Evidence-Bounded Mental Health Reasoning from Heterogeneous Speech Protocols

Chengyuan Gao [1,2,3] **Jiang Wu** [4] **Tao Lu** [3] **Jiayan Guo** [5]
Mingkun Xu[4] **Tianyi Zang**[2*] **Shangyang Li**[1,3*]
[1]Beijing University of Posts and Telecommunications [2]Harbin Institute of Technology
[3]Renyixun Health Technology Co., Ltd. [4]GDIIST [5]Tencent
{1967464882guge, wuj2386, taolu5010}@gmail.com guojy001@outlook.com
xumingkun@gdiist.cn tianyi.zang@hit.edu.cn syli@bupt.edu.cn

Abstract

Computational mental health screening using multimodal speech and text has shown great promise. However, existing models often assume all clinical speech protocols carry equivalent evidentiary validity. In reality, heterogeneous protocols, from free interviews to fixed reading tasks, support fundamentally different evidence. Forcing uniform reasoning flattens these boundaries, causing models to hallucinate symptoms from irrelevant text or overclaim support. Even advanced long chain-of-thought LLMs fail to resolve this issue, as free-form reasoning can exacerbate boundary violations. To address this, we reformulate multimodal screening as an evidence-bounded reasoning problem. We introduce the **Evidence Package Benchmark**, integrating 1,870 packages across six heterogeneous sources with explicit modality masks and evidence permissions. We further propose **EviBound**, a protocol-aware evidence control framework. Unlike direct LLM prompting, EviBound uses a profile-aware planner to restrict reasoning scope, orchestrates evidence tools via five-way acoustic consensus, and enforces a boundary critic to suppress unsupported claims. Empirical results show EviBound achieves a held-out test Depression AUROC of 0.8658, exceeding the strongest direct omni-modal baseline by +0.0811 AUROC while maintaining zero claim violations. Our work moves beyond unconstrained accuracy toward evidence-consistent, protocol-aware systems for safer clinical NLP research. Code will be released upon publication.

1 Introduction

Large language models (LLMs) offer unprecedented opportunities for scalable mental health screening. However, contemporary omni-modal systems often implicitly assume that *all speech acquisition protocols provide equivalent evidentiary validity*. In practice, diagnostic evidence is inherently constrained by its collection context. As illustrated in Figure 1, mental health datasets follow four protocol types: free interviews (supporting acoustic, semantic, and discourse claims), prompted speech (acoustic evidence permitted but semantic inference restricted), fixed reading (prosody only, text irrelevant), and text-only discourse (no prosody, semantic claims from language alone)(Gratch et al., 2014; Ringeval et al., 2019; Zou et al., 2023). Each protocol thus establishes distinct boundaries for acoustic, semantic, discourse, and missing-evidence claims (Qin et al., 2025; Shen et al., 2022; Cai et al., 2022; Cho et al., 2026; Delgaram-Nejad et al., 2023).

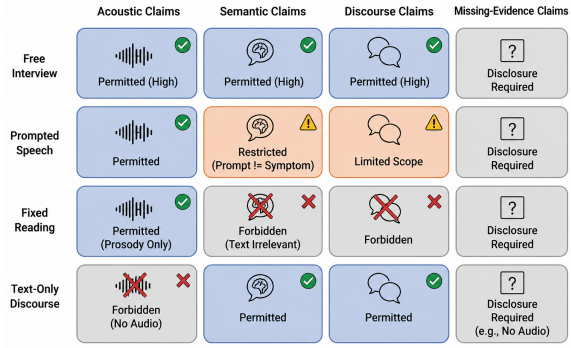

Figure 1: Protocol-specific evidence boundaries across heterogeneous speech records. Free interviews, prompted speech, fixed reading, and text-only discourse license different levels of acoustic, semantic, discourse, and missing-evidence claims.

Ignoring these boundaries leads to what we term *epistemic flattening*: the collapse of heterogeneous evidence constraints into a single unconstrained reasoning space (Qin et al., 2025; WANG et al., 2026). Under restrictive protocols, systems hallucinate symptoms by conflating semantic and acoustic cues, generate unsupported interpretations, and attribute conclusions to absent modalities. Importantly, neither larger models nor longer chain-of-thought reasoning resolves this issue; they can

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 14394–14406
October 24-29, 2026 ©2026 Association for Computational Linguistics

instead increase reasoning instability and computational cost while exacerbating boundary violations (Liu et al., 2026, 2025; Dong et al., 2025; Jiang et al., 2025). These findings indicate that the central challenge is not insufficient reasoning capability, but the absence of explicit evidence-bound control.

To address this limitation, we reformulate computational mental health screening as an evidence-bounded reasoning problem. Formally, we define each screening instance as an *evidence package* $\mathcal{E} = (P, M, \Pi, \mathcal{X})$, where P is the protocol profile, M the modality mask, $\mathcal{X}$ the multimodal observations, and Π the admissible evidence boundary for that protocol. Under this formulation, the objective is not only to maximize predictive accuracy, but also to ensure conclusions remain strictly consistent with observable evidence and protocol constraints. This view turns dataset documentation, model reporting, and medical/psychiatric benchmark principles into executable constraints rather than prose-only caveats (Gebru et al., 2021; Mitchell et al., 2019; Arora et al., 2025). Importantly, this task is framed as benchmark-level boundary control for research evaluation rather than formal clinical diagnosis or deployment readiness (Gebru et al., 2021; Mitchell et al., 2019; Arora et al., 2025).

To operationalize this paradigm, we introduce **EviBound**, a protocol-aware evidence control framework for bounded multimodal reasoning. Instead of unrestricted direct generation, EviBound uses a three-stage architecture: (1) a profile-aware planner that defines permissible reasoning scope based on protocol and modality availability; (2) parallel evidence modules that cross-validate signals via five-way acoustic consensus; and (3) a boundary critic that suppresses unsupported claims and enforces consistency with the modality mask.

Our primary contributions are:

- **Evidence-Bounded Reasoning.** We formalize protocol-aware mental health reasoning with explicit evidence constraints under heterogeneous speech settings.
- **Evidence Package Benchmark.** We build a benchmark of 1,870 standardized evidence packages from six clinical sources with modality masks and claim permissions.
- **EviBound Framework.** We propose a protocol-aware multimodal reasoning harness with explicit evidence routing and boundary validation.
- **Boundary-Aware Evaluation.** We introduce CVR, MHR, and EBCPass to measure evidence consistency under heterogeneous protocols.
- **Experimental Results.** EviBound achieves 0.8658 AUROC and 100% EBCPass on held-out packages with zero boundary violations.

2 Related Work

Multimodal Mental Health Screening and Clinical Benchmarks. Speech-based mental health research spans interview modeling, multimodal fusion, audiovisual risk assessment, and recent large multimodal reasoning systems for psychiatric analysis (Al Hanai et al., 2018; Fang et al., 2023; Dhelim et al., 2023; Zhang et al., 2026; Qin et al., 2025; WANG et al., 2026). Public resources such as DAIC-WOZ/E-DAIC, CMDC, EATD, MODMA, MMPsy, DISCOURSE-UWO, and DAIS-C differ substantially in language, modality availability, and acquisition protocol (Gratch et al., 2014; Ringeval et al., 2019; Zou et al., 2023; Shen et al., 2022; Cai et al., 2022; Cho et al., 2026; Delgaram-Nejad et al., 2023). Meanwhile, benchmarks including HealthBench, MedHELM, MentalBench, and PsychiatryBench emphasize that clinical AI evaluation should extend beyond a single predictive metric toward reliability and reasoning assessment (Arora et al., 2025; Jing et al., 2026; Song et al., 2026; Fouda et al., 2026). However, prior systems largely treat heterogeneous protocols as evidentially interchangeable, often producing what we term *epistemic flattening*. In contrast, our benchmark explicitly models modality masks and claim permissions as reasoning constraints.

Speech Evidence, Clinical Agents, and Trustworthy Medical AI. Recent work converts acoustic landmarks, SSL representations, digital phenotypes, and psychological knowledge into LLM-usable evidence for mental-health prediction (Zhang et al., 2024, 2025; Li et al., 2025; Ali et al., 2025; Pan et al., 2026). Standard speech pipelines rely on openSMILE/eGeMAPS and SSL encoders such as wav2vec 2.0, HuBERT, and WavLM (Eyben et al., 2010; Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022), while clinical language-agent systems increasingly employ structured planning and tool orchestration for biomedical reasoning (Tang et al., 2024; Yue et al., 2024; Wang et al., 2025; Liu et al., 2026; Mao et al., 2026; Zhi et al., 2026; Zheng et al., 2026). Concurrently, trustworthy clinical NLP stud-

ies investigate hallucination mitigation, verifier-guided decoding, retrieval-augmented generation, and factuality-oriented evaluation (Ni et al., 2025; Han et al., 2025). Unlike prior work that primarily expands reasoning capability or factual grounding, our focus is protocol-aware evidence validity through an Agent-Evidence Interface and boundary-aware evaluation metrics (CVR, MHR, and EBCPass).

3 Methodology

We formalize *evidence-bounded mental health reasoning* and present **EviBound**, a protocol-aware structured evidence harness for multimodal screening under heterogeneous acquisition protocols. Unlike conventional multimodal systems that implicitly treat all modalities as evidentially interchangeable, EviBound operates over explicit *evidence packages* encoding protocol profile, modality availability, missing evidence, and admissible claim scope. The framework separates three components: (1) an *evidence interface* defining protocol-conditioned routing constraints; (2) a *risk routing layer* that aggregates compatible acoustic, semantic, and feature evidence; and (3) a *boundary validator* that enforces evidence consistency during report generation.

3.1 Evidence-Bounded Reasoning

A central but largely unexamined assumption in multimodal mental health screening is that all speech protocols support equivalent forms of clinical inference. In practice, fixed-reading audio, prompted narratives, open-ended interviews, and text-only discourse expose fundamentally different evidence signals. Ignoring these distinctions produces what we term *epistemic flattening*: heterogeneous evidence constraints collapse into a single unconstrained reasoning space, allowing systems to infer symptoms from scripted text or generate acoustic claims without audio evidence.

We therefore formulate screening as an *evidence-bounded reasoning* problem. Each instance is represented as an **evidence package**

$$\mathcal{E} = (P, M, \Pi, \mathcal{X}),$$

where P denotes the protocol profile, M the modality mask, $\mathcal{X}$ the observed multimodal inputs, and Π the admissible evidence boundary defined by P and M. Protocol profiles distinguish acquisition settings such as interviews, prompted speech, fixed reading, and text-only discourse, while modality masks specify available evidence channels including raw audio (A), transcript text (T), and pre-extracted acoustic features (F).

Given an evidence package $\mathcal{E}$, the system predicts screening-oriented risk scores and generates a structured report r. A report is considered *evidence-bounded* only if every emitted claim remains consistent with the package profile and modality permissions:

$$\text{permit}(c, M, P) = 1, \quad \forall c \in \mathcal{C}(r),$$

where $\mathcal{C}(r)$ denotes the set of generated claims. The permission function is protocol-dependent: direct listening claims require raw audio, symptom-history claims require participant-centered interview evidence, and fixed-reading text cannot support semantic symptom inference. Missing evidence must also be explicitly disclosed.

The objective is therefore not only predictive performance, but also evidence-consistent reasoning under heterogeneous acquisition protocols.

3.2 Evidence Package Benchmark

We construct the **Evidence Package Benchmark**, a unified benchmark of 1,870 packages derived from six heterogeneous mental health resources: CMDC, DAIS-C, E-DAIC, EATD, MMPsy, and MODMA. Rather than concatenating raw datasets into a homogeneous cohort, each source is instantiated into the unified evidence package schema $\mathcal{E} = (P, M, \Pi, \mathcal{X})$, where observations are strictly partitioned into *model-visible inputs* and *evaluator-only labels* y_{eval} (e.g., PHQ-9 or MDD diagnosis). Evaluator-side labels remain inaccessible during inference, preventing leakage through scale memorization or protocol shortcuts.

Table 1 summarizes source composition, available modalities, languages, and frozen subject-independent splits. Packages may contain audio, transcript text, feature vectors, or text-only discourse. Feature-only records can support feature-derived risk estimation but cannot license direct listening claims. Depression screening, anxiety screening, discourse probes, and severity-oriented tasks remain distinct evaluation surfaces rather than being collapsed into a universal target.

The claim-permission matrix (Table 2) defines the evidential scope associated with each protocol–modality combination and serves as the executable specification for runtime report validation.

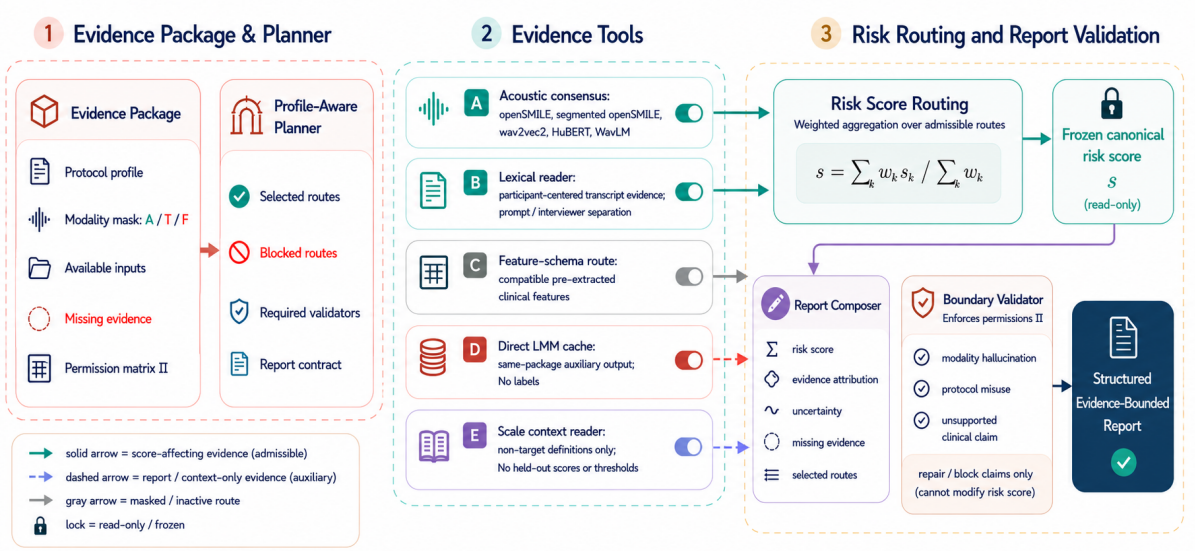

Figure 2: The EviBound framework architecture. (1) The Profile-Aware Planner parses the evidence package $\mathcal{E}$ to select admissible routes and enforce the permission matrix Π. (2) Evidence Tools extract modality-specific signals, including a five-way acoustic consensus and lexical/feature readers. (3) The system decouples Risk Score Routing from Report Validation: risk scores are aggregated from admissible routes and frozen, while a deterministic validator ensures the final structured report respects evidence boundaries without altering the read-only risk score.

Data	Lang.	N	Split	Profile	Inputs	Eval-only labels
CMDC	Zh	78	57/8/13	interview	A+T	MDD/HC; PHQ/HAMD
DAIS-C	En	28	18/4/6	discourse	T	group/disc. probe
E-DAIC	En	275	163/56/56	interview	A+T+F	PHQ-8 dep.; S
EATD	Zh	162	116/13/33	prompted	A	SDS dep.; S
MMPsy	Zh	1275	891/128/256	interview	T+F	PHQ/GAD; S
MODMA	Zh	52	39/5/8	reading	A	MDD/HC; PHQ/GAD

Table 1: Benchmark composition with fixed package splits. N counts packages, not raw segments. Inputs are model-visible A/T/F fields; labels and scale-derived targets (S) are evaluator-only.

Case	Voice claim	Text claim	Main restriction
Interview + A	direct	yes	split participant/prompt
Interview no A	F-derived	yes	no direct listening
Prompted + A	acoustic	limited	do not treat prompt text as symptoms
Reading + A	acoustic	no	no personal-history inference
Text/disc.	no	text/disc.	no acoustic claims

Table 2: Compact claim-permission matrix. Rows are profile/mask cases. A=raw audio, T=transcript, F=pre-extracted features.

3.3 Profile-Aware Planner

EviBound uses a deterministic planner to map each evidence package to a restricted set of compatible evidence routes. The planner functions as an *Agent–Evidence Interface*, determining which tools, claims, and report fields are admissible before reasoning begins.

Clinical interviews may activate lexical, acoustic, and feature routes; prompted or fixed-reading speech may use acoustic evidence but cannot interpret scripted text as symptom history; text-only records may support discourse analysis but cannot generate acoustic or prosodic claims. Missing modalities are explicitly registered before inference, ensuring that absent evidence becomes part of the report contract rather than an implicit assumption.

3.4 Evidence Tools and Risk Routing

The evidence harness integrates lightweight and independently auditable evidence modules. For audio-supported records, the primary route is a five-way acoustic consensus over openSMILE/eGeMAPS, segmented openSMILE, wav2vec2, HuBERT, and WavLM representations. Each branch outputs calibrated risk estimates and uncertainty signals, enabling conservative aggregation under disagreement.

For the applicable route set $R(x) = \{k : \rho_k(x) = 1\}$, where ρ_k is the route precondition determined by protocol and modality permissions, the final routed score is computed as

$$\hat{s}(x) = \frac{\sum_{k \in R(x)} w_k s_k(x)}{\sum_{k \in R(x)} w_k}, \qquad (1)$$

where weights $w_k \geq 0$ are historical calibration performance. Inapplicable routes are masked before aggregation, and disagreement signals propagate into report uncertainty fields.

Beyond acoustic routing, EviBound includes lexical readers for participant-centered transcripts,

feature-schema routes for compatible pre-extracted clinical features, and frozen omni-modal LLM outputs used as auxiliary evidence sources rather than authoritative decision makers. Source identifiers are used only for compatibility masking and never as direct predictors of clinical status.

3.5 Boundary Validator and Structured Reports

The report composer emits a structured schema containing risk score, uncertainty, evidence attribution, missing evidence, blocked claims, selected routes, and validator status. The final risk score is frozen before report repair, ensuring that post-hoc validation cannot alter AUROC or F1 outcomes.

The validator operates as the executable realization of the package permission matrix, transforming protocol constraints into runtime-verifiable report boundaries. It checks three major categories of violations:

1. **Modality hallucination**: claims referencing unavailable modalities are removed;
2. **Protocol misuse**: fixed-reading or prompted text cannot support symptom-history inference;
3. **Claim scope**: diagnosis, treatment, prognosis, or unsupported clinical recommendations are blocked.

Unsupported claims are removed or rewritten, blocked reasons are recorded, and schema validation is rerun before output finalization. The validator operates under a strict read-only constraint on risk scores:

$$\mathrm{score}(r') = \mathrm{score}(r), \tag{2}$$

where $r' = \mathrm{validate}(r; x)$ denotes the repaired report. Report validation is therefore evaluated independently from predictive classification performance.

3.6 Inference and Evaluation

EviBound operates as a compositional protocol-governed pipeline, as illustrated in Figure 2. During inference, the planner first selects compatible routes from the evidence package, followed by parallel evidence extraction, structured report generation, and deterministic boundary validation. This design decouples multimodal evidence aggregation from report-boundary enforcement, reducing unsupported reasoning drift during generation.

$$\mathcal{E} \xrightarrow{\text{Planner}} \tau \xrightarrow{\text{Tools}} \{\mathcal{C}_a, X_s\} \xrightarrow{\text{Decoder}} \mathcal{R} \tag{3}$$

We evaluate both predictive utility and evidence consistency. Predictive performance is measured using AUROC, Macro-F_1, and Quadratic Weighted Kappa (QWK) for severity-oriented evaluation. Evidence consistency is audited using Claim Violation Rate (CVR), Modality Hallucination Rate (MHR), and Evidence-Bound Consistency Pass (EBCPass), which measures whether generated claims remain fully contained within the admissible evidence boundary Π. Importantly, replay diagnostics and report-boundary auditing are evaluated independently from disease-label ranking metrics.

4 Experimental Setup

4.1 Baselines

We compare EviBound against four categories of baselines under the identical package-interface setting, ensuring that all systems operate only on model-visible evidence without access to evaluator-only labels or unavailable modalities.

Direct omni-modal LMMs. We evaluate direct prompting baselines using Qwen3-Omni-Flash, Qwen3.5-Omni-Plus, and Gemini 3.5 Flash (Qwen Team, 2025, 2026; Google AI for Developers, 2026). These systems receive the same package inputs, structured prompts, decoding configuration, and output schema as EviBound, but generate reports without explicit evidence-bound routing or boundary validation.

Long-reasoning baseline. To test whether extended chain-of-thought reasoning alone can recover protocol-aware behavior, we evaluate a long-reasoning baseline with expanded reasoning budgets and structured deliberation prompts. This setting follows recent multimodal reasoning analyses showing that longer reasoning traces do not necessarily improve grounded multimodal inference (Liu et al., 2025; Dong et al., 2025; Jiang et al., 2025).

Acoustic and feature baselines. We include conventional speech-analysis baselines covering openSMILE/eGeMAPS descriptors, segmented acoustic aggregation, SSL embeddings (Wav2Vec2.0, HuBERT, WavLM), and the proposed five-way acoustic consensus route (Eyben et al., 2010; Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022). We additionally evaluate feature-based logistic regression models over compatible pre-extracted clinical features, as well as LLM-facing speech-evidence pipelines that transform acoustic or psychological signals into language-readable evidence (Zhang et al., 2024, 2025; Li

et al., 2025; Ali et al., 2025; Pan et al., 2026).

Harness ablations. To isolate the contribution of each component, we evaluate controlled ablations that separately remove profile-aware planning, acoustic consensus, feature routing, and boundary validation. Predictive routing components are evaluated independently from report-repair modules to avoid conflating score improvements with post-processing effects.

All systems share the same package ordering, parser implementation, decoding limits, cached outputs, and evaluation scripts.

4.2 Metrics

We evaluate models along two complementary dimensions: *predictive performance* and *evidence consistency*. This separation allows reasoning validity and predictive utility to be evaluated independently.

Predictive Metrics. Predictive evaluation uses held-out evidence packages only. We report AUROC and validation-threshold F1 on the pooled depression-eligible split ($n = 366$, 94 positives) and the anxiety-eligible split ($n = 264$, 35 positives, dominated by MMPsy). Severity-oriented evaluation further reports Quadratic Weighted Kappa (QWK) on compatible severity subsets ($n = 320$). Secondary diagnostics include AUPRC, Brier score, and calibration-oriented operating points. Depression and anxiety evaluations are reported separately because protocol composition and class distributions differ substantially across tasks.

Evidence Consistency Metrics. Beyond predictive accuracy, we evaluate whether generated reports remain consistent with evidence permissions through protocol-aware replay diagnostics. We formalize three boundary-aware metrics:

- **Claim Violation Rate (CVR):** proportion of packages containing at least one protocol-violating claim (e.g., no-audio acoustic hallucination or fixed-reading symptom overreach);
- **Missing-Evidence Handling Rate (MHR):** proportion of packages correctly disclosing unavailable modalities;
- **EBCPass (Evidence-Bound Consistency Pass):** a binary indicator equal to 1 only when a report contains zero claim violations, correctly handles missing evidence, and satisfies schema validation.

We intentionally avoid collapsing predictive and boundary-aware metrics into a single aggregate score, since predictive accuracy and evidence consistency capture different properties of system behavior.

4.3 Implementation Details

All systems operate exclusively on package-visible inputs during held-out inference. Labels, scale values, and thresholding rules remain evaluator-side only, and audit scripts verify their absence from prompts, cached outputs, and intermediate traces (Gebru et al., 2021; Mitchell et al., 2019).

Permission rules (Π) are derived entirely from protocol metadata and modality masks rather than prediction targets. Adapter identifiers are used only for compatibility masking and evaluator bookkeeping, never as direct predictors.

Candidate evidence routes are registered before held-out evaluation with fixed modality requirements, admissible outputs, and calibration strategies. Validation splits determine consensus weights, operating thresholds, feature-main blending ratios, and canonical routing configurations prior to test evaluation. Exploratory settings that fail validation remain appendix-only analyses.

Experiments are conducted on a single NVIDIA RTX PRO 6000 GPU. Uncertainty estimates use 2,000 paired bootstrap resamples. All reported tables are reproduced from frozen manifests, cached predictions, parsed report fields, and deterministic evaluation scripts to ensure replayability.

5 Results

5.1 Main Held-Out Results

Table 3 presents the primary held-out comparison under the unified package-interface setting. Relative to direct Qwen3-Omni-Flash prompting, EviBound improves pooled depression-eligible AUROC from 0.6716 to 0.8658 (+0.1942) and improves F1 from 0.5032 to 0.6557; relative to the strongest direct LMM baseline, Gemini 3.5 Flash, the AUROC gain remains +0.0811 (Table 5). Compared with the 5-way acoustic consensus, the gain of EviBound does not primarily come from stronger acoustic representations, but from explicit evidence governance through manifest-aware routing, structured attribution, and boundary validation under the same evidence interface.

More importantly, EviBound achieves strict evidence-bound consistency across all held-out packages. The system reaches 0% CVR and 100%

EBCPass, while direct omni-modal baselines still produce unsupported acoustic references, semantic overreach under restrictive protocols, or missing-evidence hallucinations. These results suggest that evidence-bounded reasoning can preserve competitive predictive performance while eliminating protocol-boundary violations.

The pooled depression result is computed over eligible evidence profiles, whereas the anxiety benchmark is substantially narrower: 256 of 264 eligible anxiety packages originate from MMPsy interview/feature records. Relative to direct Qwen3-Omni-Flash, the evaluated long-reasoning baseline is lower by 0.0499/0.0322 AUROC on depression/anxiety-eligible tasks, indicating that unconstrained reasoning depth alone does not recover protocol-aware behavior under heterogeneous acquisition settings.

5.2 Protocol Stratification

Table 4 stratifies AUROC by evidence profile. Most depression-eligible support comes from interview-centered packages, whereas prompted and fixed-reading rows serve as smaller protocol probes. Anxiety evaluation is dominated by MMPsy interview records, with only a limited MODMA reading subset.

The largest improvements appear under restrictive acquisition settings. On prompted speech, EviBound improves AUROC from 0.5993 to 0.9779 by suppressing unsupported semantic interpretation and relying primarily on admissible acoustic evidence. Fixed-reading rows show a similar trend despite limited statistical support. These findings support the hypothesis that epistemic flattening disproportionately harms protocols where semantic symptom inference is weakly justified or explicitly invalid.

5.3 Paired Bootstrap Checks

Table 5 summarizes paired bootstrap uncertainty estimates for the primary comparisons. Depression-eligible improvements over direct omni-modal baselines remain consistently above zero. Anxiety intervals should be interpreted more cautiously because the eligible split is heavily dominated by MMPsy interview records.

The comparison against the 5-way acoustic consensus is intentionally qualified because confidence intervals overlap zero for both eligible tasks. Accordingly, EviBound should be interpreted as preserving the predictive strength of the acoustic backbone while adding explicit evidence-bound routing and report validation.

5.4 Secondary Screening Diagnostics

AUROC alone does not fully characterize screening behavior under class imbalance. Appendix diagnostics therefore additionally report AUPRC, Brier score, and validation-selected operating points. Depression-eligible AUPRC exceeds the reported Qwen direct baselines but remains close to the strongest acoustic routes, whereas anxiety-eligible AUPRC intervals are less stable because of limited positive support.

Importantly, these secondary diagnostics are reported independently from boundary-consistency metrics. A model with strong predictive accuracy may still violate protocol constraints, generate unsupported modality claims, or omit missing-evidence disclosure. EviBound is therefore evaluated jointly on predictive utility and evidence-valid reasoning rather than on classification performance alone.

6 Ablation and Analysis

6.1 What Each Component Adds

Table 6 reports depression-eligible ablations on the broader-coverage task. The table separates predictive evidence routes from report-validation components. Base Harness applies package-aware routing without strong acoustic aggregation, while subsequent rows evaluate openSMILE/eGeMAPS, segmented openSMILE, wav2vec2, HuBERT, and WavLM representations.

The 5-way acoustic consensus provides the strongest non-harness acoustic baseline, indicating that acoustic evidence contributes most of the predictive signal. The final EviBound configuration further adds manifest-aware routing, feature-level evidence attribution, and structured boundary validation while preserving essentially identical AUROC. Importantly, the validator does not alter predictive scores because risk estimates are frozen before report repair; instead, it ensures generated claims remain consistent with protocol permissions and available evidence.

6.2 Output-Scope Replay Checks

Replay diagnostics evaluate whether generated reports cite unavailable evidence or violate protocol constraints, measuring report validity independently of predictive metrics. The replay suite

Method	Family	Dep-el. AUC↑	Dep-el. F1↑	Anx-el.^{M} AUC↑	Sev. QWK↑	CVR↓	EBCPass↑
Qwen3-Omni-Flash	Direct LMM	0.6716	0.5032	0.7819	0.4209	0.271	0.714
Qwen3.5-Omni-Plus	Direct LMM	0.6808	0.5200	0.7922	0.4511	0.246	0.736
Gemini 3.5 Flash	Direct LMM	0.7848	0.5318	0.8515	0.3766	0.070	0.931
Qwen3-Omni-Flash Thinking	Thinking LMM	0.6217	0.4372	0.7497	0.3700	0.318	0.661
5-way Acoustic Consensus	Acoustic	0.8652	0.6316	0.8772	0.5232	0.083	0.912
EviBound Structured Harness	Harness	**0.8658**	**0.6557**	**0.8828**	**0.5283**	**0.000**	**1.000**

Table 3: Main held-out comparison on eligible package-interface tasks. Dep-el./Anx-el. denote depression-/anxiety-eligible tasks; Anx-el.^{M} indicates the MMPsy-dominated anxiety split. CVR denotes Claim Violation Rate and EBCPass denotes evidence-bound consistency pass rate. Best results are shown in bold and second-best results are underlined.

Profile	Task	N	Pos/Neg	Qwen3	EviBound	Δ	Note
Interview	Dep-el.	325	73/252	0.7504	**0.8456**	+0.0952	main support
Prompted	Dep-el.	33	16/17	0.5993	**0.9779**	+0.3787	small-n
Reading	Dep-el.	8	5/3	0.5000	**0.7333**	+0.2333	small-n
Interview	Anx-el.	256	30/226	0.8777	**0.8864**	+0.0086	MMP-dom.
Reading	Anx-el.	8	5/3	0.5000	**0.7333**	+0.2333	small-n

Table 4: Held-out AUROC stratified by evidence profile. Δ denotes the AUROC difference between EviBound and Qwen3-Omni-Flash.

Task	Baseline	Δ AUC	95% CI	Interpretation
Dep-el.	Qwen3-Flash	+0.1942	[0.1329, 0.2545]	CI>0
Dep-el.	Qwen3.5-Plus	+0.1851	[0.1225, 0.2502]	CI>0
Dep-el.	Gemini 3.5 Flash	+0.0811	[0.0476, 0.1175]	CI>0
Dep-el.	Qwen3-Think	+0.2442	[0.1732, 0.3153]	CI>0
Dep-el.	5-way	+0.0007	[-0.0155, 0.0170]	overlap
Anx-el.	Qwen3-Flash	+0.1009	[0.0237, 0.1810]	MMPsy-split
Anx-el.	Qwen3.5-Plus	+0.0906	[0.0117, 0.1859]	MMPsy-split
Anx-el.	Gemini 3.5 Flash	+0.0313	[-0.0160, 0.0789]	MMPsy-split
Anx-el.	Qwen3-Think	+0.1331	[0.0276, 0.2378]	MMPsy-split
Anx-el.	5-way	+0.0056	[-0.0168, 0.0267]	overlap

Table 5: Paired bootstrap analysis. Δ AUC denotes the AUROC difference between EviBound and the corresponding baseline.

records acoustic hallucinations, prompted-speech semantic overreach, unsupported listening claims, and feature-scope misuse. Under the final manifest configuration, flagged outputs are repaired or blocked before report emission, enabling EviBound to achieve 0% CVR and 100% EBCPass.

6.3 Shortcut and Source-Prior Controls

The E-DAIC transcript control illustrates one shortcut risk: interviewer prompts alone reach AUROC 0.6410, rivaling participant answers (0.6471), suggesting protocol artifacts carry predictive signal. The benchmark therefore separates participant evidence from prompt/control text at the manifest level. Source and protocol identifiers are retained only for routing and reproducibility, and are explicitly disallowed as participant-level clinical evidence.

Configuration	AUC↑	F1↑	QWK↑
Qwen3-Omni-Flash	0.6716	0.5032	0.4209
Base Harness	0.7270	0.5032	0.4547
Strong openSMILE route	0.8565	0.6243	0.5093
Segmented openSMILE route	0.8503	0.6243	0.5058
wav2vec2 route	0.8390	0.5943	0.5138
HuBERT route	0.8507	0.6036	0.5187
WavLM route	0.8316	0.6154	0.5152
5-way Acoustic Consensus	0.8652	0.6316	0.5232
EviBound Structured Harness	**0.8658**	**0.6557**	**0.5283**

Table 6: Depression-eligible held-out ablation study. Acoustic rows evaluate alternative evidence routes, while EviBound additionally introduces manifest-aware routing and structured boundary validation.

6.4 External-Baseline Tiers and Adapter Check

To ensure fair comparison, external baselines are tiered by manifest compatibility rather than merged into a single leaderboard. An adapter experiment wrapping standard acoustic backbones (openSMILE, wav2vec2, HuBERT, WavLM) within our common harness shows their performance approaches EviBound. This indicates that while acoustic backbones contribute strongly to predictive ranking, EviBound's primary contribution lies in evidence-bound routing, permission control, missing-evidence handling, and structured report validation.

7 Conclusion

We introduced evidence-bounded reasoning for heterogeneous speech-based mental health screening through the Evidence Package Benchmark and EviBound, a protocol-aware harness with explicit evidence routing and boundary validation. Our results highlight a critical insight: merely scaling up or extending LLM reasoning capacity cannot resolve the "epistemic flattening" caused by ignoring protocol boundaries. In contrast, EviBound explicitly enforces these boundaries, achieving 0% CVR and 100% EBCPass while preserving strong predictive performance. We therefore argue that

clinical multimodal systems must be evaluated not only by predictive accuracy, but also by their auditable evidence consistency and protocol-aware reasoning reliability, shifting the paradigm toward safety-constrained clinical NLP.

8 Acknowledgments

This research was supported in part by the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 32500997, S. Li), in part by the Reserve Project in the Field of Brain–Computer Interface supported by the Beijing Municipal Science and Technology Commission (Grant No. Z261100004026006, S. Li), and in part by Beijing Renyixun Health Technology Co., Ltd.

Limitations

Our findings should be interpreted within the following scopes, which also outline avenues for future work.

Benchmark scale and protocol coverage. While the Evidence Package Benchmark integrates six heterogeneous sources, its overall scale remains moderate, and anxiety evaluation is heavily dominated by MMPsy interview records. Furthermore, restrictive profiles (e.g., fixed reading) serve primarily as boundary-stress tests with limited sample sizes. Our current focus is protocol diversity for safety probing rather than raw data volume; expanding longitudinal sampling, multilingual coverage, and broader psychiatric conditions is essential for generalization.

Deterministic rules vs. learned boundaries. EviBound's boundary rules are explicitly encoded to ensure formal verifiability. This is a deliberate trade-off favoring deterministic safety over adaptive flexibility. While this guarantees zero claim violations for known protocols, the validator cannot catch failures outside the defined manifest schema such as open-domain conversational hallucinations. Future work should explore neuro-symbolic approaches that learn boundary constraints while preserving verifiable safety guarantees.

Screening utility vs. clinical deployment. This work targets offline, structured screening under controlled package interfaces, not prospective clinical deployment. EviBound generates evidence-bounded risk assessments, not clinical diagnoses. Transitioning to real-world workflows requires clinician-in-the-loop auditing, robustness to noisy real-world inputs, and rigorous prospective validation, which remain outside the current scope.

Cultural and linguistic bias. The benchmark pools English and Chinese resources acquired under different protocols and population characteristics. Both the predictive routes and the permission matrix may consequently inherit cultural and linguistic biases, and boundaries validated on our audit cohort may not transfer to other languages, cultures, or acquisition settings.

Ethical Statement

All datasets used in the Evidence Package Benchmark are de-identified public resources; we do not collect or access any private health information. While EviBound enforces strict evidence consistency to reduce hallucinations and protocol misuse, it may still inherit biases present in the underlying speech and text data across languages, genders, or cultural groups. We therefore emphasize that any practical use would require human-in-the-loop auditing, prospective clinical validation, and explicit mitigation of demographic biases before considering real-world deployment.

References

Tuka Al Hanai, Mohammad M Ghassemi, and James R Glass. 2018. Detecting depression with audio/text sequence modeling of interviews. In *Interspeech*, pages 1716–1720.

Mai Ali, Christopher Lucasius, Tanmay P. Patel, Madison Aitken, Jacob Vorstman, Peter Szatmari, Marco Battaglia, and Deepa Kundur. 2025. Speech as a multimodal digital phenotype for multi-task LLM-based mental health prediction. *Preprint*, arXiv:2505.23822.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. HealthBench: Evaluating large language models towards improved human health. *Preprint*, arXiv:2505.08775.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Hanshu Cai, Zhenqin Yuan, Yiwen Gao, Shuting Sun, Na Li, Fuze Tian, Han Xiao, Jianxiu Li, Zhengwu

Yang, Xiaowei Li, Qinglin Zhao, Zhenyu Liu, Zhijun Yao, Minqiang Yang, Hong Peng, Jing Zhu, Xiaowei Zhang, Guoping Gao, Fang Zheng, and 8 others. 2022. A multi-modal open dataset for mental-disorder analysis. *Scientific data*, 9(1):178.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Brian Cho, Estée Balles, Michael Mackinley, Paulina Dzialoszynski, Sabrina Ford, Rohit Lodhi, and Lena Palaniyappan. 2026. The DISCOURSE in psychosis (London Ontario): A speech dataset to examine communication disturbances in early-stage psychosis. *Data in Brief*, 65:112517.

Oliver Delgaram-Nejad, Dawn Archer, Gerasimos Chatzidamianos, Louise Robinson, and Alex Bartha. 2023. The DAIS-C: A small, specialised, spoken, schizophrenia corpus. *Applied Corpus Linguistics*, 3(3):100069.

Sahraoui Dhelim, Liming Chen, Huansheng Ning, and Chris Nugent. 2023. Artificial intelligence for suicide assessment using audiovisual cues: a review. *Artificial Intelligence Review*, 56(6):5591–5618.

Bowen Dong, Minheng Ni, Zitong Huang, Guanglei Yang, Wangmeng Zuo, and Lei Zhang. 2025. Mirage: Assessing hallucination in multimodal reasoning chains of mllm. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 122910–122955. Curran Associates, Inc.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. openSMILE: The Munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 1459–1462.

Ming Fang, Siyu Peng, Yujia Liang, Chih-Cheng Hung, and Shuhua Liu. 2023. A multimodal fusion model with multi-level attention mechanism for depression detection. *Biomedical Signal Processing and Control*, 82:104561.

Aya E. Fouda, Abdelrahman A. Hassan, Radwa J. Hanafy, and Mohammed E. Fouda. 2026. PsychiatryBench: A multi-task benchmark for LLMs in psychiatry. *npj Digital Medicine*, 9:320.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Google AI for Developers. 2026. Gemini 3.5 Flash. Last updated: 2026-05-19; accessed: 2026-05-26.

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. 2014. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jiuzhou Han, Wray L Buntine, and Ehsan Shareghi. 2025. Verifiagent: a unified verification agent in language model reasoning. In *EMNLP (Findings)*, pages 16410–16431.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, Bo Zhang, Chaoyou Fu, Peng Gao, and Hongsheng Li. 2025. Mme-cot: benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org.

Miao Jing, Mengting Jia, Junling Lin, Zhongxia Shen, Huan Gao, Mingkun Xu, and Shangyang Li. 2026. Beyond classification accuracy: Neural-medbench and the need for deeper reasoning benchmarks. In *The Fourteenth International Conference on Learning Representations*.

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

Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. 2025. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *Preprint*, arXiv:2505.21523.

Yunsong Liu, Zunamys I Carrero, Xiaofeng Jiang, Dyke Ferber, Georg Wölflein, Li Zhang, Sanddhya Jayabalan, Tim Lenz, Zhouguang Hui, and Jakob Nikolas Kather. 2026. Benchmarking large language model-based agent systems for clinical decision tasks. *npj Digital Medicine*.

Hongxi Mao, Wei Zhou, Mengting Jia, Tao Fang, Huan Gao, Bin Zhang, and Shangyang Li. 2026. Schema-adaptive tabular representation learning with llms for generalizable multimodal clinical reasoning. *Preprint*, arXiv:2604.11835.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson,

Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229.

Bo Ni, Zheyuan Liu, Leyao Wang, Yongjia Lei, Yuying Zhao, Xueqi Cheng, Qingkai Zeng, Luna Dong, Yinglong Xia, Krishnaram Kenthapadi, Ryan Rossi, Franck Dernoncourt, Md Mehrab Tanjim, Nesreen Ahmed, Xiaorui Liu, Wenqi Fan, Erik Blasch, Yu Wang, Meng Jiang, and Tyler Derr. 2025. Towards trustworthy retrieval augmented generation for large language models: A survey. *Preprint*, arXiv:2502.06872.

Rumo Pan, Sidu Feng, Yi Sun, Jinqiu Xu, Tianzhang Zhai, Xiaochun Wu, Liangliang Tan, Yonggui Yuan, Hanlin Cao, Maierhaba Maimaitimin, Qing Yao, Hao Guo, Mah Wing Xuan, Jian Li, and Zhi Xu. 2026. Cfrafn: A cross-feature residual attention fusion network for major depressive disorder prediction using clinical voice recordings. *IEEE Journal of Biomedical and Health Informatics*, pages 1–14.

Jinghui Qin, Changsong Liu, Tianchi Tang, Dahuang Liu, Minghao Wang, Qianying Huang, and Rumin Zhang. 2025. Mental-Perceiver: Audio-textual multimodal learning for estimating mental disorders. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):25029–25037.

Qwen Team. 2025. Qwen3-Omni technical report. *Preprint*, arXiv:2509.17765.

Qwen Team. 2026. Qwen3.5-Omni technical report. *Preprint*, arXiv:2604.15804.

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. 2019. Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, AVEC '19, page 3–12, New York, NY, USA. Association for Computing Machinery.

Ying Shen, Huiyu Yang, and Lin Lin. 2022. Automatic depression detection: an emotional audio-textual corpus and a gru/bilstm-based model. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6247–6251.

Hoyun Song, Migyeong Kang, Jisu Shin, Jihyun Kim, Chanbi Park, Hangyeol Yoo, Jihyun An, Alice Oh, Jinyoung Han, and KyungTae Lim. 2026. MentalBench: A benchmark for evaluating psychiatric diagnostic capability of large language models. *Preprint*, arXiv:2602.12871.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621.

Dingdong WANG, Shujie LIU, Tianhua Zhang, Youjun Chen, Jinyu Li, and Helen M. Meng. 2026. Emotion-thinker: Prosody-aware reinforcement learning for explainable speech emotion reasoning. In *The Fourteenth International Conference on Learning Representations*.

Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. 2025. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *Preprint*, arXiv:2503.18968.

Ling Yue, Sixue Xing, Jintai Chen, and Tianfan Fu. 2024. Clinicalagent: Clinical trial multi-agent system with large language model-based reasoning. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–10.

Wei Zhang, Juan Chen, En Zhu, Wenhong Cheng, Yun-Peng Li, Yuhan Li, and Yanbo J Wang. 2026. Mllm-dr: Towards explainable depression recognition with multimodal large language models. *ACM Transactions on Multimedia Computing, Communications and Applications*, 22(4):1–23.

Xiangyu Zhang, Hexin Liu, Kaishuai Xu, Qiquan Zhang, Daijiao Liu, Beena Ahmed, and Julien Epps. 2024. 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*, pages 146–158, Miami, Florida, USA. Association for Computational Linguistics.

Xiangyu Zhang, Hexin Liu, Qiquan Zhang, Beena Ahmed, and Julien Epps. 2025. SpeechT-RAG: Reliable depression detection in LLMs with retrieval-augmented generation using speech timing information. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10019–10030, Vienna, Austria. Association for Computational Linguistics.

Xianda Zheng, Huan Gao, Meng-Fen Chiang, Michael J. Witbrock, Kaiqi Zhao, and Shangyang Li. 2026. Evo-PI: Aligning medical reasoning via evolving principle-guided supervision. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31835–31850, San Diego, California, United States. Association for Computational Linguistics.

Weihai Zhi, Jiayan Guo, and Shangyang Li. 2026. Medgr2: Breaking the data barrier for medical reasoning via generative reward learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 28901–28909.

Bochao Zou, Jiali Han, Yingxue Wang, Rui Liu, Shenghui Zhao, Lei Feng, Xiangwen Lyu, and Huimin Ma. 2023. Semi-structural interview-based chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders. *IEEE Transactions on Affective Computing*, 14(4):2823–2838.

A Evidence Package and Same-Interface Protocol

This appendix reports implementation details supporting the main evidence-bound evaluation claims. The benchmark operates over frozen evidence packages rather than unconstrained narratives. Each package stores a source id, split id, evidence profile, modality mask, missing-evidence list, permitted claim families, and evaluator-only targets. Direct LMM baselines and EviBound receive identical model-visible fields, while PHQ, GAD, and diagnosis-derived labels remain evaluator-side only.

Package field	Role in the same-interface evaluation
Evidence profile	Acquisition protocol; controls which claim families can be reported.
Modality mask	Availability of audio, transcript, feature, and scale channels; absent channels block matching evidence claims.
Visible inputs	Audio clips, transcript excerpts, or feature summaries made available to the compared method.
Missing evidence	Explicit unavailable evidence that must be disclosed in generated reports.
Claim permissions	Non-diagnostic screening scope, source restrictions, and required blocked-claim hints.
Evaluator targets	Hidden labels or scores used only for metric computation.

Table 7: Evidence-package fields used to keep the compared methods under a common input contract.

The frozen benchmark contains 1870 packages from six sources: CMDC (78), DAIS-C (28), E-DAIC (275), EATD (162), MMPsy (1275), and MODMA (52), with 1284/214/372 train/validation/test splits. Package-level modality totals are 567 audio, 1656 transcript, 1550 feature, and 1842 scale records. Labels and scale values are never serialized into prompts or cached model-visible inputs.

The evidence-package abstraction fixes the control-variable interpretation of the baseline comparison. Qwen, Gemini, acoustic routes, and EviBound are evaluated under the same package interface with identical modality masks, missing-evidence fields, and output schemas.

B Boundary Validation and Replay

The boundary layer is deterministic with respect to report validity. It verifies modality usage, blocked claims, missing-evidence disclosure, and protocol permissions without modifying the predictive score used for AUROC, F1, or QWK. This score-preservation property ensures that boundary validation cannot improve ranking metrics post hoc.

The frozen claim-permission audit covers 1870 packages and 1870 report outputs, with zero failed records, zero score changes, and zero permission violations. Screening-risk reporting is permitted for all packages, while formal diagnosis and treatment recommendation are globally blocked. Release levels include 622 bounded automatic reports, 97 evidence-scoped reports, and 1151 manual-review reports.

The output-contract audit verifies required blocked claims, source attribution, and missing-evidence handling over all 1870 packages. The boundary-action replay audit contains 3955 action-labeled instances (786 held out). The cleanup guard changed 84 rows and removed 168 unsupported acoustic route/tool actions while removing zero gold actions. Held-out exact match, macro F1, and micro F1 are all 1.0000. The replay suite evaluates five primary failure modes: 1,870 formal-diagnosis-block checks, 595 feature-scope checks (citing absent feature sources), 1,248 uncertainty-review checks (escalating weak/conflicting evidence), 28 text-only acoustic checks (citing voice when no audio is available), and 214 weak-profile semantic checks (inferring symptom history from scripted text).

C Boundary Replay Cases

Table 8 shows representative replay cases under the frozen evidence-package interface. Cases are label-free cached fragments used for interpretability rather than model selection.

Computational environment. All local preprocessing, evidence-package construction, acoustic/feature routing, boundary replay, validation, and metric computation were run on a Linux workstation with an AMD Ryzen Threadripper PRO 7965WX CPU (24 cores, 48 threads), 188 GiB RAM, and two NVIDIA RTX PRO 6000 Blackwell Workstation Edition GPUs (97,887 MiB memory each; driver 580.126.09). The operating system used Linux kernel 6.8.0-110-generic. Experiments were executed in a Conda environment with Python 3.10.20, PyTorch 2.8.0+cu129, CUDA 12.9, NumPy 2.2.6, pandas 2.3.3, scikit-learn 1.7.2, SciPy 1.15.3, librosa 0.11.0, soundfile 0.13.1, and

Replay case	Cached boundary trace
Supported interview	E-DAIC clinical interview with audio, transcript, features, and scale context available. Acoustic evidence, feature evidence, non-diagnostic blocking, and review triage are all supported. EviBound keeps the reporting decision, preserves the risk score, and allows bounded automatic reporting.
Supported manual review	MMPsy clinical-interview package with transcript and features, but no raw audio. EviBound keeps feature/transcript attribution, reports missing raw audio and longitudinal history, preserves the score, and retains manual review.
Audio removed	E-DAIC interview counterfactual where the report still contains an acoustic claim but `audio=false`. EviBound blocks `acoustic_evidence`, escalates to manual review, and leaves the risk score unchanged.
Prompted-speech mutation	E-DAIC interview counterfactual changed to prompted-speech with transcript present. The stressor treats weak-profile text as free symptom-history evidence. EviBound blocks `transcript_semantics` for symptom reasoning, requires review, and preserves the score.
Missing non-diagnostic block	E-DAIC interview counterfactual where the report omits the required formal-diagnosis block. EviBound restores the blocked-claim requirement and routes the report to review without changing the score.

Table 8: Boundary replay case studies with label-free cached fragments. Real supported cases show when EviBound permits evidence-scoped reporting; counterfactual stressors show how unsupported claims are blocked without modifying the predictive score.

Transformers 5.8.0.dev0. Commercial direct LMM baselines were executed through vendor APIs and cached; local hardware was used for package assembly, parsing, validation, and metric computation.