RETHINK-MH: RELIABILITY-TRIGGERED MULTIMODAL RETHINKING FOR MENTAL HEALTH SCREENING

Anonymous authors
Paper under double-blind review

ABSTRACT

Given the sensitive and identifiable nature of clinical interviews, multimodal mental health screening increasingly relies on released, de-identified measurements rather than raw recordings. Most existing systems compress these measurements into a single risk score, leaving no way to inspect which observations support a decision, which contradict it, or how conflicts between them were resolved. We present RETHINK-MH, a reliability-triggered multimodal rethinking framework that turns screening into an auditable two-pass process. A deterministic compiler organizes the acoustic, visual, and transcript measurements of each session into a two-layer evidence store: a coarse segment layer for the first-pass assessment and an indexed atomic layer that remains sealed for re-examination. An audit module estimates the error risk of each first-pass judgment from source reliability and validity-weighted cross-source disagreement, and only flagged cases trigger targeted retrieval of atomic evidence under an explicit access budget. A deterministic validator closes each case as preserved, revised, or unresolved, and unresolved or invalid cases are referred to human review, so every decision carries a checkable trail from outcome back to evidence. The loop is trained on its own cross-fitted errors via cold-start supervised fine-tuning followed by iterative preference alignment. On DAIC-WOZ, E-DAIC, and D-Vlog, RETHINK-MH achieves competitive and state-of-the-art screening performance, while correcting more first-pass errors than it introduces and recovering correct decisions for cases that would otherwise be referred to human review. Code is available at `https://anonymous.4open.science/r/rethink-mh-85F3/`.

1 INTRODUCTION

Standardized questionnaires are a routine first step in identifying people at risk of depression (Kroenke et al., 2009) and are often administered during clinical interviews. Automated screening from recorded interviews and video blogs has become an active research area, with multimodal benchmarks such as DAIC-WOZ, E-DAIC, and D-Vlog (Gratch et al., 2014; Ringeval et al., 2019; Yoon et al., 2022). Systems on these corpora typically fuse acoustic and visual cues on the premise that the sources are complementary (Song et al., 2018; Wei et al., 2022; Gimeno-Gómez et al., 2024). For privacy reasons, however, the corpora release de-identified measurements rather than the recordings themselves. These measurements differ across sources in type and temporal resolution and contain tracking loss and missing data, yet every screening score must be built from them.

As reported accuracy on these benchmarks continues to rise (Wei et al., 2022; Gimeno-Gómez et al., 2024; Ye et al., 2025; Ma et al., 2026), the central question is shifting from whether a score can be produced to whether its basis can be verified. Reproduction studies show that some published classifiers learn little task-relevant signal and that shortcuts such as interviewer prompts can inflate their results (Burdisso et al., 2024). Because missed cases and false alarms carry different costs in screening, a score that cannot be inspected offers weak guidance for clinical action.

Most systems predict in a single forward pass, mapping measurements to a risk score through modality encoders and a fusion network. This design has two failure modes. First, tracking loss, noise, and missing windows are common in interview data, so a confident prediction may rest on corrupted measurements. Second, even intact channels often disagree, and latent-space fusion hides

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

this conflict rather than exposing it. Asking the model to reflect on its answer is a natural remedy, but re-prompting on the same input without new evidence largely repeats the first answer, as shown in Fig. 1. Moreover, none of these systems leaves a record that a clinician can verify.

Our key observation is that a second pass helps only if it sees evidence that the first pass did not. We therefore split the released measurements into two layers: a coarse segment layer read by the first pass, and a fine-grained atomic layer withheld for re-examination. Because retrieval costs tokens and can introduce errors, it should occur only when needed. We flag likely errors from source reliability and cross-source conflict, retrieve targeted atomic units only for flagged cases, and refer cases that remain unresolved to human review.

Building on this design, we present RETHINK-MH, a reliability-triggered rethinking framework that turns screening into an auditable two-pass loop, as illustrated in Fig. 2. A Screener reads the segment layer; an audit flags likely errors and prompts a Query Planner to retrieve targeted atomic units; and a Rethinker keeps, revises, or refers the judgment under deterministic validation. The loop learns from its own errors on held-out training folds via supervised fine-tuning of a Qwen2.5-Omni backbone (Xu et al., 2025) followed by iterative preference alignment (Hong et al., 2024) on filtered self-generated pairs (Yuan et al., 2023). On all three corpora, the loop corrects more first-pass errors than it introduces, whereas a second pass on the same evidence breaks as many decisions as it fixes. Our contributions are as follows:

- We propose a deterministic two-layer evidence store that time-aligns measurements, records their provenance, and flags missing data, so any second pass receives new, targeted evidence.
- We design a reliability-triggered rethinking loop that flags likely errors from validity-weighted cross-source conflict, retrieves targeted atomic evidence, and terminates under deterministic schema, grounding, and status checks. We further show that rethinking helps only when the second pass sees new, informative evidence and the trigger targets actual errors.
- We develop a training recipe that derives action and revision supervision from the model's own out-of-fold errors, combining dual-view fine-tuning with iterative preference optimization.
- We evaluate on DAIC-WOZ, E-DAIC, and D-Vlog, achieving the best F1 on E-DAIC among compared published methods, competitive accuracy on DAIC-WOZ and D-Vlog, and providing an auditable decision trail for every case.

2 Related Work

Multimodal Mental Health Screening. Multimodal depression screening is commonly benchmarked on DAIC-WOZ, E-DAIC, and D-Vlog (Gratch et al., 2014; Ringeval et al., 2019; Yoon et al., 2022). Methods have progressed from hand-crafted spectral features (Williamson et al., 2016) to learned fusion of non-verbal cues (Gimeno-Gómez et al., 2024; Yang et al., 2026), cross-modal Transformers and shared–private subspace models (Tsai et al., 2019; Hazarika et al., 2020; Han et al., 2021), spatiotemporal and multi-scale attention (Tao et al., 2024a; Nguyen et al., 2026; Han et al., 2026), and Mamba-based sequence models (Ye et al., 2025; Zhou et al., 2025; Ma et al., 2026). Large language models (LLMs) now also extract questionnaire items from interview transcripts (Sadeghi et al., 2024; Weber et al., 2025) and score interviews with rationales (Zhang et al., 2026). These methods map the released features to a score in a single pass, so tracking loss and cross-source conflict are absorbed into a latent representation rather than detected and handled. Reproduction studies further attribute some reported gains to interviewer-prompt leakage (Burdisso et al., 2024), which a score without checkable evidence cannot reveal.

Uncertainty, Selective Prediction, and Deferral. Calibration studies based on proper scoring rules show that modern networks are often overconfident (Gneiting & Raftery, 2007; Guo et al., 2017). Selective prediction abstains on uncertain inputs, trading coverage for accuracy (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2019); conformal prediction provides distribution-free coverage guarantees (Angelopoulos & Bates, 2023); and learning to defer jointly trains the classifier and its hand-off to human experts (Mozannar & Sontag, 2020). Semantic entropy detects unreliable LLM generations (Farquhar et al., 2024), and channel reliability can weight fusion under missing modalities (Wen et al., 2026). All of these methods decide whether to answer with the information at hand, so an uncertain case can only be answered or deferred. Active feature acquisition does acquire extra inputs before predicting (Shim et al., 2018; Li & Oliva, 2021), but it operates on fixed feature vectors with a black-box classifier and does not record how the new values changed the decision.

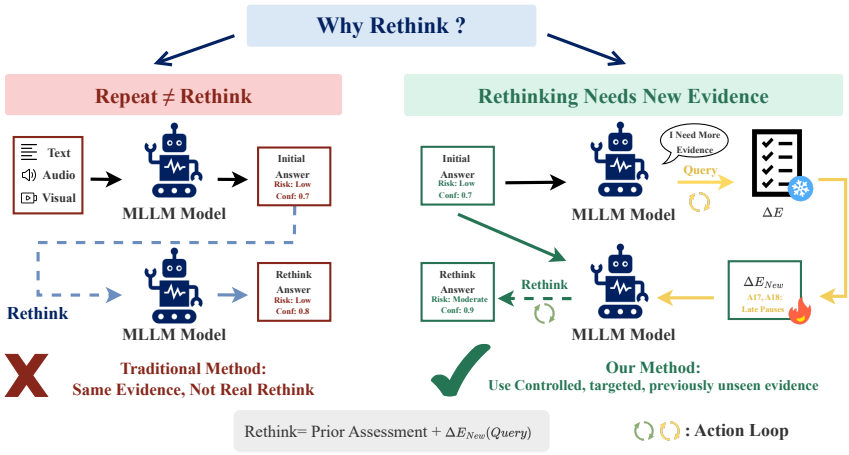

Figure 1: Rethinking requires new evidence. Re-prompting a model on the same evidence repeats its first answer, whereas a useful second pass adds evidence $\Delta\mathcal{E}_{\text{new}}$ unseen by the first pass.

Grounded Self-Correction and Preference Alignment. Self-refinement and reflection let a model revise its own output (Madaan et al., 2023; Shinn et al., 2023), and reasoning agents interleave generation with tool calls (Yao et al., 2023). Without external feedback, however, self-correction does not reliably improve reasoning (Huang et al., 2024); it helps only when verifiable conditions or tool feedback enter the loop (Wu et al., 2024; Gou et al., 2024). Retrieval-augmented generation and its active variants ground generation in retrieved documents (Lewis et al., 2020; Jiang et al., 2023), and attribution methods attach citations to generated text (Gao et al., 2023b;a), but both operate on open text corpora, and their citations are assessed post hoc rather than checked against what was actually retrieved. Preference optimization aligns models with pairwise feedback (Rafailov et al., 2023; Hong et al., 2024), and rejection-sampling fine-tuning turns filtered self-generated outputs into training data (Yuan et al., 2023). None of these lines of work addresses when a clinical judgment should be kept, revised, or handed to a human, or how such a revision can be verified against multiple measurement streams. See Appendix A for an extended review.

3 Method

3.1 Problem Formulation

We study mental health screening from clinical interviews when only de-identified measurements are available, the usual setting because interview recordings identify their participants. Each session provides measurement streams $\mathcal{M}^s$ from two sources $s \in \mathcal{S}$: audio, with acoustic descriptors, and video, with facial action units and gaze and head-pose tracks. Each stream carries per-window quality and missing-data flags, and the two interview corpora additionally provide a transcript. The label $y \in \{0, 1\}$ is positive when the released questionnaire score reaches a clinical cut-off (e.g., PHQ-8 ≥ 10); it therefore indicates screening risk, not a diagnosis.

A standard model maps all measurements to a single risk score, hiding which observations support the decision, which oppose it, and how the sources were reconciled. Instead, we let the model answer again when its first answer appears unreliable, and we require every answer to cite programmatically verifiable evidence. For each session, the system returns a decision $\hat{y}$ with a record $\mathcal{C}$ of the first answer, the audit, the retrieved evidence, and the final outcome; see the top of Fig. 2.

3.2 Two-Layer Evidence Store

A deterministic compiler Φ converts the measurements of a session into a two-layer evidence store. Throughout, we split the training participants into K folds and fit every component once per fold on the other $K - 1$ folds, so that each training participant also receives predictions from a model that never saw it. In each fold k, the compiler first fits a label-free reference $\mathcal{R}_k$ on the training participants of that fold, storing robust location, scale, and percentiles for every measurement. The *segment layer* $\mathcal{E}^{\text{seg}}$ partitions the session into at most 24 contiguous segments, each summarized as a short text brief with an ID such as `S011`. The *atomic layer* $\mathcal{E}^{\text{atom}}$ keeps one unit per measurement

window, indexed by segment, window, and source, with an ID such as `E0687`. Each unit reports its percentile under the reference, its source and time span, a quality score, and whether its data are present, partial, missing, or masked, so missing data are never read as zeros.

The first pass sees only the segment layer, whereas a second pass sees the atomic units $R(q) \subseteq \mathcal{E}^{\text{atom}}$ returned by a query q and thus gains exactly

$$\mathcal{V}^0 = \text{render}(\mathcal{E}^{\text{seg}}), \qquad \mathcal{V}^q = \text{render}\big(R(q)\big), \qquad \Delta\mathcal{E}(q) = \mathcal{V}^q \setminus \mathcal{V}^0. \tag{1}$$

Because each segment brief summarizes dozens of windows, $\Delta\mathcal{E}(q)$ contains detail that the first pass never saw. Rendering every window of a long session would exceed the token budget, so the second pass must be selective. Since Φ is deterministic and label-free, every rendered statement is reproducible and every cited ID is programmatically verifiable. Training additionally uses the *full view* $\mathcal{V}^1$ of both layers, which is never shown at inference.

3.3 Reliability-Triggered Rethinking

Initial Assessment. The *Screener* reads $\mathcal{V}^0$ and returns an initial assessment A^0 comprising a risk probability p^0, which yields the first-pass decision $\hat{y}^0 = \mathbf{1}[p^0 \geq \tau]$ at threshold τ, and a confidence c^0. For each source s, it also outputs a risk reading h^s and a reliability $r^s \in [0, 1]$; the latter rates channel quality, not the probability that the judgment is correct. Finally, it lists up to four segments that support the judgment, contradict it, or leave it uncertain, and up to four segments to re-examine.

Audit. Screening cues are weak and non-specific, so a confident judgment may rest on one fragile channel; disagreement between channels is the observable symptom. Reliability indicates whether a channel can be trusted, not whether it is informative about the label. We therefore assign each source a validity $v_s \in [0, 1]$, defined as the Gini coefficient of the out-of-fold ROC curve for its readings h^s, i.e., $v_s = 2\,A_s - 1$ clamped to $[0, 1]$, where A_s is the per-source AUROC. A well-tracked but uninformative channel thus gets little weight. The audit measures disagreement in two ways:

$$D_{\text{rel}} = \sum_{s<s'} r^s r^{s'} v_s v_{s'} \left|h^s - h^{s'}\right|, \tag{2}$$

$$D_{\text{unrel}} = \sum_{s<s'} \mathbf{1}\big[\max(r^s, r^{s'}) > \tau_r\big]\big(1 - \min(r^s, r^{s'})\big)\left|h^s - h^{s'}\right|, \tag{3}$$

where τ_r is a reliability threshold. The term D_{rel} captures conflict between reliable, informative sources, and D_{unrel} captures conflict between a reliable source and a degraded one. An L_2-regularized logistic regression g_ϕ maps these two terms, the confidence, the reliabilities, the fraction of missing sources, and the entropy of p^0 to an error risk ρ, and the case is flagged ($z = 1$) when $\rho \geq \tau_\rho$. We fit g_ϕ to predict first-pass errors on out-of-fold initial assessments, augmented with synthetically degraded cases whose targets are known by construction, and set τ_ρ on development data to achieve a target trigger rate. Unflagged cases keep their initial assessment.

Targeted Re-Query. For a flagged case, the *Query Planner* generates structured queries q from A^0 and $\mathcal{V}^0$, starting from the segments the Screener marked for re-examination. Each query specifies a segment, the measurements to inspect, and the type of pattern to look for. Retrieval returns the matching atomic units as candidate cards, from which the model selects the units to read. At most $B = 12$ units are returned, drawn from up to four segments, with at most three windows per segment and three observations per source. See Appendix H for the prompts of all steps.

Revision and Validation. Given A^0, the risk ρ, and the retrieved units, an action policy first commits to keep, revise upward, revise downward, or refer, so that this choice can be supervised separately. The *Rethinker* then writes a revised assessment A^1 containing a status $\sigma \in \{\text{preserved}, \text{revised}, \text{unresolved}\}$, a revised probability p^1 and confidence, the cited evidence IDs, each marked as kept or rejected, any remaining conflict, and a one-line summary. Because A^1 is conditioned on $\Delta\mathcal{E}(q)$ rather than on $\mathcal{V}^0$ alone, the second pass is a genuine rethink rather than a repeat. A deterministic validator checks that the output conforms to the schema, that every cited ID was retrieved, and that the status is consistent with the content; we write $\text{valid}(A^1)$ when all three checks pass. The final decision is

$$\hat{y} = \begin{cases} \mathbf{1}[p^0 \geq \tau], & z = 0 \ \text{ or } \ \sigma = \text{preserved} \wedge \text{valid}(A^1), \\ \mathbf{1}[p^1 \geq \tau], & \sigma = \text{revised} \wedge \text{valid}(A^1), \\ \text{refer}, & \text{otherwise}. \end{cases} \tag{4}$$

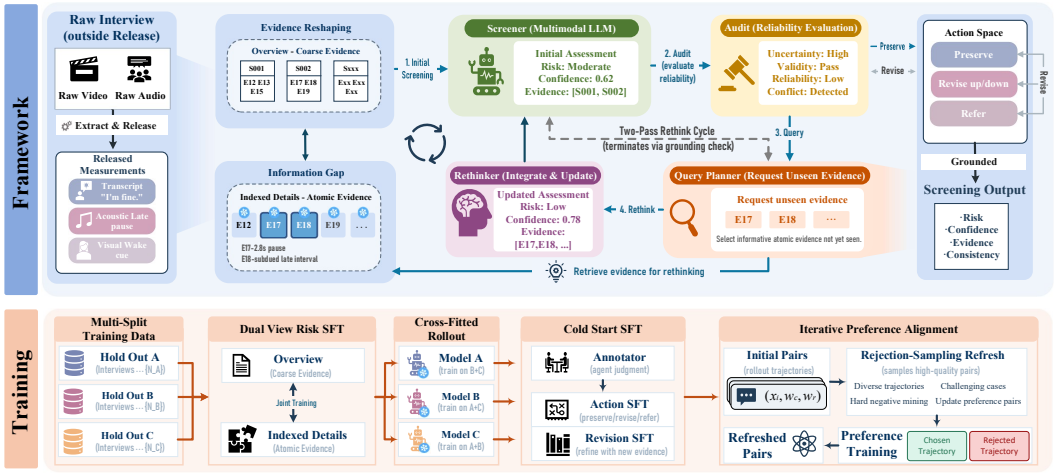

Figure 2: Overview of RETHINK-MH. Top, inference: in Step 1, the Screener assesses the segment layer; in Step 2, the audit estimates the risk that this assessment is wrong; in Step 3, the Query Planner retrieves unseen atomic units for flagged cases; and in Step 4, the Rethinker keeps, revises, or refers the case under a deterministic validator. Bottom, training stages.

Unresolved and invalid cases are thus always referred to human review, because a missed case costs more than an extra review, and the record $\mathcal{C}$ is verifiable end to end.

3.4 Training on Cross-Fitted Errors

Training consists of a cold-start supervised fine-tuning (SFT) phase followed by a preference-alignment phase, as illustrated at the bottom of Fig. 2. The corpora label only the outcome y, not whether a judgment should be kept, revised, or referred, so we derive this supervision from the model's own out-of-fold errors.

Dual-View Risk SFT. From the base model, we train risk prediction on both views with equal weight, $\mathcal{L}_{\mathrm{dv}} = \frac{1}{2}\,\mathrm{BCE}(p^0, y) + \frac{1}{2}\,\mathrm{BCE}(\tilde{p}^1, y)$, where $\tilde{p}^1$ is predicted from the full view $\mathcal{V}^1$. The model thus learns which cases the segment briefs can decide and which require atomic detail.

Supervision from Cross-Fitted Errors. Running each fold's dual-view model on its held-out participants yields an out-of-fold pair $(p^0, \tilde{p}^1)$ for every training participant. When the two predictions fall on opposite sides of τ, we shrink the revision toward the threshold:

$$p^* = \begin{cases} \tau + \alpha\,(\tilde{p}^1 - \tau), & (p^0 - \tau)(\tilde{p}^1 - \tau) < 0, \\ \tilde{p}^1, & \text{otherwise}, \end{cases} \tag{5}$$

with $\alpha = 0.5$ in both directions to avoid directional bias. A deterministic annotator reads the labels only after all rollouts are frozen. It labels a case *keep* if the first pass is correct; *revise* if p^* is correct, lowers both the Brier score and the log loss, and shifts the probability by at least 0.05; and *refer* otherwise. From these labels, it constructs short action targets, full revision targets, and preference pairs whose two outputs share a prompt.

Action-SFT and Revision-SFT. From the dual-view model, Action-SFT trains it to choose among keep, revise, and refer under class-balanced sampling. Revision-SFT then trains it to write the revised assessment from the retrieved units, supervising the action, its direction, the schema, and the cited evidence. After these two stages, the model can run the entire loop by imitation.

Preference Alignment. We then apply odds-ratio preference optimization (Hong et al., 2024),

$$\mathcal{L}_{\mathrm{pref}} \;=\; -\log P_{\boldsymbol{\theta}}(w_c \mid x) \;-\; \beta \log \sigma\!\left(\log \frac{P_{\boldsymbol{\theta}}(w_c|x)}{1-P_{\boldsymbol{\theta}}(w_c|x)} - \log \frac{P_{\boldsymbol{\theta}}(w_r|x)}{1-P_{\boldsymbol{\theta}}(w_r|x)}\right), \tag{6}$$

where $P_{\boldsymbol{\theta}}$ denotes the model likelihood, and the chosen output w_c and rejected output w_r respond to the same prompt x. The chosen output is valid, grounded, and consistent with the outcome. The rejected output is grounded but wrong, such as a harmful revision of a case that should be kept, a missed or wrong-direction revision of a case that should be revised, or a confident answer on a case

that should be referred. We run two rounds with this objective: the first uses the annotator's pairs, and the second rolls out the current model once on the frozen training prompts and retains outputs that pass outcome, schema, and grounding filters; see Appendix C. Because the prompts form a fixed offline set, this procedure is offline preference optimization rather than online reinforcement learning. Finally, we fit a calibrator on a separate held-out split.

3.5 When Does Rethinking Help?

We now ask when a second pass can beat the first. For a loss ℓ and an observable X, let $L^\star_\ell(X) = \inf_f \mathbb{E}[\ell(Y, f(X))]$ denote the Bayes risk of predicting Y from X. See Appendix B for proofs.

Proposition 1 (No-New-Evidence Ceiling). *Let $A^0 = f_0(\mathcal{V}^0, U)$ and $A^1 = f_1(\mathcal{V}^0, A^0, U)$, where the randomness U is independent of $(Y, \mathcal{V}^0)$. Then $L^\star_\ell(A^0, A^1) \geq L^\star_\ell(\mathcal{V}^0)$ for any ℓ.*

However long the model reflects, re-prompting on the same view can at best close the gap between the first pass and the Bayes-optimal predictor on $\mathcal{V}^0$. A well-trained Screener leaves little of this gap, so re-prompting fixes about as many decisions as it breaks, as Sec. 4.4 confirms.

Proposition 2 (Value of Retrieved Evidence). *Under log loss, for any query q,*

$$L^\star_{\log}(\mathcal{V}^0) - L^\star_{\log}\big(\mathcal{V}^0, \Delta\mathcal{E}(q)\big) = I\big(Y; \Delta\mathcal{E}(q) \mid \mathcal{V}^0\big) \geq 0, \tag{7}$$

so the ideal query under budget B is $q^\star \in \arg\max_{|R(q)|\leq B} I(Y; \Delta\mathcal{E}(q) \mid \mathcal{V}^0)$.

Units that merely restate their brief add no information, so the retrieval budget pays off only where the brief hides label-relevant detail. The ideal query also depends on the joint distribution of the label and the evidence, so a retrieval rule that ignores the outcome need not approach it. An outcome-informed diagnostic on DAIC-WOZ supports both points: the withheld layer contains useful evidence, yet a fixed label-free probe underperforms keeping the first pass; see Appendix E.

Proposition 3 (Error-Precision Condition). *Let $\pi_T = P(z = 1)$ be the trigger rate and $\varepsilon_T = P(\hat{y}^0 \neq Y \mid z = 1)$ its error precision. Let κ_r be the probability that a flagged case with a wrong first pass is accepted with a correct decision (repair), and κ_h the probability that one with a correct first pass is accepted with a wrong decision (harm). The expected net correction, i.e., the fraction of cases turned from wrong to correct minus that turned from correct to wrong, satisfies*

$$\mathbb{E}[\mathrm{NetCorr}] = \pi_T\big[\kappa_r\,\varepsilon_T - \kappa_h\,(1 - \varepsilon_T)\big] \leq \pi_T\,\varepsilon_T, \tag{8}$$

which is positive if and only if $\varepsilon_T/(1 - \varepsilon_T) > \kappa_h/\kappa_r$.

Thus, a trigger must select errors, not merely uncertain cases, and referral removes a case from both terms, trading coverage for safety. On E-DAIC at the same trigger rate, an entropy trigger achieves only half the net correction of the audit, although the two predict first-pass errors about equally well. Eq. 8 thus implies that the audit selects cases with a higher repair-to-harm ratio κ_r/κ_h, namely cases in which a reliable source contradicts the first pass and new evidence can resolve the conflict.

4 Experiments

4.1 Datasets and Protocol

We evaluate on three feature-only corpora. DAIC-WOZ (Gratch et al., 2014) contains 189 clinical interviews with acoustic, visual, and transcript features. E-DAIC (Ringeval et al., 2019) provides 275 interviews with eGeMAPS features (all test sessions automated). D-Vlog (Yoon et al., 2022) contains 1,000 video blogs with acoustic features and facial landmarks, without transcripts. Labels follow PHQ-8 $\geq$ 10 on interview corpora and released labels on D-Vlog. Each corpus is trained separately with outer-fold reference fitting. Following evaluation protocols in Appendix I, we evaluate DAIC-WOZ on official dev, and E-DAIC and D-Vlog on official test. Decisions use threshold 0.5, except the D-Vlog first pass which tunes threshold on validation data; due to higher prevalence on D-Vlog, F1 is compared within each dataset.

Metrics. We report AUROC and Brier score for ranking and calibration, respectively, and precision, recall, and F1 for decisions. For the loop, let N be the number of evaluated cases, $\mathcal{A}$ the set of accepted (non-referred) cases, and $N_{w\to c}$ and $N_{c\to w}$ the numbers of accepted cases that the loop

Table 1: Comparison with published methods. Modalities: A, audio; V, visual; T, transcript. Coverage is the fraction of cases decided without referral; RETHINK-MH is scored on the cases it accepts. †Five-fold cross-validation instead of the official partitions. ‡Custom evaluation protocol. §Mean of five runs. **Bold**: best; underline: second best.

Dataset	Method	Modality	Coverage	Precision	Recall	F1
DAIC-WOZ	Gimeno-Gómez et al. (2024)	V	1.000	0.660	0.610	0.620
	Gimeno-Gómez et al. (2024)	AV	1.000	0.680	0.660	0.670
	VTA-DepressNet† (Nguyen et al., 2026)	AVT	1.000	–	–	**0.830**
	MSCDV‡ (Han et al., 2026)	AVT	1.000	0.800	**0.860**	**0.830**
	RETHINK-MH first pass (Ours)	AVT	1.000	0.727	0.667	0.696
	RETHINK-MH (Ours)	AVT	0.857	**0.818**	0.750	0.783
E-DAIC	Gimeno-Gómez et al. (2024)	V	1.000	0.580	0.490	0.500
	Gimeno-Gómez et al. (2024)	AV	1.000	0.590	0.580	0.560
	VTA-DepressNet† (Nguyen et al., 2026)	AT	1.000	–	–	0.770
	MSCDV‡ (Han et al., 2026)	AVT	1.000	0.680	**0.880**	0.770
	MLlm-DR§ (Zhang et al., 2026)	AVT	1.000	0.770	0.810	0.688
	RETHINK-MH first pass (Ours)	AVT	1.000	0.552	0.762	0.640
	RETHINK-MH (Ours)	AVT	0.875	**0.800**	0.800	**0.800**
D-Vlog	Depressformer (He et al., 2024)	AV	1.000	0.640	0.540	0.590
	CAIINET (Zhou et al., 2023)	AV	1.000	0.666	0.670	0.666
	Attention-MIL (Shangguan et al., 2023)	AV	1.000	0.673	0.678	0.666
	DepMSTAT (Tao et al., 2024a)	AV	1.000	0.715	0.756	0.735
	STST (Tao et al., 2024b)	AV	1.000	0.725	0.777	0.750
	MDAVIF (Ling et al., 2024)	AV	1.000	0.743	0.760	0.753
	DepMamba (Ye et al., 2025)	AV	1.000	0.682	0.870	0.764
	STCM-Mamba (Zhou et al., 2025)	AV	1.000	0.737	0.821	0.777
	BHDD (Ma et al., 2026)	AV	1.000	**0.760**	0.760	0.759
	DepMPK (Yang et al., 2026)	AV	1.000	0.708	**0.908**	**0.796**
	RETHINK-MH first pass (Ours)	AV	1.000	0.664	0.805	0.728
	RETHINK-MH (Ours)	AV	0.849	0.675	0.830	0.744

turns from wrong to correct and from correct to wrong, respectively. We report

$$\mathrm{NetCorr} = \frac{N_{w\to c} - N_{c\to w}}{N}, \quad \mathrm{Coverage} = \frac{|\mathcal{A}|}{N}, \quad \mathrm{FNR}_{\mathcal{A}} = \frac{|\{i \in \mathcal{A} : y_i = 1, \hat{y}_i = 0\}|}{|\{i \in \mathcal{A} : y_i = 1\}|}, \quad (9)$$

where $\mathrm{FNR}_{\mathcal{A}}$, the false-negative rate on accepted cases, captures the costliest screening error.

4.2 Implementation Details

We use Qwen2.5-Omni-7B (Xu et al., 2025) with one adapter per outer fold across $K = 5$ folds and greedy decoding with constrained JSON output. At test time, all five fold adapters process each case. A case is referred if at least three adapters refer it; otherwise, its risk is the mean of their final probabilities. The audit flags about one third of the cases on each corpus, and only these cases run the query, selection, action, and revision turns, so the loop costs far less than a second pass on every case. Appendix C details the training schedules, optimizer choices, and hyperparameters.

4.3 Comparison with Existing Methods

Table 1 compares RETHINK-MH with methods published since 2023. On E-DAIC, the full loop achieves the highest F1 of 0.800, narrowly ahead of the mean F1 0.688 of MLlm-DR. On DAIC-WOZ, VTA-DepressNet and MSCDV both report 0.830, above our 0.783, but neither evaluates on the development partition we use: VTA-DepressNet pools the corpus under five-fold cross-validation, and MSCDV scores the official test participants with the custom utterance-level protocol described in the caption. On D-Vlog, which lacks transcripts, RETHINK-MH trails the best published result by DepMPK by 0.052 in F1 but provides cited evidence and referral. Because RETHINK-MH refers between 12% and 15% of uncertain cases, we verify that its gain does not stem from referral alone. When the first pass naively refers its least confident predictions at matched coverage, it attains F1 of only 0.706 on DAIC-WOZ, 0.667 on E-DAIC, and 0.731 on D-Vlog, compared with full-coverage baselines of 0.696, 0.640, and 0.728. In contrast, RETHINK-MH achieves 0.783, 0.800, and 0.744 on these benchmarks, demonstrating that the improvement stems from targeted evidence retrieval rather than coverage shrinkage, as further analyzed in Appendix D.1.

Table 2: Closed-loop analysis. Direct-refer refers every flagged case; no-new-evidence rethinking re-prompts without retrieval. Metrics follow Eq. 9; arrows mark the better direction.

Dataset	System	AUROC ↑	Brier ↓	$N_{w\to c}$ / $N_{c\to w}$	Net correction ↑	Coverage ↑	$\text{FNR}_{\mathcal{A}}$ ↓
DAIC-WOZ	First pass	0.826	0.150	–	–	1.000	0.333
	Direct-refer	–	–	0 / 0	0.000	0.714	0.273
	No-new-evidence rethinking	0.828	0.149	1 / 1	0.000	0.971	0.333
	RETHINK-MH	0.842	0.141	3 / 1	+0.057	0.857	0.250
E-DAIC	First pass	0.772	0.178	–	–	1.000	0.238
	Direct-refer	–	–	0 / 0	0.000	0.661	0.211
	No-new-evidence rethinking	0.775	0.176	2 / 2	0.000	0.964	0.250
	RETHINK-MH	0.791	0.170	11 / 1	+0.179	0.875	0.200
D-Vlog	First pass	0.676	0.222	–	–	1.000	0.195
	Direct-refer	–	–	0 / 0	0.000	0.698	0.190
	No-new-evidence rethinking	0.680	0.220	2 / 2	0.000	0.948	0.232
	RETHINK-MH	0.694	0.211	6 / 2	+0.019	0.849	0.170

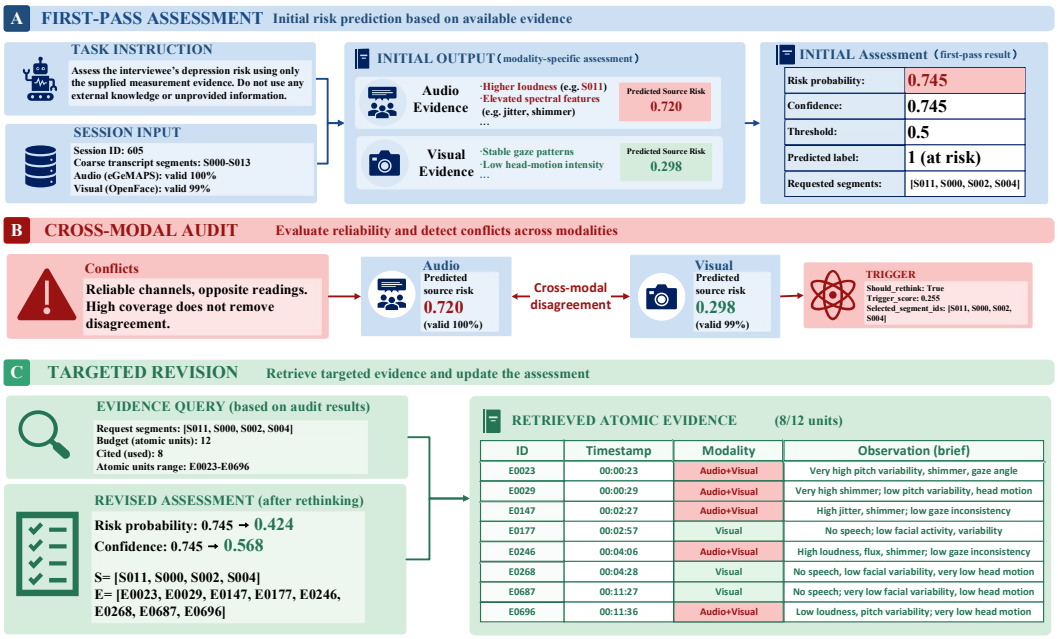

Figure 3: Case study of E-DAIC test session 605 with a negative label. (A) The first pass predicts a risk of 0.745, a false positive. (B) The audit flags the case. (C) The Rethinker cites eight retrieved atomic units and lowers the risk to 0.424, correcting the decision.

4.4 Closed-Loop Analysis

Table 2 compares the loop with the frozen first pass and with two controls sharing its trigger. The loop corrects more errors than it introduces on all three corpora. The gain is significant on E-DAIC under an exact McNemar test with $p = 0.006$, while smaller gains on DAIC-WOZ and D-Vlog do not reach significance at these sample sizes, as reported in Appendix D.7. As Proposition 1 predicts, rethinking without new evidence breaks as many decisions as it fixes on every corpus. Compared with referring every flagged case, the loop decides more cases automatically and misses fewer positives among them, so cases it reclaims are decided reliably rather than merely accepted. The first pass itself matches or exceeds numeric and flat-text baselines, as established in Appendix D.2.

The decisions are also verifiable. Every cited identifier lies in the retrieved packet, and deleting cited units alters 81.5% of revisions compared with only 4.8% when deleting uncited units, as shown in Appendix F. Frontier audits with Gemini-3.1-Pro and GPT-5.6-Sol confirm an 88.3% factual support rate for revision statements with substantial agreement at Cohen's $\kappa = 0.764$, and 97.2% of outputs strictly conform to the JSON schema. Under injected tracking loss up to 30% and acoustic noise down to 0 dB SNR, the loop fails closed: referral rates rise monotonically from 15% to 45% while accepted false negatives remain flat, erring toward human review as detailed in Appendix D.6.

Table 3: Ablation on E-DAIC. Each row adds one component; the last row is the full loop.

Configuration	AUROC	Brier	F1	$N_{w\to c}$ / $N_{c\to w}$	Net correction	Coverage	$\text{FNR}_{\mathcal{A}}$
First pass	0.772	0.178	0.640	–	–	1.000	0.238
+ Audit	–	–	–	0 / 0	0.000	0.661	0.211
+ Retrieval	0.780	0.172	0.727	7 / 1	+0.107	1.000	0.238
+ SFT	0.784	0.170	0.762	9 / 1	+0.143	0.875	0.238
+ Align	0.791	0.170	0.800	11 / 1	+0.179	0.875	0.200

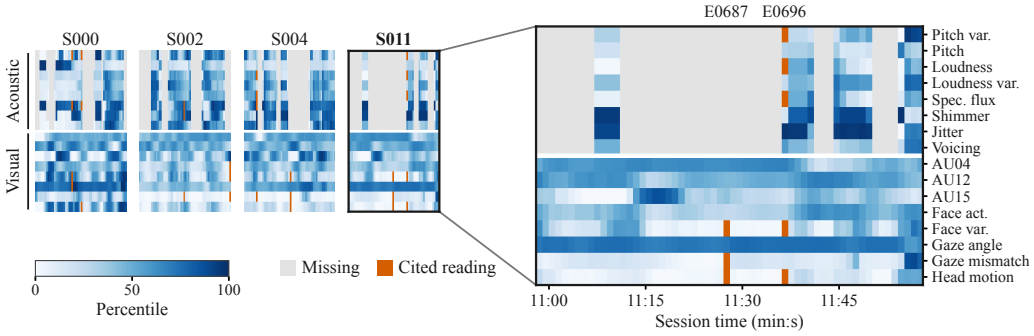

Figure 4: Evidence map for the case in Fig. 3. Left: the four retrieved one-minute segments. Right: segment S011 enlarged. Cells show percentiles; gray: missing data; orange: cited units.

4.5 Ablations and Case Study

Table 3 adds one component at a time on E-DAIC. The audit alone can only refer; retrieval with a fixed policy yields the first net correction and the largest F1 gain; the two SFT stages add referral; and preference alignment repairs two more cases. Two further controls test the conditions of Sec. 3.5. Removing validity weighting lowers net correction, consistent with Proposition 3, and replacing the Query Planner with random or earliest-segment retrieval under the same budget removes most of the gain, consistent with Proposition 2. In line with Proposition 3, the validity-weighted audit consistently outperforms first-pass entropy gating at predicting natural errors across all three corpora, achieving error AUROCs of 0.66 versus 0.61 on DAIC-WOZ, 0.70 versus 0.69 on E-DAIC, and 0.62 versus 0.58 on D-Vlog. Furthermore, the audit reliably detects synthetic corruptions with AUROC between 0.88 and 0.93, as documented in Appendix D.4.

Fig. 3 and Fig. 4 trace E-DAIC test session 605, which has a true negative label, end to end. The first pass predicts a false-positive risk of 0.745 due to elevated vocal features in segment summaries. Because the visual channel indicates normal engagement with a lower predicted risk of 0.298, cross-modal disagreement triggers targeted retrieval spanning segments S011, S000, S002, and S004. The Rethinker inspects the atomic layer and cites eight units: visual units confirm intact gaze contact and facial dynamics including action units 4 and 12, while acoustic units reveal that elevated pitch occurred during conversational responsiveness rather than depressive agitation. Consequently, four of the five fold adapters lower the risk to 0.424 and raise confidence to 0.568, correcting the decision through verified evidence that Proposition 2 requires.

5 Conclusion

We presented RETHINK-MH, a reliability-triggered rethinking framework turning automated screening into an auditable two-pass process. By structuring measurements into an evidence store, auditing predictions via validity-weighted cross-source disagreement, and selectively retrieving atomic evidence under an explicit budget, RETHINK-MH achieves competitive screening across three diverse benchmarks. Crucially, the loop corrects more errors than it introduces, grounds revisions in verifiable behavioral observations, and refers unresolved cases to clinicians. Evidence withholding, conflict-driven auditing, and constrained deliberation thus provide a transparent, fail-closed foundation for trustworthy multimodal healthcare screening.

Reproducibility Statement

Evidence compilation is specified in Sec. 3.2, including the fitting scope of the reference distribution and the construction and rendering rules of the two evidence layers; fold definitions and cross-fitting relations are pinned per fold by configuration manifests. Training schedules, learning rates, preference-pair construction, and decoding settings are given in Sec. 3.4 and Appendix C, metric definitions in Sec. 4.1, proofs of the propositions in Appendix B, and training seeds are fixed at 42, 43, and 44. Under the corpus access agreements of DAIC-WOZ, E-DAIC, and D-Vlog, we release the evidence-compilation specification, training and evaluation configurations, fold definitions, and evaluation scripts; model predictions and derived metrics are released to the extent the agreements permit, and no raw recordings or identifiable information are released.

Ethics Statement

This system is intended for questionnaire-based mental health screening and does not constitute a clinical diagnosis, nor does it replace professional assessment. The three corpora are used under their respective data access agreements, and we process only de-identified released features, processing and redistributing no raw audio or video. Screening errors are asymmetric in cost, with false negatives the primary concern, so the system always refers cases it cannot resolve to human review, and the intended deployment is human-in-the-loop rather than fully automatic. Any prospective deployment should re-evaluate the system's trigger rate, referral burden, and error structure in the real workflow, and users should be informed that the output is a checkable second opinion rather than a diagnosis.

AI Use Statement

Large language models were used for language polishing of the manuscript and for coding assistance during development. All research ideas, experimental designs, analyses, and results are the authors' own work, every AI-assisted output was reviewed and verified by the authors, and the authors take full responsibility for the content of this paper.

References

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.*, 16(4):494–591, 2023. ISSN 1935-8237.

Sergio Burdisso, Ernesto Reyes-Ramírez, Esaú Villatoro-Tello, Fernando Sánchez-Vega, A. Pastor López-Monroy, and Petr Motlicek. DAIC-WOZ: On the validity of using the therapist's prompts in automatic depression detection from clinical interviews. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pp. 82–90, 2024.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 16477–16508, 2023a.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6465–6488, 2023b.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning (ICML)*, pp. 2151–2159, 2019.

David Gimeno-Gómez, Ana-Maria Bucur, Adrian Cosma, Carlos-D. Martínez-Hinarejos, and Paolo Rosso. Reading between the frames: Multi-modal depression detection in videos from non-verbal cues. In *European Conference on Information Retrieval (ECIR)*, pp. 191–209, 2024.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations (ICLR)*, volume 2024, pp. 57734–57811, 2024.

Jonathan Gratch, Ron Artstein, Gale M. Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Albert Rizzo, and Louis-Philippe Morency. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pp. 3123–3128, 2014.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

Dongfang Han, Guo-Xing Xiang, Jingyu Zhu, Yuanyuan Liao, Jihong Zhu, Askar Hamdulla, and Turdi Tohti. A multi-scale transformer framework with consistency and dual-view for depression detection. *Biomedical Signal Processing and Control*, 112:108461, 2026.

Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9180–9192, 2021.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1122–1131, 2020.

Lang He, Zheng Li, Prayag Tiwari, Cui Cao, Jize Xue, Feng Zhu, and Di Wu. Depressformer: Leveraging video swin transformer and fine-grained local features for depression scale estimation. *Biomedical Signal Processing and Control*, 96:106490, 2024.

Lang He, Kai Chen, Junnan Zhao, Yimeng Wang, Ercheng Pei, Haifeng Chen, Jiewei Jiang, Shiqing Zhang, Jie Zhang, Zhongmin Wang, Tao He, and Prayag Tiwari. LMVD: A large-scale multi-modal vlog dataset for depression detection in the wild. *Information Fusion*, 126:103632, 2026.

Jiwoo Hong, Noah Lee, and James Thorne. ORPO: Monolithic preference optimization without reference model. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 11170–11189, 2024.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations (ICLR)*, volume 2024, pp. 32808–32824, 2024.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7969–7992, 2023.

Kurt Kroenke, Tara W. Strine, Robert L. Spitzer, Janet B. W. Williams, Joyce T. Berry, and Ali H. Mokdad. The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, 114(1-3):163–173, 2009.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 9459–9474, 2020.

Yang Li and Junier Oliva. Active feature acquisition with generative surrogate models. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6450–6459, 2021.

Tianfei Ling, Deyuan Chen, and Baobin Li. MDAVIF: A multi-domain acoustical-visual information fusion model for depression recognition from vlog data. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8115–8119, 2024.

Mengyuan Ma, Chengji Wang, Sijia Liu, Weizhong Zhao, and Xingpeng Jiang. BHDD: Facial behaviors modeling and hierarchical multimodal fusion for depression detection from vlog data. *IEEE Transactions on Affective Computing*, 17(3):2714–2728, 2026.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, and Yiming Yang. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 46534–46594, 2023.

David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning (ICML)*, pp. 7076–7087, 2020.

Nguyen Nguyen, Minh Nguyen, Tri Doan, Hau Vo, Nha Tran, and Hung Nguyen. VTA-DepressNet: A dual-level attention fusion for multimodal depression detection. *CCF Transactions on Pervasive Computing and Interaction*, pp. 1–20, 2026. doi: 10.1007/s42486-026-00239-x.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 27730–27744, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 53728–53741, 2023.

Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maja Pantic, Maximilian Schmitt, Shahin Amiriparian, Eva-Maria Messner, et al. AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition. In *Proceedings of the 9th International Audio/Visual Emotion Challenge and Workshop*, pp. 3–12, 2019.

Misha Sadeghi, Robert Richer, Bernhard Egger, Lena Schindler-Gmelch, Luisa Rupp, Farnaz Rahimi, Matthias Berking, and Björn Eskofier. Harnessing multimodal approaches for depression detection using large language models and facial expressions. *npj Mental Health Research*, 3(1):66, 2024. doi: 10.1038/s44184-024-00112-8.

Zixuan Shangguan, Xiaxi Li, Yanjie Dong, and Xiaoyan Yuan. Automatic depression detection using attention-based deep multiple instance learning. In *International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness*, pp. 40–51, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Hajin Shim, Sung Ju Hwang, and Eunho Yang. Joint active feature acquisition and classification with variable-size set encoding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 8634–8652, 2023.

Siyang Song, Linlin Shen, and Michel Valstar. Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features. In *13th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pp. 158–165, 2018.

Yongfeng Tao, Minqiang Yang, Huiru Li, Yushan Wu, and Bin Hu. DepMSTAT: Multimodal spatio-temporal attentional transformer for depression detection. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):2956–2966, 2024a.

Yongfeng Tao, Minqiang Yang, Yushan Wu, Kevin Lee, Adrienne Kline, and Bin Hu. Depressive semantic awareness from vlog facial and vocal streams via spatio-temporal transformer. *Digital Communications and Networks*, 10(3):577–585, 2024b.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6558–6569, 2019.

Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pp. 3–10, 2016.

Samantha Weber, Nicolas Deperrois, Robert Heun, Laura Frühschütz, Anna Monn, Stephanie Homan, Alina Häfliger, Erich Seifritz, Tobias Kowatsch, Lena A. Jäger, et al. Using a fine-tuned large language model for symptom-based depression evaluation. *npj Digital Medicine*, 8(1):598, 2025. doi: 10.1038/s41746-025-01982-8.

Ping-Cheng Wei, Kunyu Peng, Alina Roitberg, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhagen. Multi-modal depression estimation based on sub-attentional fusion. In *European Conference on Computer Vision (ECCV) Workshops*, pp. 623–639, 2022.

Yuhua Wen, Yingying Zhou, Qifei Li, Yingming Gao, Zhengqi Wen, Jianhua Tao, and Ya Li. MIDAS: Mutual information disentanglement with uncertainty-aware fusion for incomplete multimodal sentiment analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. doi: 10.1109/TPAMI.2026.3713694.

James R. Williamson, Elizabeth Godoy, Miriam Cha, Adrianne Schwarzentruber, Pooya Khorrami, Youngjune Gwon, Hsiang-Tsung Kung, Charlie Dagli, and Thomas F. Quatieri. Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pp. 11–18, 2016.

Zhenyu Wu, Qingkai Zeng, Zhihan Zhang, Zhaoxuan Tan, Chao Shen, and Meng Jiang. Large language models can self-correct with key condition verification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 12846–12867, 2024.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Yunfei Chu, et al. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Beiying Yang, Lu Zhou, Haixin Wang, Yingying Chen, Jinqiao Wang, and Kehua Su. DepMPK: Multi-perspective knowledge fusion for multimodal depression detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12582–12586, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Jiaxin Ye, Junping Zhang, and Hongming Shan. Depmamba: Progressive fusion mamba for multimodal depression detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025.

Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. D-Vlog: Multimodal vlog dataset for depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 12226–12234, 2022.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*, 2023.

Wei Zhang, Juan Chen, En Zhu, Wenhong Cheng, Yunpeng Li, Yuhan Li, and Yanbo J. Wang. MLlm-DR: Towards explainable depression recognition with multimodal large language models. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 22(4):111:1–111:23, 2026.

Bowen Zhou, Marc-André Fiedler, and Ayoub Al-Hamadi. STCM-Mamba: Multimodal spatio-temporal cross-modal mamba for depression detection. *IEEE Access*, 13:189494–189505, 2025.

Li Zhou, Zhenyu Liu, Xiaoyan Yuan, Zixuan Shangguan, Yutong Li, and Bin Hu. CAIINET: Neural network based on contextual attention and information interaction mechanism for depression detection. *Digital Signal Processing*, 137:103986, 2023.

Appendix

RETHINK-MH: Reliability-Triggered Multimodal Rethinking for Mental Health Screening

A Extended Related Work

Multimodal Mental Health Screening. Depression screening from clinical interviews is supported by a series of multimodal corpora (Gratch et al., 2014; Valstar et al., 2016; Ringeval et al., 2019; Yoon et al., 2022; He et al., 2026). Early methods analyze hand-crafted statistics and spectral features (Song et al., 2018; Williamson et al., 2016), and later work moves to learned fusion structures, including sub-attentional fusion (Wei et al., 2022) and video modeling of non-verbal cues (Gimeno-Gómez et al., 2024; Yang et al., 2026), as well as general multimodal fusion machinery such as cross-modal Transformers (Tsai et al., 2019), modality-invariant and modality-specific subspaces (Hazarika et al., 2020), hierarchical mutual-information maximization (Han et al., 2021), spatiotemporal and multi-scale attention (Tao et al., 2024a; Nguyen et al., 2026; Han et al., 2026), and Mamba-based sequence models (Ye et al., 2025; Zhou et al., 2025; Ma et al., 2026). Large language models have recently entered screening as well, extracting questionnaire-related indicators from interview text (Sadeghi et al., 2024; Weber et al., 2025) and scoring interviews with rationales (Zhang et al., 2026). At the same time, reproduction studies document shortcuts on these benchmarks, where the reported performance of some published models is inflated by interviewer-prompt leakage (Burdisso et al., 2024). These systems all map features to a score in a single forward pass, so source quality and cross-source conflict are absorbed into the representation rather than acted upon, whereas we attach an auditable second pass to the judgment.

Reliability, Calibration, and Selective Prediction. Probabilistic predictions are evaluated against proper scoring rules (Gneiting & Raftery, 2007), modern networks produce systematically biased confidence (Guo et al., 2017), and deep ensembles offer a pragmatic uncertainty estimate (Lakshminarayanan et al., 2017). Selective prediction formalizes the risk–coverage trade-off and defers low-confidence cases to humans (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; 2019), conformal prediction supplies distribution-free coverage guarantees (Angelopoulos & Bates, 2023), and learning-to-defer methods train the routing together with the model (Madras et al., 2018; Mozannar & Sontag, 2020). For generative models, semantic entropy detects unreliable generations (Farquhar et al., 2024), and in incomplete multimodal learning, channel reliability already acts as a fusion weight (Wen et al., 2026). Active feature acquisition purchases further inputs before predicting (Shim et al., 2018; Li & Oliva, 2021), but over fixed feature vectors and without a record of how the acquired values changed the decision. These methods decide whether to answer at the current information state, whereas we use the same reliability signals to decide whether to acquire more evidence first, leaving referral as the terminal fallback rather than the only option.

Test-Time Revision and Retrieval. Self-correction methods let a model iteratively refine its own output (Madaan et al., 2023; Shinn et al., 2023), and reasoning–acting agents interleave generation with tool calls (Yao et al., 2023). Without external signals, such reflection does not reliably correct reasoning (Huang et al., 2024), and it becomes effective when verifiable conditions or external feedback are introduced (Wu et al., 2024; Gou et al., 2024). Retrieval-augmented generation requeries external corpora before generation (Lewis et al., 2020), and active variants decide when and what to retrieve (Jiang et al., 2023), while another line makes long-form answers carry verifiable citations (Gao et al., 2023b;a). In mental health, interview signals have also been transcribed into text for large language models to judge (Sadeghi et al., 2024). Our second pass instead re-queries an indexed evidence store compiled from the same session under a budget, and every citation is programmatically checked against what was retrieved.

Preference Alignment and Post-Training. Post-training aligns a pre-trained model with human or synthetic preferences, from reinforcement learning from human feedback (Ouyang et al., 2022) to reference-free variants such as direct preference optimization (Rafailov et al., 2023), monolithic preference optimization (Hong et al., 2024), and group-relative policy optimization (Shao et al., 2024), with rejection sampling turning a model's own filtered generations into training pairs (Yuan et al., 2023). Most of these methods rely on an online or semi-online rollout–reward loop. Our

post-training instead runs on a frozen offline replay, where rollouts produced by the current policy on frozen prompts are deterministically filtered before entering the preference pairs, so no online policy-gradient update is needed.

B PROOFS OF THE PROPOSITIONS

Throughout, Y is the binary screening label, all observables are random variables on a common probability space, predictors are measurable, and the loss ℓ is nonnegative. For an observable X, $L^\star_\ell(X) = \inf_f \mathbb{E}[\ell(Y, f(X))]$ is the Bayes risk of Sec. 3.5.

Proof of Proposition 1. Let g be any predictor of Y from (A^0, A^1) and define $\tilde{g}(v, u) = g\big(f_0(v, u),\ f_1(v, f_0(v, u), u)\big)$, which is measurable and satisfies $g(A^0, A^1) = \tilde{g}(\mathcal{V}^0, U)$. Because U is independent of $(Y, \mathcal{V}^0)$, the joint law of $(Y, \mathcal{V}^0, U)$ is the product of the law of $(Y, \mathcal{V}^0)$ and the law P_U of U, and Tonelli's theorem gives

$$\mathbb{E}\big[\ell\big(Y, g(A^0, A^1)\big)\big] = \int \mathbb{E}\big[\ell\big(Y, \tilde{g}(\mathcal{V}^0, u)\big)\big]\, dP_U(u).$$

For every fixed u, the map $v \mapsto \tilde{g}(v, u)$ is a predictor of Y from $\mathcal{V}^0$, so the integrand is at least $L^\star_\ell(\mathcal{V}^0)$, and so is the integral. Taking the infimum over g yields $L^\star_\ell(A^0, A^1) \geq L^\star_\ell(\mathcal{V}^0)$. □

Greedy decoding corresponds to a constant U, and sampling, self-consistency voting, or any number of reflection rounds are covered by a richer U, since each round is again a function of $\mathcal{V}^0$, the earlier outputs, and fresh independent randomness. The improvement of any such re-prompting scheme over the first pass is therefore bounded by the excess risk of the first pass over the Bayes rule on $\mathcal{V}^0$.

Proof of Proposition 2. Under log loss a predictor outputs a distribution $Q(\cdot \mid X)$ over the label and incurs $-\log Q(Y \mid X)$. Writing $P(\cdot \mid X)$ for the true posterior,

$$\begin{aligned}\mathbb{E}\big[-\log Q(Y \mid X)\big] &= \mathbb{E}\big[-\log P(Y \mid X)\big] + \mathbb{E}\Big[\log \tfrac{P(Y|X)}{Q(Y|X)}\Big] \\ &= H(Y \mid X) + \mathbb{E}_X\big[\mathrm{KL}\big(P(\cdot \mid X) \,\|\, Q(\cdot \mid X)\big)\big].\end{aligned}$$

The divergence is nonnegative and vanishes exactly when Q equals the posterior, so $L^\star_{\log}(X) = H(Y \mid X)$, which is finite because Y is binary. Applying this to $X = \mathcal{V}^0$ and to $X = (\mathcal{V}^0, \Delta\mathcal{E}(q))$ gives $L^\star_{\log}(\mathcal{V}^0) - L^\star_{\log}(\mathcal{V}^0, \Delta\mathcal{E}(q)) = H(Y \mid \mathcal{V}^0) - H(Y \mid \mathcal{V}^0, \Delta\mathcal{E}(q)) = I(Y; \Delta\mathcal{E}(q) \mid \mathcal{V}^0)$, which is nonnegative as a conditional mutual information. Since $L^\star_{\log}(\mathcal{V}^0)$ does not depend on q, minimizing the second-pass Bayes risk over queries with $|R(q)| \leq B$ is equivalent to maximizing $I(Y; \Delta\mathcal{E}(q) \mid \mathcal{V}^0)$. □

The Query Planner chooses q as a function of the first-pass view and its assessment. Conditional on $\mathcal{V}^0 = v$ the query is then fixed, and the same argument shows that the ideal policy picks, for each v, a feasible query that maximizes the expected reduction of $H(Y \mid \mathcal{V}^0 = v)$ by $\Delta\mathcal{E}(q)$. If the retrieved units are a function of $\mathcal{V}^0$, the conditional mutual information is zero, which formalizes why atomic detail already implied by its segment summary cannot help.

Proof of Proposition 3. Draw the N evaluation cases independently from the evaluation distribution, and for case i let W_i be the event that it is accepted with $\hat{y}^0_i \neq y_i$ and $\hat{y}_i = y_i$, and H_i the event that it is accepted with $\hat{y}^0_i = y_i$ and $\hat{y}_i \neq y_i$. Then $N_{w\to c} = \sum_i \mathbf{1}[W_i]$ and $N_{c\to w} = \sum_i \mathbf{1}[H_i]$, and linearity of expectation gives $\mathbb{E}[\mathrm{NetCorr}] = P(W) - P(H)$ for a generic case. When $z = 0$, Eq. 4 returns $\hat{y} = \mathbf{1}[p^0 \geq \tau] = \hat{y}^0$, so neither event can occur, and referred cases are not accepted. Both events therefore require $z = 1$, and by the chain rule

$$\begin{aligned}P(W) &= P(z = 1)\, P(\hat{y}^0 \neq Y \mid z = 1)\, P(W \mid \hat{y}^0 \neq Y, z = 1) = \pi_T\, \varepsilon_T\, \kappa_r, \\ P(H) &= P(z = 1)\, P(\hat{y}^0 = Y \mid z = 1)\, P(H \mid \hat{y}^0 = Y, z = 1) = \pi_T\, (1 - \varepsilon_T)\, \kappa_h.\end{aligned}$$

Subtracting gives the identity of Eq. 8, and the upper bound follows from $\kappa_r \leq 1$ and $\kappa_h(1 - \varepsilon_T) \geq 0$. For $\pi_T > 0$, $\kappa_r > 0$, and $\varepsilon_T < 1$, the expectation is positive exactly when $\kappa_r \varepsilon_T > \kappa_h(1 - \varepsilon_T)$, that is, when $\varepsilon_T/(1 - \varepsilon_T) > \kappa_h/\kappa_r$. □

Direct-refer sets $\kappa_r = \kappa_h = 0$ and thus has zero expected net correction, matching its $0/0$ transitions in Table 2. At a fixed trigger rate, two triggers can differ only through the share of errors they select and through the repair-to-harm ratio that the downstream Rethinker achieves on the cases they select.

C Extended Implementation Details

All experiments use Qwen2.5-Omni-7B (Xu et al., 2025) as the backbone and train a separate adapter in each outer fold. The first pass sees at most 24 segments, and a re-query returns at most 12 atomic units. Cross-fitting uses $K = 5$ outer folds, with the reference refit in every fold on training participants only. The optimizer is AdamW. The cold-start schedule runs Action-SFT for 3 epochs and Revision-SFT for 2 epochs. The post-training schedule runs initial-pair alignment for 2 epochs at learning rate 2×10^{-5}, then rollout refresh and refreshed-pair alignment for 1 epoch each at 1×10^{-5}, with $\beta = 0.10$ in both rounds. Uncertainty intervals use 2,000 participant-level, label-stratified bootstrap resamples.

The audit of Sec. 3.3 is an L_2-regularized logistic regression on standardized features. The validity of each source is the rescaled out-of-fold AUROC of its risk readings on the training folds. The audit is fit on the out-of-fold initial assessments of the training folds, with the target of whether the first-pass decision was wrong, together with synthetic degradations whose targets are known, and its threshold is the quantile of development-set risks that gives the trigger rate in Table A2, which needs no development labels. At test time, a case is referred when at least three of the five fold adapters refer it, and otherwise its risk is the mean of their final probabilities.

The refreshed pairs of Sec. 3.4 are built as follows. The current policy is rolled out once on the frozen fit prompts, and generations are screened by three deterministic filters in sequence. The outcome filter requires the terminal decision to be correct and compatible with the target transition, the schema filter requires all fields, value ranges, and status consistency to be valid, and the grounding filter requires all cited IDs to lie within the participant's evidence allowlist. Generations passing all three filters become the new chosen completions, and the grounded but wrong alternatives remain as rejected completions. Contract guards, which are rejected outputs with missing fields, unknown evidence IDs, or out-of-scope clinical content, are rebuilt at the original sparse ratio, yielding a preference set closer to the current policy distribution without crowding out hard negatives.

The annotator of Sec. 3.4 marks a case as keep when $\mathbf{1}[p^0 \geq \tau] = y$. It marks a case as revise, in the direction of $p^* - p^0$, when $\mathbf{1}[p^* \geq \tau] = y$, the change lowers both the Brier score and the log loss, and $|p^* - p^0| \geq 0.05$. All other cases are marked as refer.

D Additional Empirical Results

D.1 Closed-Loop Gains

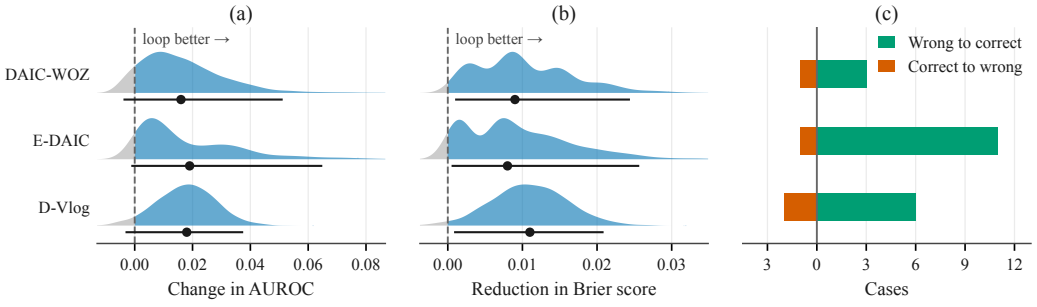

Figure A1: Gains of the loop over the first pass. (a) Change in AUROC and (b) reduction in Brier score over paired bootstrap resamples, blue where the loop is better, with the point estimate and 95% interval below. (c) Cases turned from wrong to correct and from correct to wrong.

In Fig. A1, we show the distributions behind Table 2. On all three corpora, about 95% of bootstrap resamples favor the loop in AUROC and nearly all favor it in Brier score, although the 95% intervals for AUROC extend slightly below zero, and repairs outnumber harms on each of them.

Coverage-Matched Control. RETHINK-MH refers some cases and is scored on the rest, so part of its F1 gain could come from referral alone. To test this, we let the first pass refer its least confident cases, without any rethinking, until its coverage equals that of RETHINK-MH. This control reaches F1 of 0.706, 0.667, and 0.731 on DAIC-WOZ, E-DAIC, and D-Vlog, against 0.783, 0.800, and 0.744 for RETHINK-MH and 0.696, 0.640, and 0.728 for the first pass at full coverage, so most of the gain comes from rethinking rather than from referral.

D.2 Representation Controls

We test whether the text-based two-layer store is a good interface at all, by comparing it with a numeric-feature classifier on the same measurements and folds and with a flat single-layer rendering under a matched token budget.

Table A1: Representation controls for the first pass, on the cross-fitted training folds of DAIC-WOZ and E-DAIC and the test partition of D-Vlog. Numeric features feed per-session statistics to a logistic regression, and the flat rendering puts all measurements in one text layer. AP is average precision.

Dataset	Representation	AUROC	AP	Brier	Median tokens
DAIC-WOZ	Numeric features	0.742	0.588	0.171	–
	Flat single-layer text	0.761	0.605	0.166	5.8k
	Two-layer store	0.777	0.631	0.162	6.2k
E-DAIC	Numeric features	0.748	0.590	0.184	–
	Flat single-layer text	0.755	0.598	0.181	4.1k
	Two-layer store	0.772	0.612	0.178	4.3k
D-Vlog	Numeric features	0.661	0.730	0.230	–
	Flat single-layer text	0.668	0.739	0.226	2.7k
	Two-layer store	0.676	0.746	0.222	2.9k

In Table A1, the two-layer store matches or beats both controls on every corpus at a similar token count, so the loop does not rest on a weak first pass.

D.3 Ablation Ladder

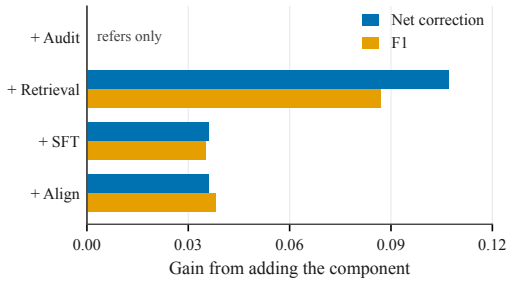

Figure A2: Gain in net correction and F1 from adding each component of Table 3.

Fig. A2 shows the gains behind Table 3, where we add one component at a time. The trigger alone only refers cases, retrieval brings most of the net correction, the two SFT stages add referral and a valid output format, and alignment brings most of the gain in F1.

D.4 Audit Quality

The audit serves two purposes that we evaluate separately, namely detecting genuinely degraded inputs and predicting natural first-pass errors.

In Table A2 and Fig. A3, the audit separates degraded from clean sessions well, but natural first-pass errors are much harder to predict, for the audit and for the entropy of the first-pass probability alike.

D.5 Trigger and Retrieval Controls

We test the two components of the loop with matched controls on the E-DAIC test partition.

Table A2: Audit quality. Corruption AUROC is for detecting synthetic degradations, and error AUROC is for predicting natural first-pass errors, by the audit and by the entropy of the first-pass probability. Δ is the difference in error AUROC.

Dataset	Corruption AUROC	Error AUROC, audit	Error AUROC, entropy	Δ	95% CI	Trigger rate
DAIC-WOZ	0.91	0.66	0.61	+0.05	[0.01, 0.09]	0.29
E-DAIC	0.93	0.70	0.69	+0.01	[-0.03, 0.05]	0.34
D-Vlog	0.88	0.62	0.58	+0.04	[0.01, 0.08]	0.30

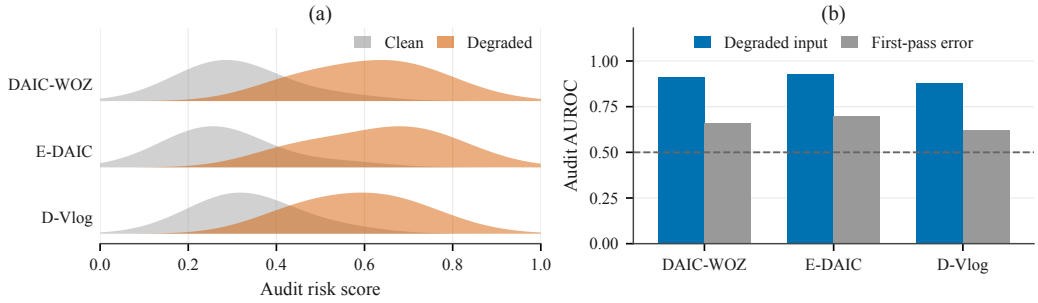

Figure A3: Audit quality. (a) Audit risk scores of clean and synthetically degraded sessions. (b) AUROC of the audit for detecting degraded inputs and for predicting first-pass errors, with the dashed line at chance.

Trigger Controls. At the matched trigger rate of 0.34, triggering by normalized entropy alone attains a net correction of $+0.089$ and $\mathrm{FNR}_{\mathcal{A}}$ of 0.250, against $+0.179$ and 0.200 for the full audit. Removing validity weighting by setting $v_s \equiv 1$ lowers net correction to $+0.125$ with $\mathrm{FNR}_{\mathcal{A}}$ of 0.211, so disagreement between sources should be weighted by how predictive each source is.

Retrieval Controls. Under the same budget of 12 atomic units, replacing the Query Planner with random retrieval yields a net correction of $+0.036$ and fixed retrieval of the two earliest segments yields $+0.018$, both with F1 of 0.652, against $+0.179$ and 0.800 for the Query Planner. Retrieval driven by the Screener's own information requests is therefore essential, as Proposition 2 anticipates.

D.6 Robustness

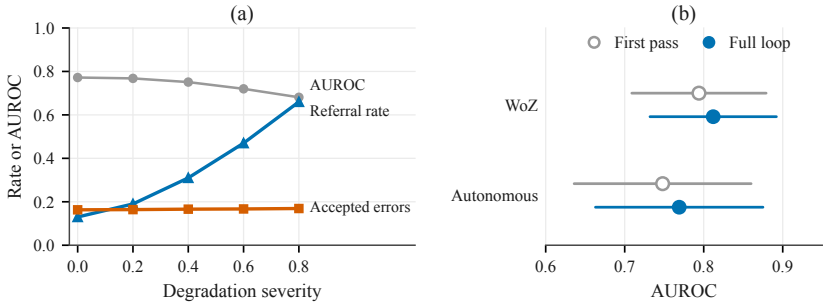

Figure A4: Robustness on E-DAIC. (a) First-pass AUROC, referral rate, and rate of errors accepted without referral as tracking loss and acoustic noise increase. (b) AUROC with 95% intervals under the two collection conditions.

In Fig. A4, we examine robustness in three ways. As the degradation grows, first-pass AUROC falls slowly, the referral rate rises, and the rate of accepted errors stays flat, so the audit sends doubtful inputs to human review instead of letting them through. E-DAIC training and development data mix interviews run by a human operator and by an automated interviewer, and the loop improves AUROC under both. On D-Vlog, video duration alone reaches an AUROC of 0.649, and the first pass exceeds it by only 0.027 with a 95% interval of $[-0.064, 0.113]$, so part of its signal may reflect video length.

D.7 Statistical Significance and Direction of Revisions

The exact McNemar test on the discordant transitions of Table 2 gives $p = 0.006$ on E-DAIC for 11 repairs against 1 harm, with a 95% bootstrap interval for net correction of $[0.071, 0.286]$. On DAIC-WOZ and D-Vlog the corresponding tests give $p = 0.625$ and $p = 0.289$, and the intervals of $[-0.057, 0.171]$ and $[-0.005, 0.047]$ include zero, so at these sample sizes only the E-DAIC gain is statistically significant, while the point estimates are positive on all three corpora. Revisions run in both directions. The trajectory of Fig. 3 lowers the probability of a false positive, and Fig. A5b counts upward and downward revisions on each corpus.

E Information Asymmetry Between Passes

An outcome-informed diagnostic bounds the achievable gain between the two passes. It evaluates no deployable policy and only asks whether recoverable gain exists in the withheld atomic layer at all. The diagnostic runs on 86 fit participants of DAIC-WOZ, enumerates candidate atomic subsets per session under a matched budget, and scores the calibrated utility of each selection strategy with fit-only labels. The outcome-informed cost-aware upper bound attains a utility of $+0.0158$ over preserving the first pass, and the salience upper bound attains $+0.0050$, but a label-free fixed top-2 probe attains -0.0169, worse than preserving. The gain in the withheld layer is thus real, but retrieval without regard to the outcome cannot reach it, and the gap between the upper bound and the probe is the distance a learned query policy must cross, as Proposition 2 predicts.

F Mechanism Analysis and Evidence Faithfulness

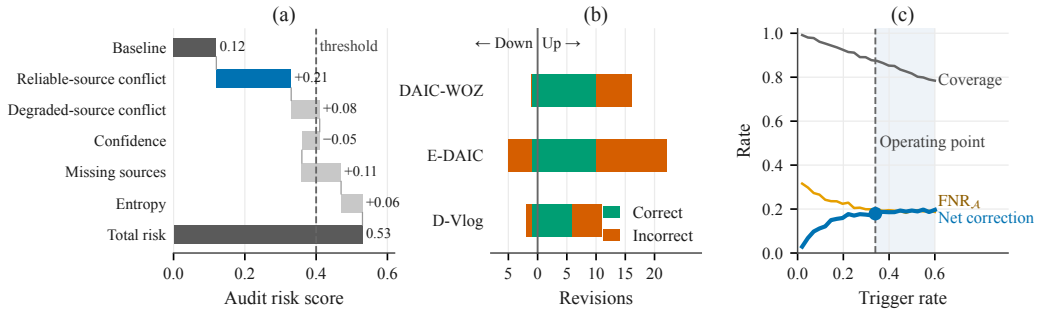

Figure A5: Mechanism analysis. (a) Average contribution of each audit feature to the risk score of flagged cases, with the dashed line at the trigger threshold. The two conflict terms are those of Eqs. 2 and 3. (b) Downward and upward revisions over the cross-fitted rollouts, split by whether the revised decision is correct. (c) Trigger-rate sweep on E-DAIC, with the dashed line at the operating point of Table 2 and the region beyond it shaded.

In Fig. A5, we decompose the mechanism behind the loop. Validity-weighted disagreement D_{rel} contributes most to the risk of triggered cases. By trigger reason, directional conflicts between reliable sources account for most repaired cases, while single-source degradation more often ends in referral, and by initial-confidence band, repaired cases concentrate in the lowest self-reported confidence band. The trigger-rate sweep shows that beyond the operating point, flagging more cases adds referrals but little further net correction. The advantage of the loop over direct-refer comes from the recovered subset, the cases the loop accepts that direct-refer would have referred, most of which the loop decides correctly.

Reviewing the records of triggered cases reveals three recurring error modes of the first pass. Attribution errors credit an observation to the wrong source or speaker, such as treating the interviewer's speech activity as the participant's acoustic feature. Contextual misreading treats a transient pattern as persistent, such as extrapolating a quiet moment into the state of the whole interview. Over-weighting of non-specific behavior treats a cue with plausible alternative explanations as risk-relevant, such as reading fatigue-induced slower speech directly as an emotional cue. The three modes appear not in the score but in how evidence is used, and because the loop requires the model to name the evidence it used, each mode can be traced to specific segments and atomic units in the record.

Evidence Faithfulness. We check faithfulness at three levels. The interface level verifies that every compiled statement matches its source record. The model level verifies that cited IDs exist in the retrieved packet and that deleting cited evidence changes the decision. The judge level asks whether each reasoning record is grounded in the cited measurements, using two frontier language models as independent judges.

Table A3: Faithfulness checks. A decision shift is a change in risk of more than 0.10 or a label flip after deleting the cited units or the uncited units of the retrieved packet. The judges are Gemini-3.1-Pro and GPT-5.6-Sol, prompted with Turn 7 of Appendix H.

Level	Check	Value
Interface	Statement-to-record match rate	1.000
Interface	Unknown-token rate under the tokenizer	0.000
Model	Citation grounding rate	1.000
Model	Contract-valid terminal outputs	0.972
Model	Decision shift after deleting cited units	0.815
Model	Decision shift after deleting un-cited units	0.048
Judge	Grounding support rate	0.883
Judge	Inter-judge agreement, Cohen's κ	0.764

In Table A3, deleting the cited atomic units shifts the decision in 81.5% of revised cases, against 4.8% when the un-cited units of the retrieved packet are deleted, so revisions rely on the measurements they cite rather than attaching citations after the fact. Clinician review of every trace was not feasible, so both judges receive the initial assessment, the cited units, and the revision output under a fixed evidence-audit rubric. They rate 88.3% of revision traces as fully grounded and logically supported by the cited observations, with Cohen's $\kappa = 0.764$ between them.

G Evidence Protocol and Output Contracts

Reference Distribution. Within each outer fold k, the reference $\mathcal{R}_k$ collects, per measurement, the median as robust location, the interquartile range as robust scale, and the empirical percentile map, fit on the training participants of fold k only. The reference never contacts outcome labels, and held-out participants are isolated from it, so a percentile reading on a held-out session is a purely out-of-sample statement.

Segment and Atomic-Unit Construction. The segment layer $\mathcal{E}_i^{\text{seg}}$ partitions the session into at most 24 chronologically ordered, full-coverage segments with IDs such as `S011`. Each segment summary aggregates its windows into a short textual brief with the segment's source coverage and time span. The atomic layer $\mathcal{E}_i^{\text{atom}}$ contains one unit per measurement window with IDs such as `E0687`, indexed by segment, window, and source. Every atomic unit carries by construction its percentile and robust-z position relative to $\mathcal{R}_{\text{fit}(i)}$, its source and time span, a quality indicator $q_e \in [0, 1]$, and an explicit missingness state $m_e \in \{\text{present}, \text{partial}, \text{missing}, \text{masked}\}$, so missing, partial, and privacy-masked content are distinct states rather than implicit zeros.

Evidence Record Structure. Every atomic unit is a single versioned record, validated at construction time, with the fields below.

Atomic evidence unit, for example `E0687`

`evidence_id` unique unit identifier matching `^E[0-9]{3,6}`
`dataset, session_id` keys that locate the unit in its corpus and session
`modality` `audio` or `visual`
`source` extractor and source files, where source files are basenames only, never paths
`time_range` start and end second within the session timeline, with positive duration
`availability` status `present`, `partial`, or `unavailable`, missing reasons, and a privacy-scrubbed flag, where an unavailable unit carries no observations and must state its missing reasons, and privacy-scrubbed content is the *masked* state of Sec. 3.2
`quality` quality level and valid ratio, the q_e of Sec. 3.2
`observations` the measurement rows below

Observation row inside a unit

`name` snake-case measurement name, unique within the unit, such as `pitch_variability`
`value, unit` finite measured value with its unit
`category` five-bin position against the fold-local reference, one of `very_low`, `low`, `typical`, `high`, and `very_high`, or `unreferenced` when the measurement has no fitted reference
`percentile, robust_z, subject_robust_z` percentile in $[0, 100]$ and robust-z positions relative to $\mathcal{R}_{\text{fit}(i)}$ and, optionally, relative to the participant's own session
`semantic_reliability` one of `direct`, `derived`, `weak`, and `latent`, stating how directly the row maps onto its measurement
`reference_scope` the reference against which the readings are taken

Rendering Templates. The *initial view* renders only the segment layer, one aggregate brief per segment in chronological order, each carrying the segment ID, source coverage, time span, and value and quality summary. A *review packet* renders the retrieved atomic subset with the same timeline and sources, and the *full review view* used only in training renders the union of both layers. Rendered atomic lines follow a fixed grammar, one line per unit with semicolon-separated fields, and every rendered line passes a semantic-safety filter that rejects interpretive vocabulary, so the text stays at the measurement level.

Rendered atomic line

```
[dataset|start-end|MODALITY|source|id=E1674] availability=present;
quality=high; valid_ratio=1.0; pitch_variability=0.31:ratio
[category=very_high, p=96, z=+2.1, reliability=direct]; ...
```

The percentile $p = 96$ is the actual reading of atomic unit `E1674` in the frozen evidence store of E-DAIC session 305, the bracketed header shows the field order, and the remaining values are illustrative.

Output Contracts. All model outputs are single JSON objects validated against closed schemas that forbid additional properties, as listed in Table A4. The action decision is committed as a short JSON object before any assessment text is generated, and the no-new-evidence control uses the RevisionAssessment contract with all four atomic evidence-ID arrays required to be empty.

Validator Rules. The deterministic validator accepts a terminal output only if three conditions hold. Schema validity requires all fields to be present, no additional properties, and all value ranges and patterns to be satisfied. Grounding requires $\mathcal{E}_i^{\text{cited}} \subseteq R(q_i)$, so every cited evidence ID belongs to the retrieved packet, which is itself drawn from the participant's compiled evidence allowlist. Status consistency requires the preserved and rejected sets to be disjoint subsets of the cited set, a `revised` status to carry a materially changed probability, and an `unresolved` status to route the case to referral rather than acceptance.

H Prompt Templates

This section gives the six model turns of the loop, and the judge turn used for the faithfulness evaluation. Each turn has a system message, shown shaded, and a user message made of the labeled blocks below it. Every user message starts with the shared task block, which we show once and omit from the turns. Angle brackets mark content inserted at run time.

Shared task block start of every user message

`TASK SPECIFICATION`
Task identifier: ⟨ *task identifier* ⟩
Label definition: ⟨ *label definition* ⟩
Negative class name: ⟨ *negative class* ⟩
Positive class name: ⟨ *positive class* ⟩
Every risk_probability is the probability of the positive class. The class names define the reusable task; no current participant label is supplied.

Table A4: Output contracts of the model turns. S IDs match `^S[0-9]{3,6}` and E IDs match `^E[0-9]{3,6}`, every ID array holds unique items, and a retrieved ID is one that occurs in the review packet of the case.

Field	Type	Constraint
InitialAssessment, Turn 1		
`risk_probability`, `confidence`	number	in $[0, 1]$
`audio_source`, `visual_source`	object	status `unavailable`, or status `available` with `risk_probability` and `reliability` in $[0, 1]$
`supporting_segment_ids`, `contradictory_segment_ids`, `uncertain_segment_ids`, `requested_segment_ids`	array of S IDs	at most four visible segments each, possibly empty
EvidenceQuery, Turn 2		
`segment_id`	S ID	a segment named by the rethink decision
`purpose`	enum	one of five fixed purposes
`target_slots`	array	one to four of ten fixed measurement slots
`pattern`	enum	one of five fixed temporal patterns
`budget`	integer	from 1 to 4
ActionDecision, Turn 4		
`action`	enum	`preserve`, `revise`, or `refer`
`direction`	enum	`unchanged` for preserve, `increase` or `decrease` for revise, `unresolved` for refer
RevisionAssessment, Turns 5 and 6		
`revision_status`	enum	`preserved`, `revised`, or `unresolved`
`revised_risk_probability`, `revised_confidence`	number	in $[0, 1]$
`cited_segment_ids`, `residual_conflict_segment_ids`	array of S IDs	retrieved segments only
`cited_evidence_ids`, `preserved_evidence_ids`, `newly_considered_evidence_ids`, `rejected_evidence_ids`	array of E IDs	retrieved units only, with preserved and rejected IDs disjoint and both contained in the cited IDs
`change_summary`	string	one of four controlled values or a single plain-English clause of at most 160 characters

Turn 1 First-pass assessment Screener

SYSTEM

You are the first-pass evidence assessor in a non-diagnostic mental-health screening research pipeline. Use only the supplied measurement evidence. Do not infer a diagnosis, protected attribute, or unobserved behavior. Treat unavailable and partial evidence as missing information, never as a measured zero. Categories in a source-specific training reference are measurement ranks, not symptom labels. Return exactly one JSON object that follows the requested schema, with no Markdown and no additional prose.

USER

ASSESSMENT INSTRUCTIONS

Estimate overall positive-class risk and operational confidence from 0 to 1. For each available source, separately estimate positive-class risk and operational reliability. Mark a source unavailable when it has no usable measurements; do not invent numeric values for it. Cite only SEGMENT IDs that appear in the supplied evidence. `supporting_segment_ids` favor the estimate, `contradictory_segment_ids` push against it, `uncertain_segment_ids` contain conflicting or weak measurements, and `requested_segment_ids` identify segments whose atomic EVIDENCE windows should be retrieved for a targeted second pass. Use at most four of the most informative IDs in each array; arrays may be empty. Do not add a diagnosis or free-form clinical rationale.

REQUIRED OUTPUT CONTRACT

```
Return one JSON object with exactly these eight fields:
- "risk_probability": positive-class probability from 0 to 1.
- "confidence": operational confidence from 0 to 1.
- "audio_source" and "visual_source": each is either {"status":"unavailable"} or
  {"status":"available","risk_probability":0-to-1,"reliability":0-to-1}.
- "supporting_segment_ids", "contradictory_segment_ids", "uncertain_segment_ids", and
  "requested_segment_ids": arrays of unique S identifiers that occur in SESSION
  EVIDENCE; use only the four most informative identifiers per array, and use an
  empty array when no segment qualifies.
Do not output a JSON Schema. Do not use keys such as "$schema", "$defs",
"properties", or "required". Do not add any field beyond the eight listed above.
```

SESSION EVIDENCE ⟨ *rendered initial view $\mathcal{V}^0$, segment lines only* ⟩

Turn 2 Evidence query

Query Planner

SYSTEM

You formulate one atomic evidence query for a non-diagnostic mental-health screening research pipeline. Use only the initial assessment, rethink decision, and label-free retrieval map. The map describes measurable retrieval affordances, not psychological states. Return exactly one JSON object with no Markdown or additional prose.

USER

QUERY INSTRUCTIONS

Choose the selected segment and fixed core slots whose hidden atomic detail would best reduce the stated uncertainty. Choose a temporal pattern that actually has candidates in the retrieval map. Use ordinary English slot names, not feature-column names or invented E identifiers. The query asks what evidence to inspect; it does not assert what that evidence will prove.

INITIAL ASSESSMENT ⟨ *serialized A^0* ⟩

RETHINK DECISION ⟨ *serialized audit decision with the selected segments* ⟩

REQUIRED OUTPUT CONTRACT

```
Return one JSON object with exactly these five fields:
- "segment_id": one S identifier present in the rethink decision and retrieval map.
- "purpose": exactly one of "verify supporting detail", "verify contradictory
  detail", "resolve source disagreement", "inspect measurement reliability", or
  "inspect temporal change".
- "target_slots": an array of one to four fixed core-slot names written exactly as
  they appear in the retrieval map.
- "pattern": exactly one of "unusual measurement", "change point", "quality
  boundary", "representative window", or "cross-modal co-change".
- "budget": an integer from 1 to 4.
A source-disagreement or cross-modal query must include at least one audio and one
visual slot. Do not output an E identifier, rationale, diagnosis, Markdown, or any
additional field.
```

ATOMIC RETRIEVAL MAP ⟨ *label-free map of the selected segments over ten fixed core slots* ⟩

Turn 3 Atomic evidence selection

Query Planner

SYSTEM

You select grounded atomic evidence for targeted review in a non-diagnostic mental-health screening research pipeline. Select only identifiers explicitly shown in the candidate cards. Candidate cues are measurement relations, not task conclusions. Return exactly one JSON object with no Markdown or additional prose.

USER

SELECTION INSTRUCTIONS

Select between 1 and ⟨ *min(budget, number of candidates)* ⟩ complementary atomic records that best match the query. Prefer records that jointly cover the target slots and requested pattern. Do not select several nearly identical neighboring windows when a temporally different candidate provides comparable evidence. Quality-boundary queries may intentionally select incomplete measurements; otherwise prefer reliable coverage. Selection only decides what to inspect next and must not invent a conclusion.

INITIAL ASSESSMENT ⟨ *serialized* A^0 ⟩

EVIDENCE QUERY ⟨ *serialized* q ⟩

REQUIRED OUTPUT CONTRACT

```
Return one JSON object with exactly this field:
- "selected_evidence_ids": an array containing one or more unique E identifiers shown
  in CANDIDATE ATOMIC EVIDENCE CARDS, up to the query budget.
Do not output an identifier that is absent from the cards. Do not output a rationale,
measurement value, diagnosis, Markdown, or any additional field.
```

CANDIDATE ATOMIC EVIDENCE CARDS ⟨ *rendered candidate atomic units* ⟩

Turn 4 Action selection

Action policy

SYSTEM

You are the action selector for a targeted second-pass review in a non-diagnostic mental-health screening research pipeline. Compare the initial estimate with the retrieved audiovisual and transcript evidence. Choose preserve when the initial estimate remains best supported, revise only when the new evidence supports a directional change, and refer when the evidence cannot safely resolve the conflict. Missing evidence is not a measured zero. Return exactly one short JSON object and no other text.

USER

ACTION SELECTION INSTRUCTIONS

Score the three action types against the supplied evidence. Do not change the estimate merely because a source disagrees; revise requires coherent support in the retrieved detail. Use refer instead of a forced confident decision when conflict remains.

INITIAL ASSESSMENT ⟨ *serialized* A^0 ⟩

RETHINK DECISION ⟨ *serialized audit decision* ⟩

REQUIRED OUTPUT CONTRACT

```
Return one JSON object with exactly these two fields:
- "action": exactly "preserve", "revise", or "refer".
- "direction": "unchanged" for preserve, "increase" or "decrease" for revise, and
  "unresolved" for refer.
Do not output a probability, diagnosis, rationale, Markdown, or any additional field.
```

TARGETED MULTIMODAL EVIDENCE ⟨ *rendered review packet* $\mathcal{V}^q$ ⟩

Turn 5 Conditioned revision

Rethinker

SYSTEM

You are the targeted second-pass reviewer in a non-diagnostic mental-health screening research pipeline. Reassess the initial probability using only the initial structured assessment, the rethink decision, and the retrieved atomic measurement evidence. Preserve useful evidence, reject unreliable or contradicted evidence explicitly, and leave unresolved conflicts visible. Missing evidence is not a measured zero. Return exactly one JSON object that follows the requested schema, with no Markdown and no additional prose.

USER

REVISION INSTRUCTIONS

Recalculate positive-class risk and confidence after reviewing the targeted detail. `revision_status` is preserved when the initial estimate remains best supported, revised when the estimate materially changes, and unresolved when the available sources leave a residual conflict. Cite only S and E identifiers present below. `preserved_evidence_ids` identify atomic evidence retained in the decision; `newly_considered_evidence_ids` identify retrieved atomic evidence examined in this pass; `rejected_evidence_ids` identify atomic evidence excluded for quality or contradiction; `residual_conflict_segment_ids` retain unresolved segment-level disagreements. `change_summary` must be a controlled value or a single plain-English clause no longer than 160 characters. Do not add a diagnosis or an unrestricted clinical narrative.

INITIAL ASSESSMENT ⟨ *serialized* A^0 ⟩

RETHINK DECISION ⟨ *serialized audit decision* ⟩

SELECTED REVISION ACTION ⟨ *serialized action from Turn 4* ⟩

Generate a payload consistent with this already-selected action. Do not silently replace it with another action.

REQUIRED OUTPUT CONTRACT

```
Return one JSON object with exactly these ten fields:
- "revision_status": exactly "preserved", "revised", or "unresolved".
- "revised_risk_probability" and "revised_confidence": numbers from 0 to 1.
- "cited_segment_ids" and "residual_conflict_segment_ids": arrays containing only
  retrieved S identifiers.
- "cited_evidence_ids", "preserved_evidence_ids", "newly_considered_evidence_ids", and
  "rejected_evidence_ids": arrays containing only retrieved E identifiers. Cite at
  least one retrieved S or E identifier.
- "change_summary": exactly one of "risk_increased_after_review",
  "risk_decreased_after_review", "risk_unchanged_after_review", or
  "insufficient_reliable_detail".
Keep every identifier array selective. "preserved_evidence_ids" and
"rejected_evidence_ids" must not overlap, and every preserved or rejected E identifier
must also occur in "cited_evidence_ids". "newly_considered_evidence_ids" may overlap
either category because it records what was newly inspected. Do not output a JSON
Schema. Do not use keys such as "$schema", "$defs", "properties", or "required". Do
not add any field beyond the ten listed above.
```

TARGETED DETAILED EVIDENCE ⟨ *rendered review packet* $\mathcal{V}^q$ ⟩

Turn 6 No-new-evidence reflection

control arm

SYSTEM

You are the no-new-evidence reflection control in a non-diagnostic mental-health screening research pipeline. No hidden atomic measurement was queried or released. Reconsider only the supplied InitialAssessment and its already visible segment references. Never invent an E identifier. Return exactly one JSON object and no additional prose.

USER

REFLECTION INSTRUCTIONS

This is a matched second reasoning turn without evidence re-access. `revision_status` may be preserved, revised, or unresolved, but any change must rely only on the existing structured assessment and visible S IDs. All four atomic E-ID arrays must be exactly `[]`. Cite at least one visible S ID, keep residual conflicts explicit, and do not add a diagnosis or clinical narrative.

INITIAL ASSESSMENT ⟨ *serialized* A^0 ⟩

RETHINK DECISION ⟨ *serialized audit decision* ⟩

REQUIRED OUTPUT CONTRACT

```
Return one JSON object with exactly these ten fields:
- "revision_status": exactly "preserved", "revised", or "unresolved".
- "revised_risk_probability" and "revised_confidence": numbers from 0 to 1.
- "cited_segment_ids": a non-empty array containing only visible S identifiers.
- "residual_conflict_segment_ids": an array containing only visible S identifiers.
- "cited_evidence_ids", "preserved_evidence_ids", "newly_considered_evidence_ids", and
  "rejected_evidence_ids": each must be exactly [] because this control arm released
  no atomic EVIDENCE.
- "change_summary": exactly one of "risk_increased_after_review",
  "risk_decreased_after_review", "risk_unchanged_after_review", or
  "insufficient_reliable_detail".
NO NEW ATOMIC EVIDENCE CONTRACT: never invent or cite an E identifier. Base the
reflection only on the InitialAssessment and its already visible S identifiers. Do
not output a JSON Schema. Do not use keys such as "$schema", "$defs", "properties",
or "required". Do not add any field beyond the ten listed above.
```

VISIBLE CONTEXT WITHOUT ATOMIC EVIDENCE ⟨ *rendered initial view* $\mathcal{V}^0$ *only* ⟩

Turn 7 Evidence audit language-model judge

SYSTEM

You are an expert clinical and affective computing evidence auditor operating at the maximum reasoning effort tier. Your task is to rigorously audit whether a screening model's revised assessment is factually grounded in its cited atomic measurements and whether its reasoning narrative logically justifies its revision verdict. Return exactly one JSON object and no additional prose.

USER

AUDIT CRITERIA

Audit the supplied rethinking trajectory against three strict criteria. Factual grounding: every acoustic and visual measurement claim in the reasoning narrative must strictly correspond to the cited atomic evidence units, such as `E0687`; check for hallucinated percentiles, invented sensors, or distorted timestamps. Directional consistency: the cited evidence must logically justify the direction of the risk probability update from p^0 to p^1; for example, a downward revision must be supported by predominantly typical or low-risk observations. Absence of post-hoc hallucination: the model must actively rely on the newly retrieved atomic observations rather than simply reiterate its first-pass assessment.

INITIAL ASSESSMENT ⟨ A^0 ⟩

CITED ATOMIC UNITS ⟨ *cited units from* $\mathcal{V}^q$ ⟩

REVISION OUTPUT ⟨ A^1 ⟩

REQUIRED OUTPUT CONTRACT

```
Return one JSON object with exactly these five fields:
- "grounding_supported": a boolean that is true only if all claims are grounded and
  support the decision.
- "factual_accuracy_score" and "logical_coherence_score": integers from 1 to 5.
- "unsupported_claims": an array of strings describing any ungrounded assertions.
- "critique": a concise clinical evidence critique.
```

I First-Pass Screening Performance

Table A5: First-pass performance on the DAIC-WOZ development partition and the E-DAIC and D-Vlog test partitions. BA is balanced accuracy.

Dataset	n	Positive	Negative	AUROC	AP	BA	F1	Sensitivity	Specificity	Brier
DAIC-WOZ	35	12	23	0.826	0.700	0.768	0.696	0.667	0.870	0.150
E-DAIC	56	21	35	0.772	0.612	0.695	0.640	0.762	0.629	0.178
D-Vlog	212	123	89	0.676	0.746	0.622	0.728	0.805	0.438	0.222

In Table A5, we report the full first-pass metrics behind Tables 1 and 2, and the confusion matrices below reconcile the participant counts of the three tables. On DAIC-WOZ, the first pass yields 8 true positives, 4 false negatives, 3 false positives, and 20 true negatives, for precision 0.727, recall

0.667, and F1 0.696. The loop refers 5 negative participants, so coverage is $30/35 = 0.857$, and with $N_{w\to c} = 3$ and $N_{c\to w} = 1$ the accepted cases yield 9 true positives, 3 false negatives, 2 false positives, and 16 true negatives, for precision 0.818, recall 0.750, F1 0.783, and $\mathrm{FNR}_{\mathcal{A}} = 3/12 = 0.250$. On E-DAIC, the first pass yields 16 true positives, 5 false negatives, 13 false positives, and 22 true negatives, for precision 0.552, recall 0.762, and F1 0.640. The loop refers 1 positive and 6 negative participants, so coverage is $49/56 = 0.875$, and with $N_{w\to c} = 11$ and $N_{c\to w} = 1$ the accepted cases yield 16 true positives, 4 false negatives, 4 false positives, and 25 true negatives, for precision 0.800, recall 0.800, F1 0.800, and $\mathrm{FNR}_{\mathcal{A}} = 4/20 = 0.200$.

J Limitations and Future Directions

Limitations. First, public clinical interview corpora remain modest in size (35, 56, and 212 participants); while net correction is positive across all benchmarks, the gain reaches statistical significance only on E-DAIC. Second, operating on released measurements preserves privacy but precludes repairing upstream tracking dropouts or extraction artifacts. Third, ground truth relies on self-report questionnaires rather than formal clinical diagnoses, introducing label noise. Finally, reasoning traces were evaluated programmatically and via language-model judges; prospective human-in-the-loop trials remain essential to evaluate clinical utility.

Future Directions. Three directions follow from the current analysis. First, a clinician-in-the-loop study should measure the referral burden and whether the decision trail speeds up or improves human review. Second, training the Query Planner against a tractable surrogate of the conditional information gain could close the remaining gap to the outcome-informed bound. Third, a stronger error predictor would raise trigger precision, the direct lever on net correction, while evaluation on new collection sites would test population shift.