the labels: they come from self-report scales rather than clinician diagnosis, and datasets with verified clinical diagnoses remain scarce, so the benchmark can measure recognition but cannot establish diagnostic validity.