Unveiling the Landscape of Clinical Depression Assessment: From Behavioral Signatures to Psychiatric Reasoning

Zhuang Chen[1], Guanqun Bi[2*], Wen Zhang[3*], Jiawei Hu[4], Aoyun Wang[1], Xiyao Xiao[5], Kun Feng[6*], Minlie Huang[2*]
[1]School of Computer Science and Engineering, Central South University
[2]CoAI Group, DCST, IAI, BNRIST, Tsinghua University
[3]University of International Relations
[4]Central China Normal University
[5]Lingxin AI
[6]Yuquan Hospital, Tsinghua University
zhchen18@foxmail.com, biguanqun@mail.tsinghua.edu.cn

Abstract

Depression is a widespread mental disorder that affects millions worldwide. While automated depression assessment shows promise, most studies rely on limited or non-clinically validated data, and often prioritize complex model design over real-world effectiveness. In this paper, we aim to unveil the landscape of clinical depression assessment. We introduce C-MIND, a *clinical multimodal neuropsychiatric diagnosis* dataset collected over two years from real hospital visits. Each participant completes three structured psychiatric tasks and receives a final diagnosis from expert clinicians, with informative audio, video, transcript, and functional near-infrared spectroscopy (fNIRS) signals recorded. Using C-MIND, we first analyze *behavioral signatures* relevant to diagnosis. We train a range of classical models to quantify how different tasks and modalities contribute to diagnostic performance, and dissect the effectiveness of their combinations. We then explore whether LLMs can perform *psychiatric reasoning* like clinicians and identify their clear limitations in realistic clinical settings. In response, we propose to guide the reasoning process with clinical expertise and consistently improve LLM diagnostic performance by up to 10% in Macro-F1 score. We aim to build an infrastructure for clinical depression assessment from both data and algorithmic perspectives, enabling C-MIND to facilitate grounded and reliable research for mental healthcare.

1 Introduction

Depression is a widespread and serious mental disorder that places a heavy burden on individuals and public health systems worldwide. While automated assessment shows promise for offering objective and scalable support, its real-world clinical utility remains limited due to a lack of clinically grounded data (Cummins et al. 2015; Sarsam et al. 2024). Most widely used datasets rely on self-reported questionnaires rather than diagnoses made by trained clinicians (Gratch et al. 2014; Tadesse et al. 2019). Even the few pioneering studies that include clinical diagnoses often suffer from small sample sizes (< 30 patients) and limited behavioral tasks or modalities (Cai et al. 2022; Zou et al. 2022). These constraints lead many studies to focus on sophisticated model design in controlled settings instead of addressing the full complexity of real clinical data. As a result, a clear picture of what effective automated clinical depression assessment entails has yet to emerge (Sarsam et al. 2024).

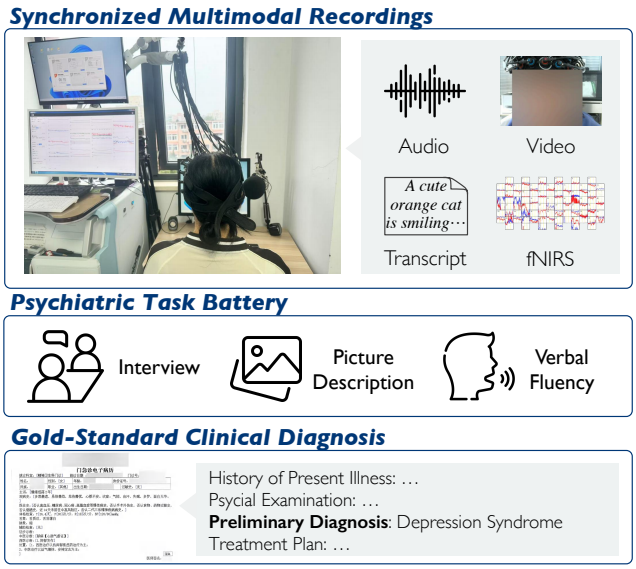

Figure 1: C-MIND integrates multimodal recordings from psychiatric tasks with clinical diagnosis.

In this paper, we aim to unveil this landscape through a three-pronged investigation: 1) establishing a new, clinically grounded data foundation, 2) analyzing the core behavioral signatures, and 3) advancing clinically guided psychiatric reasoning for diagnosis. First, we introduce **C-MIND**: the Clinical Multimodal Neuropsychiatric Diagnosis dataset. Over a two-year period, we build this dataset from a real hospital setting. It comprises 169 participants who each complete three distinct psychiatric tasks, including Interview (Gratch et al. 2014), Picture Description (Ramponi et al.

*Corresponding author.

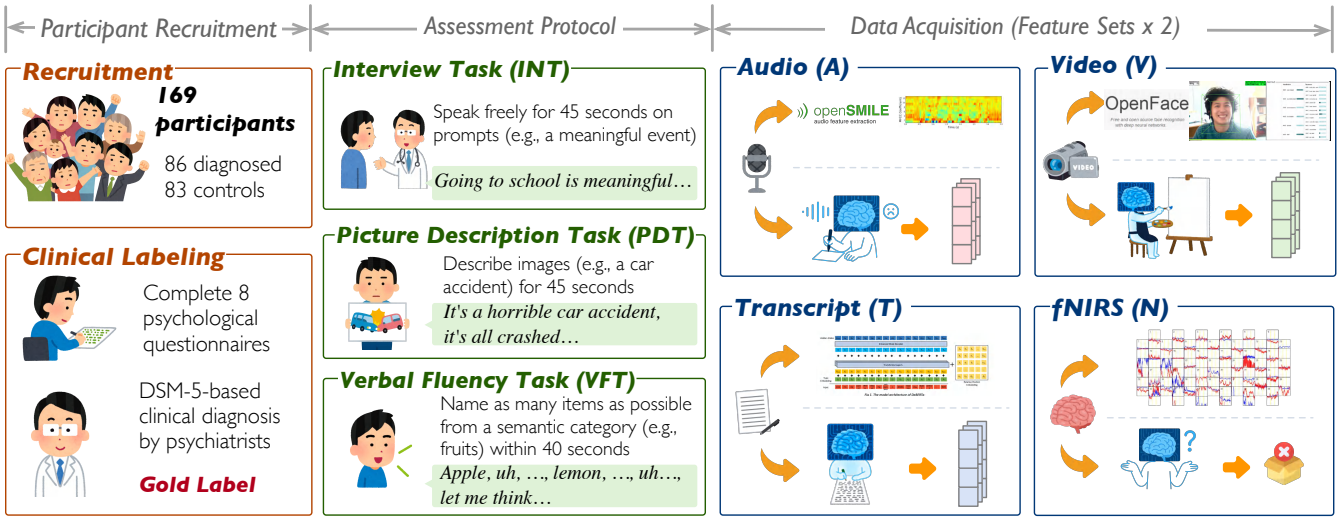

Figure 2: C-MIND collection pipeline, outlining participant recruitment, assessment protocol, and data acquisition.

2010a), and Verbal Fluency (Fossati et al. 2003b). We capture four synchronized modalities (Audio, Video, Transcript, and fNIRS (Cui et al. 2011)) for each session. Crucially, every participant receives a face-to-face diagnostic interview with senior psychiatrists, whose final clinical diagnosis, made according to DSM-5 (American Psychiatric Association 2013) criteria, serves as the gold-standard ground truth. The dataset is further enriched with detailed medical records and a battery of eight psychometric questionnaires. C-MIND's scale, clinical grounding, and richness in tasks and modalities far exceed previously available resources.

Leveraging C-MIND, we conduct an in-depth analysis of **behavioral signatures**, defined as observable patterns in speech, facial expression, and neural activity indicative of depressive states. We train a range of modeling backbones to systematically quantify the diagnostic value of different tasks and modalities, revealing that audio and video are the most informative, while the picture description task best elicits depressive markers. Fusing modalities or tasks further enhances performance and robustness, providing clear empirical guidance for designing future assessment systems.

Beyond analyzing predictive signals, we explore whether Large Language Models (LLMs) perform **psychiatric reasoning**. We evaluate seven top-tier text and multimodal LLMs and find clear limitations in their ability to handle real-world clinical data. In response, we propose a novel method that guides the LLM's reasoning process using structured clinical expertise. This approach significantly boosts diagnostic performance by up to 10% in Macro-F1 score, demonstrating a promising direction for developing clinically informed computational models.

Our main contributions can be summarized as follows:

- We introduce C-MIND, a clinically validated depression diagnosis dataset with rich tasks and modalities.
- We provide a comprehensive analysis of behavioral signatures, offering clear, data-driven insights into the discriminative power of different tasks and modalities.
- We demonstrate the limitations of LLMs in clinical assessment and propose a novel psychiatric reasoning mechanism that significantly boosts performance.

We believe this work builds a critical infrastructure for the field and provides a blueprint for developing computational systems that are not only effective, but also clinically grounded and trustworthy.

2 C-MIND Collection

To ensure ecological validity, data quality, and clinical reliability of C-MIND, we follow a comprehensive collection protocol. Below, we describe participant recruitment, assessment procedures, and data acquisition in detail.

2.1 Participant Recruitment and Demographics

We recruited participants from December 2022 to April 2025 at the psychiatric department of Yuquan Hospital, Tsinghua University. Recruitment was conducted through internal announcements. Volunteers who met the inclusion criteria and were evaluated by a chief psychiatrist together with an associate chief psychiatrist according to DSM-5 (American Psychiatric Association 2013) were invited to participate after providing written informed consent.

Statistic	Depression	Control	Total
Subject	86	83	169
Gender (M/F %)	36.05 / 63.95	44.58 / 55.42	40.34 / 59.66
Age (Mean ± SD)	33.49 ± 16.47	32.47 ± 10.90	32.99 ± 13.98
Duration (s)	171.84	125.01	152.16
Word Count	392	519	445

Table 1: Detailed statistics of C-MIND.

All procedures receive full approval from the university's Institutional Review Board (IRB), and strict measures are in place to protect participant confidentiality. The final cohort consists of 169 participants, including 86 individuals

Availability	Dataset	Language	Subj. (MDD/HC)	Tasks	Modalities	Ground Truth Labels
No	Oizys (Lin et al. 2022)	Chinese	103 (56/47)	READ	A	C.D., HAMD-17
	Guo et al. (2021)	Chinese	208 (104/104)	INT, READ, PDT	A, V	PHQ-9, BDI
	Liu et al. (2021)	Chinese	50 (25/25)	INT	A	BDI
	DEPAC (Tasnim et al. 2022)	English	552 (134/418)	VFT, PDT, READ	A	PHQ-9, GAD-7
Yes (*w/o C.D.*)	DAIC-WOZ (Gratch et al. 2014)	English	142 (42/100)	INT	A, V, T	PHQ-8
	EATD (Shen et al. 2022)	Chinese	162 (30/132)	INT	A, T	SDS
Yes (*with C.D.*)	MODMA (Cai et al. 2022)	Chinese	53 (24/29)	INT, READ, PDT	A, EEG	C.D., PHQ-9
	CMDC (Zou et al. 2022)	Chinese	78 (26/52)	INT	A, V, T	C.D.$^-$, HAMD-17, PHQ-9
	C-MIND (Ours)	Chinese	**169 (86/83)**	**INT, PDT, VFT**	**A, V, T, fNIRS**	**C.D.$^+$, 8 Questionnaires**

Table 2: Comparison of depression datasets. C.D.=Clinical Diagnosis, INT=Interview, PDT=Picture Description, VFT=Verbal Fluency, READ=Reading, A=Audio, V=Video, T=Transcript, fNIRS=functional near-infrared spectroscopy. "C.D.$^-$" means the control group is not confirmed by clinical diagnosis. "C.D.$^+$" means our C-MIND further provides detailed medical records.

diagnosed with Major Depressive Disorder (MDD) and 83 healthy controls (HC). Table 1 presents detailed statistics of the C-MIND cohort, including group size, gender distribution, age, average speech duration, and word count.

2.2 Assessment Protocol

The assessment protocol for each participant includes two main parts: a formal face-to-face clinical diagnosis (used to obtain clinically validated labels), and a series of psychiatric tasks (used to collect rich multimodal behavioral signatures). Due to space constraints, we provide detailed experimental materials, guidelines, and procedures in the Technical Appendix.

Clinical Diagnosis Participants are interviewed by a clinical team comprising a chief and an associate chief psychiatrist, each with over ten years of experience. The team conducts a face-to-face diagnostic interview and makes a high-confidence diagnosis based on DSM-5 criteria. A detailed medical record is maintained for each participant. Participants also complete a battery of eight psychometric questionnaires, including: *HAMD* (Hamilton 1960), *HAMA* (Hamilton 1959), *SDS* (Zung 1965), *SAS* (Zung 1971), *PSQI* (Buysse et al. 1989), *16PF* (Cattell, Eber, and Tatsuoka 1970), *SCL-90* (Derogatis 1977), and *HCL-32* (Angst et al. 2005). These instruments assess depressive symptoms, anxiety, sleep quality, and personality traits. In this study, we use only the clinical diagnosis as the ground-truth label; questionnaire data are reserved for future work.

Psychiatric Tasks All tasks take place in a quiet, controlled laboratory with ambient noise below 60dB. Participants sit in front of a monitor displaying instructions and are asked to remain seated, minimize movement, and maintain a fixed distance from the microphone (approx. 20 cm). We design three structured tasks that elicit cognitive and emotional markers of depression:

- **Interview Task (INT):** Participants speak for 45 seconds in response to autobiographical prompts. This task elicits emotional expression and narrative patterns indicative of depression (e.g., negative sentiment, frequent first-person use) (Rinaldi, Tree, and Chaturvedi 2020). Prompts include: "something that made you angry," "a meaningful event," and "your favorite food."
- **Picture Description Task (PDT):** Participants describe a given image for 45 seconds. This task captures visual interpretation and emotional valence. Depressed individuals may show negative bias or limited detail (Ramponi et al. 2010b). Images include: "a cat," "a car accident," and "a spaceship."
- **Verbal Fluency Task (VFT):** Participants list items in a semantic category (e.g., "fruits") within 40 seconds. This evaluates semantic memory and executive function, which are often impaired in depression (Akiyama et al. 2018; Fossati et al. 2003b). Categories include: "four-legged animals," "fruits," "cities," and "vegetables."

In total, each participant completes ten tests across the three tasks. We record four synchronized modalities during the psychiatric tasks. We use a studio microphone to capture high-quality audio (44.1kHz, 24-bit, WAV). A camera records video (640x480, 30fps), and a functional near-infrared spectroscopy (fNIRS) device measures blood oxygenation changes in the prefrontal cortex. Each of the 10 tasks is recorded separately in audio. Using timestamps, we segment the continuous video and fNIRS recordings to align all modalities. Transcripts are generated using a commercial speech recognition system and are manually proofread by human annotators to ensure accuracy.

2.3 Comparison with Existing Datasets

To situate C-MIND within the current research landscape, we compare it with other depression datasets collected in controlled environments, while excluding those based on social media data (Yoon et al. 2022).

As summarized in Table 2, many existing datasets (e.g., DAIC-WOZ (Gratch et al. 2014)) rely on self-report instruments like PHQ-8 and lack clinical validation. While MODMA (Cai et al. 2022) and CMDC (Zou et al. 2022) include clinical diagnoses, they have smaller sample sizes (53/78 participants with only 24/26 patients, respectively), limited tasks, and fewer modalities.

In contrast, C-MIND goes far beyond existing resources in every aspect: it offers a larger and balanced sample size, more diverse psychiatric tasks (INT, PDT, VFT),

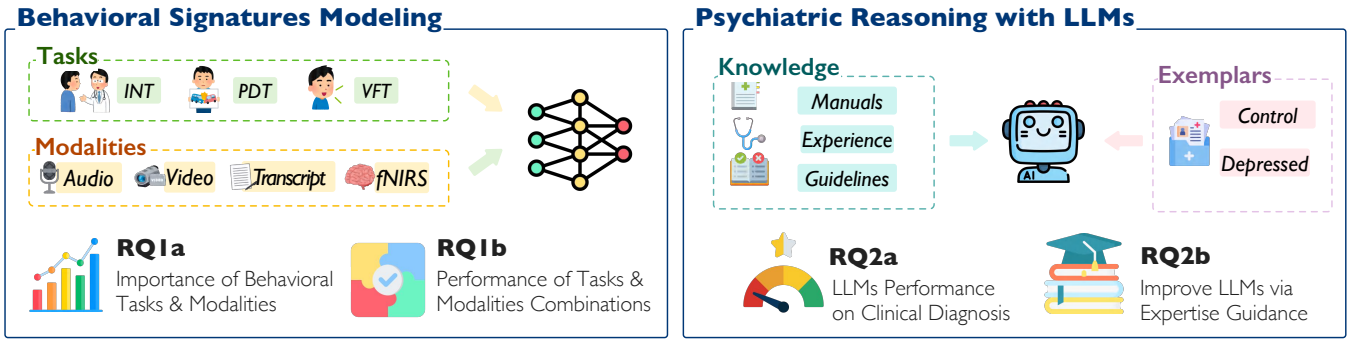

Figure 3: Overview of the two-part research methodology, addressing: 1) the diagnostic value of behavioral signatures (RQ1) and 2) the performance and enhancement of psychiatric reasoning in LLMs (RQ2).

richer modalities (Audio, Video, Text, fNIRS), expert-verified clinical diagnoses, and comprehensive psychometric data—making it a uniquely comprehensive and clinically grounded benchmark for depression research.

3 Methodology

As shown in Figure 3, we design a two-part methodological framework to uncover the mechanisms of clinical depression assessment: 1) modeling behavioral signatures across tasks and modalities, and 2) simulating psychiatric reasoning through guided LLMs. Due to space limitations, we present only core formulations here; details of models and prompts can be found in Technical Appendix.

3.1 Behavioral Signature Modeling

We define behavioral signatures as the informational cues derived from different psychiatric tasks and data modalities. To quantify the diagnostic relevance of different behavioral cues, we follow a structured modeling pipeline. First, we extract feature representations from four modalities (audio, video, text, fNIRS) for each psychiatric task. Then, we train learning models to predict depression status based on feature sets. By evaluating each task-modality combination and their fusions, we aim to uncover which behavioral signatures most robustly reflect clinical diagnoses. The entire process is designed to simulate and analyze how observable behaviors align with psychiatric assessment.

Feature Representations We extract two feature sets from four synchronized modalities: audio, video, transcript, and fNIRS. **1) Classical Feature Set.** We apply standard feature extraction pipelines. Audio signals are encoded using OpenSmile's eGeMAPS (88 dimensions), while video-based facial behavior is represented using OpenFace (4,963 dimensions), including action units, gaze, and head pose. Textual transcripts are embedded using DeBERTa. For fNIRS, statistical features are computed from 45 optical channels, resulting in a 630-dimensional vector. **2) Foundation Model Feature Set.** We also extract semantic-level embeddings using pretrained foundation models: Qwen2-Audio-7B (Chu et al. 2023) for audio, Qwen2.5-VL-72B for video (Bai et al. 2025), and Qwen3-235B-A22B for transcript (Yang et al. 2025). Each modality is segmented by task, encoded through the model's final hidden layer, and globally max-pooled over time to yield fixed-length vectors (4096 dimensions for audio and transcript, 8192 for video). fNIRS remains represented by classical statistics due to the absence of public pretrained encoders.

Task & Modality Modeling We denote the task set as $\mathcal{T} = \{\text{INT}, \text{PDT}, \text{VFT}\}$ and the modality set as $\mathcal{M} = \{\text{Audio}, \text{Video}, \text{Transcript}, \text{fNIRS}\}$. For each subject i, let $X^{(i)}_{m,t}$ represent the features from modality m and task t. We train a classifier to map them to the clinical label:

$$f_\theta : X^{(i)}_{m,t} \to y^{(i)}, \quad y^{(i)} \in \{0, 1\}$$

This allows us to quantify the diagnostic value of each task-modality pair. To further explore whether aggregating behavioral evidence enhances performance, we conduct two types of fusion experiments. In *task fusion*, we fix the modality and concatenate features across all tasks: $X^{(i)}_{\text{fused}} = \text{Concat}(X^{(i)}_{m,t_1}, \ldots, X^{(i)}_{m,t_k})$ for every $m \in \mathcal{M}$. This setup allows us to assess whether different task designs provide complementary cognitive and affective signals. In *modality fusion*, we fix the task set and concatenate features across all modalities: $X^{(i)}_{\text{fused}} = \text{Concat}(X^{(i)}_{m_1,t}, \ldots, X^{(i)}_{m_k,t})$ for every $t \in \mathcal{T}$, followed by aggregation across tasks. This setting evaluates how multimodal observations enhance detection when applied consistently across structured clinical tasks. In both settings, the fused representation is passed to the same predictive function f_θ for classification.

3.2 Psychiatric Reasoning with LLMs

We examine whether LLMs can emulate clinician-like diagnostic reasoning based solely on transcripts (for text LLMs) or multimodal signals (for MLLM) from three structured psychiatric tasks: interview, picture description, and verbal fluency, spanning a total of ten tests (T1–T10). Each subject's signals are concatenated into a single prompt, and the LLM is tasked with predicting a binary diagnostic label.

We compare three reasoning strategies. **1) Direct Prediction** asks the model to infer the diagnosis directly from the input without any explanation. **2) Vanilla Reasoning** encourages the model to engage in free-form, step-by-step

Modality	Feature Set	Interview (INT)						Picture Description (PDT)						Verbal Fluency (VFT)					
		LSTM	CNN	MLP	k-NN	RF	SVM	LSTM	CNN	MLP	k-NN	RF	SVM	LSTM	CNN	MLP	k-NN	RF	SVM
Audio (A)	OpenSmile	72.15	84.25	82.31	85.18	91.17	91.11	69.49	88.24	81.21	94.10	88.24	85.18	70.90	84.28	79.39	76.14	88.24	82.29
	Qwen-Audio	72.57	78.41	58.39	55.54	69.26	76.39	74.07	71.54	67.58	69.64	72.47	79.39	76.93	82.33	61.63	78.96	76.43	76.39
Video (V)	OpenFace	70.41	71.47	69.00	72.94	72.40	73.32	72.25	75.46	65.42	69.64	74.49	64.58	74.64	74.49	69.54	69.64	75.43	64.71
	Qwen-VL	79.97	83.31	76.92	85.28	84.28	85.28	79.91	86.27	80.32	88.24	83.29	85.28	78.73	84.25	74.26	67.39	81.28	85.28
Transcript (T)	DeBERTa	60.93	64.33	51.92	46.32	58.02	52.28	66.36	62.16	48.54	55.54	52.87	51.43	57.83	56.54	56.37	62.64	56.44	55.84
	Qwen	60.57	64.69	58.16	60.92	61.37	61.73	67.69	68.38	56.81	58.88	67.58	67.39	55.07	59.29	53.83	76.14	62.13	64.21
fNIRS (N)	Statistics	62.54	71.27	59.61	43.33	73.49	61.46	65.98	69.06	54.83	42.05	73.30	45.36	62.55	64.05	65.57	50.18	75.43	46.88

Table 3: Performance (Macro-F1) of different models and feature sets, evaluated per task-modality combination. The color of each block corresponds to the average performance of the results within that block (a darker shade indicates a higher average).

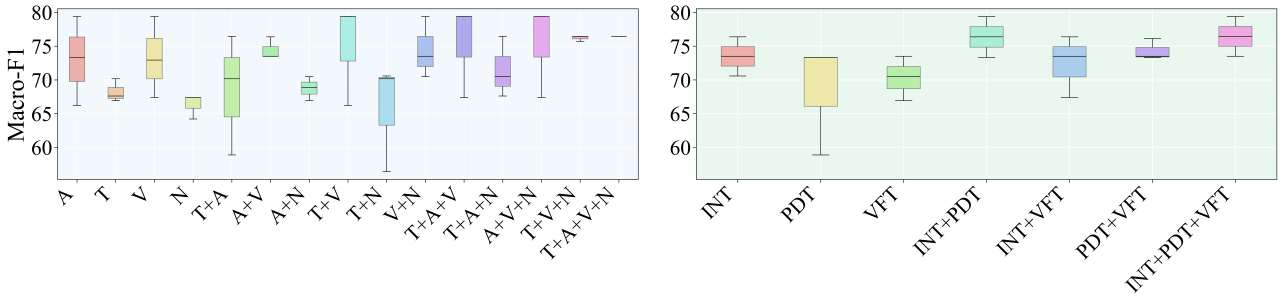

Figure 4: Performance (Macro-F1) of modality fusion (left) and task fusion (right). Combining signals generally improves the Macro-F1 score and reduces performance variance.

reasoning before making a prediction. **3) Psychiatric Reasoning** is our proposed method, which guides the model through a structured reasoning process grounded in clinical expertise. This approach reflects our key insight: effective psychiatric diagnosis depends not only on what is said, but also on how it is expressed under different task demands. To operationalize this, the prompt incorporates task definitions and expert-informed behavioral expectations, helping the LLM attend to symptom-relevant cues in a way that aligns with real clinical reasoning. The essential structure is summarized in the Technical Appendix. Let $\mathcal{P}^{(i)} = \{T_1^{(i)}, \ldots, T_{10}^{(i)}\}$ represent the transcript prompt for subject i. The LLM performs a binary classification $g_\phi(\mathcal{P}^{(i)}) \rightarrow y^{(i)} \in \{0, 1\}$. When clinical guidance $\mathcal{K}$ is included, the model performs $g_\phi(\mathcal{P}^{(i)}, \mathcal{K}) \rightarrow y^{(i)}$, using the embedded knowledge to attend to diagnostically meaningful patterns.

4 Experiments & Analysis

4.1 Experimental Settings

We conduct all experiments on C-MIND. To ensure robust evaluation, we randomly split the dataset into training, validation, and test sets following a 6:2:2 ratio. We report Macro-F1 as the main evaluation metric. Full metrics, including Precision, Recall, and per-class F1 scores, are available in the Technical Appendix. All results are averaged over five independent runs with different random seeds. Our analysis aims to answer two key research questions (RQs):

- **RQ1:** What are the contributions of different behavioral tasks and modalities to depression assessment?
- **RQ2:** Can LLMs reason like clinical psychiatrists, and how can knowledge injection improve their performance?

To address RQ1, we benchmark a suite of classical learning backbones, including LSTM, CNN, MLP, k-NN, Random Forest (RF), and SVM. To address RQ2, we evaluate several leading LLMs. The text-based LLMs, including `GPT-4o`, `GPT-o3`, `DeepSeek-V3`, `DeepSeek-R1`, and `Qwen3-235B-A22B-T/NT` (thinking/non-thinking mode), use only the transcript as input. In contrast, the multimodal model `Qwen2.5-Omni` processes a combination of audio, video, and transcript. Due to space limitations, detailed model architectures, parameters, and versions are provided in the Technical Appendix.

4.2 RQ1: The Power of Behavioral Signatures

Tasks and Modalities As shown in Table 3, we evaluate each task and modality using two feature sets. Our analysis reveals that Audio and Video are the most informative modalities, though their effectiveness is deeply intertwined with the psychiatric task being performed. Each task is designed to probe different cognitive and emotional facets, and their diagnostic power comes from how well these probes elicit observable, depression-related behavioral markers.

The Picture Description Task (PDT), for instance, excels in this regard, proving to be the most effective probe in our analysis. This is clinically intuitive as it assesses for emotional and attentional biases. Depressed individuals may exhibit a negative interpretation bias or provide less detailed descriptions, which is reflected not only in word choice (Transcript) but crucially in a flat vocal tone (Au-

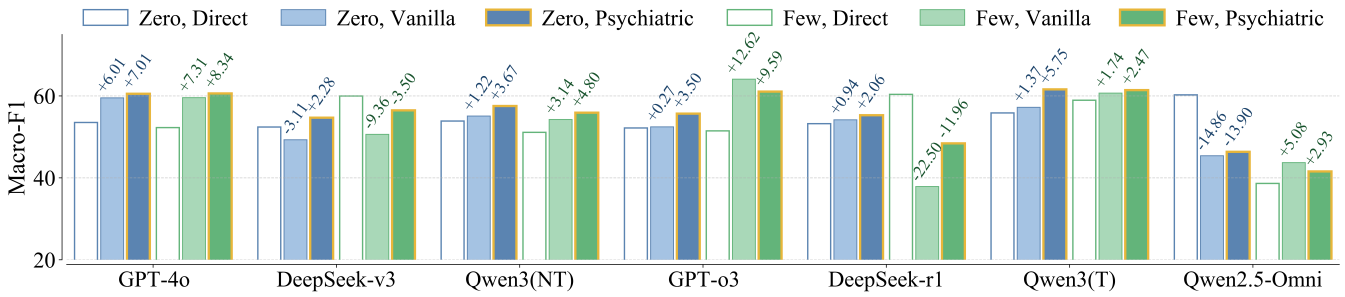

Figure 5: Performance comparison of LLMs using three reasoning strategies in zero/few-shot settings. Numbers above the bars indicate the performance change relative to the corresponding "Direct" baseline.

dio) and blunted affect (Video). This is evidenced by the top-performing model, which achieved a 94.10% Macro-F1 score using audio features from the PDT. Similarly, the Verbal Fluency Task (VFT) also shows remarkable performance, particularly with audio features. VFT assesses executive functions and semantic memory, which are often impaired in depression. This cognitive deficit doesn't just manifest as a lower word count, but more saliently as acoustic patterns like longer pauses, frequent hesitation markers (e.g., "uh", "um"), and reduced prosodic variation. These are precisely the signals captured by audio analysis, explaining its success. The Interview task (INT) remains a robust baseline because its autobiographical prompts are effective at eliciting narratives laden with depressive markers like negative sentiment and overgeneralization, signals that are present across Audio, Video, and Transcript.

Fusion Improves Robustness As illustrated in Figure 4, a clear and consistent finding is that fusing evidence from multiple sources enhances diagnostic performance. Combining modalities (e.g., Audio and Video) or integrating tasks (e.g., INT and PDT) consistently leads to higher Macro-F1 scores and, critically, more stable and reliable predictions by reducing variance. This underscores the value of a holistic assessment strategy, where a richer, multi-faceted view of a participant's behavior provides a more robust foundation for clinical inference than any single signal alone.

4.3 RQ2: Psychiatric Reasoning with LLMs

Reasoning Strategies We evaluate seven leading LLMs under three prompting strategies: *Direct Prediction*, *Vanilla Reasoning*, and our proposed *Psychiatric Reasoning*, across both zero-shot and few-shot conditions (Figure 5). Several consistent patterns emerge.

1) Psychiatric Reasoning consistently improves zero-shot performance. Across most non-thinking models (e.g., GPT-4o, GPT-o3, Qwen3(NT)), the structured psychiatric prompt yields stable gains (e.g., +7.01% for GPT-4o, +3.67% for Qwen3(NT)), outperforming both Direct and Vanilla strategies. 2) Few-shot performance gains vary, and can conflict with structured guidance. For GPT-4o, Vanilla Reasoning under few-shot improves significantly (+7.31%), but the gain from Psychiatric Reasoning (+8.34%) suggests that explicit guidance remains beneficial. In contrast, DeepSeek-v3 and DeepSeek-r1 show degradation under few-shot reasoning, likely due to incompatibility between pretrained reasoning paths and injected prompts. 3) Models with internal reasoning protocols may conflict with external prompts. DeepSeek-r1 exhibits a significant drop when reasoning is added, especially under few-shot settings (-22.5% with Vanilla and -11.96% with Psychiatric prompts), highlighting potential interference from overlaying external logic on built-in reasoning. 4) Multimodal input does not guarantee improved performance. The multimodal model Qwen2.5-Omni consistently underperforms across all prompting settings, achieving just 46.36% with Psychiatric Reasoning (zero-shot), which is worse than most transcript-only models. This suggests that general-purpose multimodal LLMs currently lack the fine-grained capability to utilize clinical non-verbal cues effectively without task-specific tuning.

Notably, even the best transcript-based LLM with Psychiatric Reasoning (GPT-4o, 60.53%) still falls short of the 68.24% achieved by a supervised model trained directly on transcript Qwen features, further emphasizing the performance gap between prompting-based and discriminative approaches in high-stakes diagnosis.

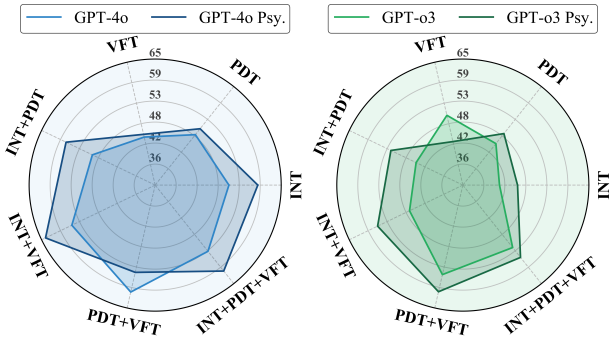

Figure 6: Performance gains of task combinations on LLMs.

Task Fusion On LLMs To examine whether LLMs benefit from task-level information integration, we aggregate transcripts from multiple psychiatric tasks (Figure 6). Results align with RQ1: combining tasks improves performance. For example, GPT-4o's Macro-F1 improves from

50.48% (INT only) to 55.68% (INT+VFT), and further to 60.53% when Psychiatric Reasoning is applied. The best-performing configuration under transcript-only input is GPT-4o with INT+VFT and Psychiatric prompt. Similar trends are observed with GPT-o3. However, tasks like VFT remain underutilized in isolation, likely due to the loss of timing and repetition patterns during transcription. This supports our earlier findings that linguistic transcripts alone cannot fully capture the cognitive and affective richness of certain psychiatric tasks.

4.4 Case Study

The case study presented in Figure 7 serves as a clear illustration that explicitly integrating domain-specific psychiatric knowledge enhance the performance of depression assessment. In our observation, domain knowledge contributes in two crucial ways. First, it guides clinical interpretation of signals. In the case of PDT, rather than overreacting to raw metrics *"low word count"*, reasoning with psychiatric knowledge assesses recognize protective factors like emotional expressiveness, therefore correctly identifying non-pathological cases. Second, knowledge prevents over-weighting isolated negative signals. For example, when encountering *"Good people don't get good rewards,"* the baseline model treats it as a core depressive marker. Psychiatric reasoning, however, draws on clinical reasoning to distinguish between fleeting complaints and the pervasive negativity typical of depression. By noting the lack of elaboration or supporting cues, it correctly down-weights the phrase.

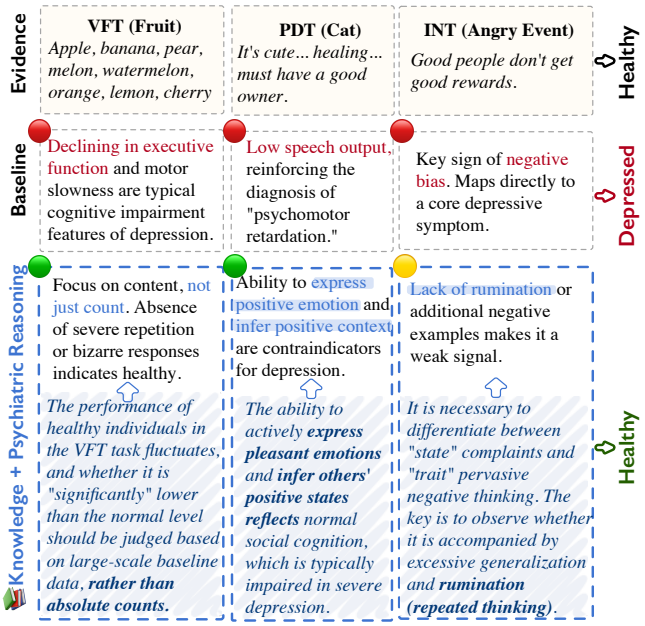

Figure 7: A case contrasting a superficial, baseline interpretation with a nuanced, knowledge-guided assessment for the same healthy participant. Red marks the incorrect assumptions of the baseline, green highlights the correct interpretations guided by psychiatric expertise, and yellow identifies points that are considered but ultimately de-emphasized.

5 Related Work

Corpus The foundation of depression detection is its data corpora, which have evolved along a hierarchy of evidence, trading scale for clinical validity. Early research leveraged large-scale social media corpora with labels based on user self-disclosure (Shen et al. 2017; Tadesse et al. 2019; Zirikly et al. 2019; Bucur et al. 2025). To improve signal quality, subsequent work introduced datasets collected in controlled settings, where ground truth was typically derived from self-report questionnaire scores (Gratch et al. 2014; Valstar et al. 2016; Guo et al. 2021; Tasnim et al. 2022). Representing a move toward the clinical gold standard, the most recent corpora have begun to incorporate formal diagnoses from trained psychiatrists (Cai et al. 2022; Zou et al. 2022; Lin et al. 2022). These pioneering efforts often feature smaller or imbalanced cohorts and are focused on a limited set of behavioral tasks or modalities. Our work therefore introduces a new clinically-validated resource featuring a balanced cohort across diverse tasks and modalities.

Method Paralleling the evolution of datasets, detection methods have shifted from analyzing handcrafted features within single modalities to learning complex data representations through sophisticated, multimodal architectures. Initial approaches relied on handcrafted features from single modalities, such as text, audio, or video (Fossati et al. 2003a; Cummins et al. 2015; Ma et al. 2016). A consensus has since formed around multimodal fusion models, which integrate these channels using sophisticated attention or transformer architectures to achieve stronger performance (Fan et al. 2019; Wei et al. 2023; Chen et al. 2024; Jia et al. 2025; Wu et al. 2025). The latest frontier involves applying LLMs to this task. While powerful, research highlights challenges in adapting these general-purpose models for clinical use, noting the need to imbue them with specialized, domain-specific knowledge beyond what is learned from web-scale text (Guo et al. 2024; Wang, Inkpen, and Kirinde Gamaarachchige 2024; Hua et al. 2025; Bi et al. 2025). Our work contributes to this frontier by conducting a comprehensive analysis across different tasks and modalities to clarify their discriminative power, and then proposing a novel psychiatric reasoning mechanism to enhance the clinical awareness of LLMs.

6 Conclusion

We present the clinical multimodal neuropsychiatric diagnosis (C-MIND) dataset, a clinically validated resource collected from real hospital settings, featuring diverse behavioral signals across structured tasks and synchronized modalities. Through systematic analysis, we reveal how specific combinations of tasks and modalities enhance diagnostic stability, providing empirical guidance for system design. We also show that large language models, when guided by structured psychiatric knowledge, can better approximate expert reasoning in complex diagnostic scenarios. By integrating high-quality clinical data with interpretable and knowledge-informed modeling, this work offers a concrete step toward computational systems that are accurate, trustworthy, and deployable in real-world mental healthcare.

Acknowledgments

This work was supported by Beijing Natural Science Foundation (L252009), the National Key Research and Development Program of China 2024YFC3606800, and the NSFC projects 62441614. This work was also supported by the National Science Foundation for Distinguished Young Scholars (No. 62125604) and the Beijing Municipal Natural Science Foundation (No. 7244509).

References

Akiyama, T.; Koeda, M.; Okubo, Y.; and Kimura, M. 2018. Hypofunction of left dorsolateral prefrontal cortex in depression during verbal fluency task: A multi-channel near-infrared spectroscopy study. *Journal of Affective Disorders*, 231: 83–90.

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders*. Arlington, VA: American Psychiatric Publishing, 5th edition.

Angst, J.; Adolfsson, R.; Benazzi, F.; Gamma, A.; Hantouche, E.; Meyer, T. D.; Skeppar, P.; Struening, E.; Vieta, E.; and Scott, J. 2005. The HCL-32: Towards a self-assessment tool for hypomanic symptoms in outpatients. *Journal of Affective Disorders*, 88(2): 217–233.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bi, G.; Chen, Z.; Liu, Z.; Wang, H.; Xiao, X.; Xie, Y.; Zhang, W.; Huang, Y.; Chen, Y.; Peng, L.; and Huang, M. 2025. MAGI: Multi-Agent Guided Interview for Psychiatric Assessment. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 24898–24921. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Bucur, A.-M.; Moldovan, A.-C.; Parvatikar, K.; Zampieri, M.; KhudaBukhsh, A. R.; and Dinu, L. P. 2025. Datasets for Depression Modeling in Social Media: An Overview. *arXiv preprint arXiv:2503.21513*.

Buysse, D. J.; Reynolds III, C. F.; Monk, T. H.; Berman, S. R.; and Kupfer, D. J. 1989. The Pittsburgh Sleep Quality Index: A new instrument for psychiatric practice and research. *Psychiatry Research*, 28(2): 193–213.

Cai, H.; Yuan, Z.; Gao, Y.; Sun, S.; Li, N.; Tian, F.; Xiao, H.; Li, J.; Yang, Z.; Li, X.; et al. 2022. A multi-modal open dataset for mental-disorder analysis. *Scientific Data*, 9(1): 178.

Cattell, R. B.; Eber, H. W.; and Tatsuoka, M. M. 1970. *Handbook for the Sixteen Personality Factor Questionnaire (16PF)*. Champaign, IL: Institute for Personality and Ability Testing.

Chen, Z.; Deng, J.; Zhou, J.; Wu, J.; Qian, T.; and Huang, M. 2024. Depression Detection in Clinical Interviews with LLM-Empowered Structural Element Graph. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 8181–8194. Mexico City, Mexico: Association for Computational Linguistics.

Chu, Y.; Xu, J.; Zhou, X.; Yang, Q.; Zhang, S.; Yan, Z.; Zhou, C.; and Zhou, J. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Cui, X.; Bray, S.; Bryant, D. M.; Glover, G. H.; and Reiss, A. L. 2011. A quantitative comparison of NIRS and fMRI across multiple cognitive tasks. *NeuroImage*, 54(4): 2808–2821.

Cummins, N.; Scherer, S.; Krajewski, J.; Schnieder, S.; Epps, J.; and Quatieri, T. F. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71: 10–49.

Derogatis, L. R. 1977. *SCL-90: Administration, scoring and procedures manual-I for the R(evised) version and other instruments of the Psychopathology Rating Scale Series*. Baltimore: Johns Hopkins University School of Medicine.

Fan, W.; He, Z.; Xing, X.; Cai, B.; and Lu, W. 2019. Multi-modality depression detection via multi-scale temporal dilated cnns. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 73–80.

Fossati, P.; Ergis, A.-M.; Allilaire, J.-F.; et al. 2003a. Qualitative analysis of verbal fluency in depression. *Psychiatry research*, 117(1): 17–24.

Fossati, P.; Le Bastard, G.; Small, S.; Rigaud, A.-S.; Kahn, J.-P.; Pilliod, S.; Jaafari, N.; Allilaire, J.-F.; and Dubois, B. 2003b. Verbal fluency and clustering in patients with mood disorders. *Psychiatry Research*, 117(3): 187–207.

Gratch, J.; Artstein, R.; Lucas, G. M.; Stratou, G.; Scherer, S.; Nazarian, A.; Wood, R.; Boberg, J.; DeVault, D.; Marsella, S.; et al. 2014. The distress analysis interview corpus of human and computer interviews. In *LREC*, volume 14, 3123–3128. Reykjavik.

Guo, W.; Yang, H.; Liu, Z.; Xu, Y.; and Hu, B. 2021. Deep neural networks for depression recognition based on 2D and 3D facial expressions under emotional stimulus tasks. *Frontiers in neuroscience*, 15: 609760.

Guo, Z.; Lai, A.; Thygesen, J. H.; Farrington, J.; Keen, T.; Li, K.; et al. 2024. Large language models for mental health applications: systematic review. *JMIR mental health*, 11(1): e57400.

Hamilton, M. 1959. The assessment of anxiety states by rating. *British Journal of Medical Psychology*, 32(1): 50–55.

Hamilton, M. 1960. A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry*, 23(1): 56–62.

Hua, Y.; Liu, F.; Yang, K.; Li, Z.; Na, H.; Sheu, Y.-h.; Zhou, P.; Moran, L. V.; Ananiadou, S.; Clifton, D. A.; et al. 2025. Large language models in mental health care: a scoping review. *Current Treatment Options in Psychiatry*, 12(1): 1–18.

Jia, X.; Chen, J.; Liu, K.; Wang, Q.; and He, J. 2025. Multimodal depression detection based on an attention graph convolution and transformer. *Mathematical biosciences and engineering: MBE*, 22(3): 652–676.

Lin, Y.; Liyanage, B. N.; Sun, Y.; Lu, T.; Zhu, Z.; Liao, Y.; Wang, Q.; Shi, C.; and Yue, W. 2022. A deep learning-based model for detecting depression in senior population. *Frontiers in psychiatry*, 13: 1016676.

Liu, Y.; Lu, X.; Shi, D.; Yuan, J.; Pan, T.; and An, H. 2021. Improved depression recognition using attention and multi-task learning of gender recognition. In *2021 International Conference on Asian Language Processing (IALP)*, 57–61. IEEE.

Ma, X.; Yang, H.; Chen, Q.; Huang, D.; and Wang, Y. 2016. DepAudionet: an efficient deep model for audio based depression classification. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 35–42. ACM.

Ramponi, C.; Collinson, S. L.; Worters, R.; and Breen, N. 2010a. Picture perception in depression: A systematic review and meta-analysis of neuroimaging studies. *Journal of Psychiatric Research*, 44(15): 1002–1014.

Ramponi, C.; Murphy, F. C.; Calder, A. J.; and Barnard, P. J. 2010b. Recognition memory for pictorial material in subclinical depression. *Acta Psychologica*, 135(3): 293–301.

Rinaldi, A.; Tree, J. E. F.; and Chaturvedi, S. 2020. Predicting depression in screening interviews from latent categorization of interview prompts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7–18. Association for Computational Linguistics.

Sarsam, S. M.; Al-Samarraie, H.; Alzahrani, A. I.; and Wright, B. 2024. Multimodal machine learning in mental health: A survey of data, algorithms, and challenges. *Information Fusion*, 106: 102517.

Shen, G.; Jia, J.; Nie, L.; Feng, F.; Zhang, C.; Hu, T.; Chua, T.-S.; Zhu, W.; et al. 2017. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI*, volume 2017, 3838–3844.

Tadesse, M. M.; Lin, H.; Xu, B.; and Yang, L. 2019. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7: 44883–44893.

Tasnim, M.; Ehghaghi, M.; Diep, B.; and Novikova, J. 2022. DEPAC: a Corpus for Depression and Anxiety Detection from Speech. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, 1–16.

Valstar, M.; Gratch, J.; Schuller, B.; Ringeval, F.; Lalanne, D.; Torres Torres, M.; Scherer, S.; Stratou, G.; Cowie, R.; and Pantic, M. 2016. AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 3–10. ACM.

Wang, Y.; Inkpen, D.; and Kirinde Gamaarachchige, P. 2024. Explainable Depression Detection Using Large Language Models on Social Media Data. In Yates, A.; Desmet, B.; Prud'hommeaux, E.; Zirikly, A.; Bedrick, S.; MacAvaney, S.; Bar, K.; Ireland, M.; and Ophir, Y., eds., *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, 108–126. St. Julians, Malta: Association for Computational Linguistics.

Wei, Y.; Zhang, Y.; Zhang, S.; and Zhang, H. 2023. Canamrf: An attention-based model for multimodal depression detection. In *Pacific Rim International Conference on Artificial Intelligence*, 111–116. Springer.

Wu, Z.; Zhou, L.; Li, S.; Fu, C.; Lu, J.; Han, J.; Zhang, Y.; Zhao, Z.; and Song, S. 2025. DepMGNN: Matrixial Graph Neural Network for Video-based Automatic Depression Assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1610–1619.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yoon, J.; Kang, C.-y.; Kim, S.; and Han, J. 2022. D-vlog: Multimodal vlog dataset for depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 12226–12234.

Zirikly, A.; Resnik, P.; Uzuner, O.; and Hollingshead, K. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, 24–33. ACL.

Zou, B.; Han, J.; Wang, Y.; Liu, R.; Zhao, S.; Feng, L.; Lyu, X.; and Ma, H. 2022. Semi-structural interview-based Chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders. *IEEE Transactions on Affective Computing*, 14(4): 2823–2838.

Zung, W. W. K. 1965. A self-rating depression scale. *Archives of General Psychiatry*, 12(1): 63–70.

Zung, W. W. K. 1971. A rating instrument for anxiety disorders. *Psychosomatics*, 12(6): 371–379.