EVIDIAL: WHEN CORRECT DEPRESSION PREDICTIONS LACK EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Depression dialogue benchmarks typically evaluate models by comparing predicted symptom labels against clinical scale scores. Yet a prediction can match the benchmark label even when the dialogue does not contain sufficient evidence to support that assessment. We introduce EVIDIAL, a benchmark for measuring this distinction between *label agreement* and *evidence-grounded assessment*. EVIDIAL covers 382 depression dialogues from three sources and 2,744 participant–symptom cells, each characterized under an evidence taxonomy of *absent*, *partial*, *conflicting*, or *sufficient* evidence. Because evidence sufficiency itself has no directly observable gold standard, we construct calibrated evidence references from multiple judgments using generalizability theory, many-facet Rasch measurement, and random-effects latent class analysis. Auditing existing benchmarks with these references reveals a substantial evidence gap: 82.8% of participant–symptom cells do not contain sufficient textual evidence for the corresponding symptom assessment, with evidence availability varying sharply across data sources. More importantly, across evaluated model families and prompting strategies, 69.2–96.9% of predictions that are *correct according to the benchmark label* remain unsupported by sufficient evidence in the dialogue. Additional measurement analyses show that high agreement among LLM evaluators does not necessarily imply independent evidence judgments, motivating explicit calibration of automated reference panels. Conventional label-based evaluation can substantially overstate evidence-grounded performance by conflating agreement with clinical labels and support from the observed dialogue. EVIDIAL provides a measurement layer for separating these two constructs and enables more faithful auditing of LLM behavioral assessment in depression dialogues.

1 INTRODUCTION

Depression dialogue benchmarks record real or simulated conversations between clinicians and individuals with potential depression, and often evaluate models by comparing predicted symptom labels with clinical or self-report scores. As shown in Figure 1 A., consider an interview paired with a recorded sleep-symptom label. A model is asked to evaluate the respondent on the sleep disturbance dimension. It predicts *no sleep disturbance*-and matches the native symptom label exactly. However, label accuracy is a success: one more correct cell in the accuracy table. But the interview never discussed sleep. The model produced the right answer without supporting evidence in the released dialogue–one label–support divergence we term **Textually Unsupported Correctness**. The term locates a mismatch between a credited prediction and the observable input; it is not a mechanism claim that the model ignored evidence, used priors, or reasoned in any particular way.

Benchmarks such as PDCH (Cao et al., 2025), E-DAIC (Gratch et al., 2014), EATD (Shen et al., 2022), and their derivatives report label accuracy, macro F1, or mean absolute error. These metrics do not directly distinguish two fundamentally different outcomes: a prediction grounded in textual evidence and one that matches the label without support for the predicted symptom state. An explicit denial and silence about a symptom can receive the same negative label, although only the denial provides negative evidence within its stated scope. Label agreement alone therefore cannot establish that an assessment is justified by the available dialogue. A native clinical score may incorporate information outside the released transcript, and the issue here is what the text supports, rather than whether that score is wrong.

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.

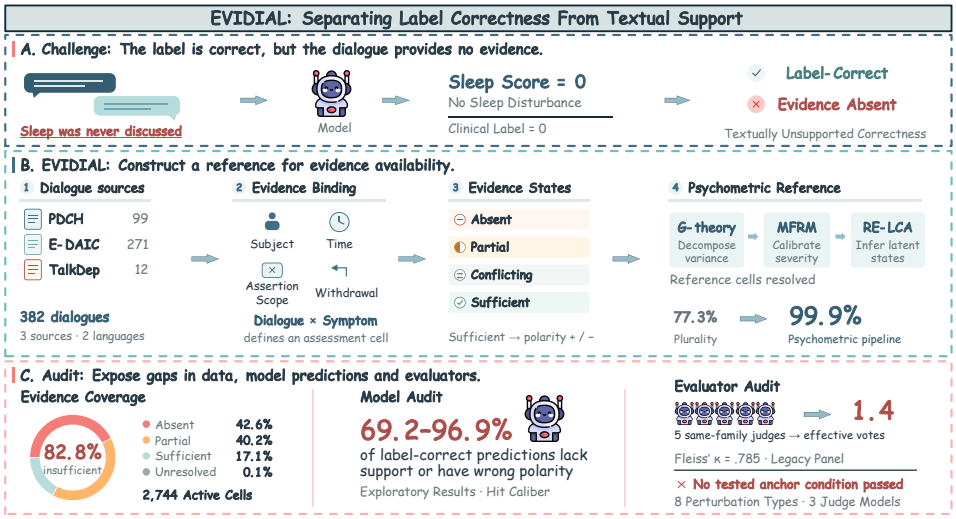

Figure 1: Existing depression dialogue benchmarks have one **challenge**: Textually Unsupported Correctness *(A.)*. We propose EVIDIAL, a reference benchmark for evidence availability *(B.)*, and audit this phenomenon by exposing gaps in data, model predictions, and evaluators *(C.)*.

We introduce EVIDIAL (Evidence-Dialogue), a benchmark designed to make this distinction explicit. Across 2,744 assessment cells from 343 clinical participants, with eight source-specific dimensions per participant, 82.8% of cells are assigned absent or partial evidence by the candidate reference (95% CI: 82.1–83.5%). This pooled share is composition-dominated: sufficient evidence is assigned to 74.1% of PDCH cells but only 1.9% of E-DAIC cells, so **evidence availability behaves as a property of the benchmark–instrument–transcript interface rather than of the clinical label itself**. The gap is also not confined to negative labels, where negative evidence is structurally harder to establish (§3.2): 76.0% of positive-native-label cells lack sufficient evidence (95% CI: 73.9–78.1%), against 91.7% of negative-label cells. Among predictions that do match native labels, 69.2–96.9% lack reference support for the predicted direction across three model families and four prompting methods on 87 development cells. As a fraction of all planned predictions, the corresponding incidence is 20.7–54.0%—a quantity we term the unsupported-correct rate (UCR). These findings motivate an audit of both model predictions and the evidence judgments used to evaluate them. Textually Unsupported Correctness matters when benchmark accuracy is interpreted as evidence of clinical assessment capability. If a model achieves high accuracy on depression severity without an audit of textual support, we cannot tell which credited predictions are warranted by its input. A model that has learned "most people in this corpus score 0 on sleep disturbance" could receive credit without establishing that a particular respondent denied the symptom (Geirhos et al., 2020). If that regularity changes in another population, the original accuracy score alone gives no assurance that those predictions will remain reliable. Evaluation therefore needs to distinguish supported predictions, unsupported label matches, and cases in which the available dialogue warrants abstention or further inquiry. Making these differences visible clarifies what benchmark performance can substantiate before it is used to argue for clinical readiness.

EVIDIAL adds this evidence-support layer to label-based evaluation. The core collection comprises 382 dialogues from three sources: PDCH (99 dialogues, Chinese), E-DAIC (271 dialogues, English), and TalkDep (12 dialogues, English). Repeated PDCH visits and simulated TalkDep dialogues are outside the main clinical reference pool. Each dialogue–dimension pair constitutes an *evidence cell*, annotated under a four-state taxonomy: ABSENT (no eligible evidence), PARTIAL (some indicators present but insufficient for the scoped binary target), CONFLICTING (contradictory signals), or SUFFICIENT (enough evidence to support one target polarity). Annotations are grounded through formal binding operations that enforce constraints on subject identity, temporal reference, negation scope, and symptom withdrawal over judge-extracted evidence frames, which makes the distinction between missing evidence and evidence of symptom absence explicit in the evaluation target.

Since no single annotator or rubric can serve as an uncontested gold standard (Artstein & Poesio, 2008), we audit how the evidence reference is constructed. Agreement among judges does not by itself establish that their evidence assignments are valid. Random-effects latent class analysis

(RE-LCA) (Qu et al., 1996) combines the new panel's categorical votes into a model-consensus candidate reference. Separately, generalizability theory (G-theory) (Brennan, 2001) quantifies variance components across cells, participants, and protocols in a legacy panel; many-facet Rasch measurement (MFRM) (Linacre, 1989) calibrates legacy protocol severity. Aggregation comparisons expose sensitivity to reference choices, while perturbation anchors test judgments against controlled changes to evidence. Our findings remain conditional on the candidate reference; a blinded human validation study (§5) supports the main evidence gap, though uncertainty calibration is incomplete. Our work makes three contributions: **1. Evidence-support taxonomy and benchmark**. We introduce EVIDIAL, which augments conventional symptom-label evaluation with a four-state taxonomy of textual evidence support, enabling direct assessment of whether a prediction is justified by the dialogue. **2. Psychometric reference construction and audit**. We develop a psychometric framework for constructing and auditing evidence references when no single judge can be treated as ground truth, explicitly accounting for judge dependence and reference uncertainty. **3. Empirical evidence-gap audit.** Using EVIDIAL, we show that label correctness and evidence support can diverge substantially across models and prompting strategies, revealing failure modes that conventional accuracy alone cannot capture.

2 RELATED WORK

Depression and psychiatric dialogue benchmarks. Clinical dialogue resources such as DAIC (Gratch et al., 2014), PDCH (Cao et al., 2025), and EATD (Shen et al., 2022) support depression assessment from interviews or elicited speech, while newer benchmarks examine robustness across splits, models, and symptom distributions (Ishikawa & Duke, 2026). Broader psychiatric benchmarks evaluate diagnostic capabilities of LLMs and multi-agent systems (Song et al., 2026; Xu et al., 2026). These benchmarks primarily evaluate agreement with clinical labels or diagnostic targets. EVIDIAL instead asks a complementary question: whether the processed dialogue itself contains sufficient evidence to justify the assessed symptom label.

LLM evaluators and reference construction. LLMs are increasingly used as evaluators (Zheng et al., 2023; Tan et al., 2025; Kim et al., 2024; KC, 2026), including multi-judge panels that can improve reliability over individual judges (Verga et al., 2024). However, correlated errors and shared model biases can make apparent agreement overstate the amount of independent information in a panel (Kohli, 2026). Recent work therefore applies psychometric tools such as item-response theory, generalizability theory (Brennan, 2001), and Rasch measurement (Linacre, 1989) to characterize LLM-rater reliability and severity (Choi et al., 2026; Wang et al., 2025; Song et al., 2025; Sunkavalli, 2026). Separately, latent-class and observer-error models estimate reference states when ground truth is unavailable (Dawid & Skene, 1979; Qu et al., 1996; Albert & Dodd, 2004). We bring these perspectives together to construct and audit evidence references without treating either a single LLM judge or simple panel agreement as ground truth.

Evidence support, faithfulness, and selective prediction. Our distinction between label correctness and textual support is related to faithfulness and attribution in generation (Maynez et al., 2020; Rashkin et al., 2023; Gao et al., 2023; Honovich et al., 2022; Min et al., 2023), where an output may appear correct while lacking support from its source. Clinical NLP has likewise modeled contextual properties such as negation, temporality, and experiencer (Chapman et al., 2001; Harkema et al., 2009), while work on context sufficiency and selective prediction studies whether available information is adequate for answering and when a model should abstain (Joren et al., 2025; Geifman & El-Yaniv, 2017; Kamath et al., 2020). EVIDIAL adapts these ideas to symptom-level behavioral assessment by explicitly measuring whether benchmark-correct predictions are supported by the evidence available in the dialogue.

3 THE EVIDIAL BENCHMARK

To audit the gap between label correctness and evidence-grounded assessment, EVIDIAL defines an assessment unit, an evidence taxonomy with explicit binding rules, a multi-judge reference construction procedure, and evaluation metrics that separate label agreement from textual support.

3.1 ASSESSMENT CELLS AND DATA SOURCES

The unit of analysis is an *assessment cell* $u = (x, d)$, pairing a dialogue transcript x with a symptom dimension d and asking whether the transcript contains evidence bearing on that symptom for the participant. The benchmark contains 382 depression-dialogue transcripts from three sources. **PDCH**

Table 1: Binding operations applied to a sleep-disturbance cell. Each row shows a candidate evidence frame from the dialogue, the binding outcome, and its effect on the eligible evidence set.

Dialogue excerpt	Binding test	Result	Status
"My mother can't sleep at all"	Subject	Other person	Excluded
"I had terrible insomnia in college"	Temporal	Outside window	Excluded
"I haven't been sleeping well this week"	All pass	Affirms sleep difficulty	Eligible (+)
"Well, actually I meant last month"	Withdrawal	Retracts prior frame	Prior removed

$\mathcal{E}(u) = \emptyset$ after withdrawal $\Rightarrow s(u) =$ ABSENT

contains 99 Chinese psychiatric interviews from 72 participants, including 27 repeated visits, scored with HAMD-17 (Hamilton, 1986). **E-DAIC** contains 271 English semi-structured distress interviews scored with PHQ-8 (Gratch et al., 2014; Kroenke et al., 2009). **TalkDep** contains 12 scripted dialogues generated from BDI-II persona profiles; because these profiles are construction targets rather than independent clinical ground truth, TalkDep is excluded from the main reference pool. We align the sources to nine symptom dimensions: depressed mood, anhedonia, sleep disturbance, fatigue/energy, appetite/weight change, psychomotor change, guilt/worthlessness, concentration difficulty, and suicidal ideation. PDCH excludes concentration and E-DAIC excludes suicidality, yielding $72 \times 8 + 271 \times 8 = 2{,}744$ main-reference cells. Participants are split without overlap into 182 development and 161 held-out test participants; the model audit in §4.2 uses 87 development cells from 76 participants. E-DAIC transcripts were reformatted with `claude-3-5-sonnet-latest` (271/275 succeeded). Because preservation of clinical meaning has not been independently verified, our audit concerns the processed model-visible input rather than guaranteed coverage of the original interviews. Appendix A provides source statistics.

3.2 Evidence Taxonomy and Binding Operations

A prediction is evidence-grounded only if the dialogue contains eligible evidence for the target symptom. Candidate evidence frames are filtered by four binding operations: **subject binding** requires that the statement concern the participant; **temporal binding** restricts evidence to the instrument's target window; **assertion scope** determines whether the statement covers the relevant symptom components; and **withdrawal** removes subsequently retracted evidence. Presence requires at least one qualifying component, whereas establishing absence requires coverage of all relevant components, making sufficient-negative evidence structurally harder to establish. We audit this asymmetry separately in §4.3 and §4.1.

After binding, the surviving evidence frames $\mathcal{E}(u)$ determine compatible binary target values $K(u) \subseteq \{0, 1\}$:

$$s(u) = \begin{cases} \text{ABSENT}, & \mathcal{E}(u) = \emptyset, \\ \text{SUFFICIENT}, & \mathcal{E}(u) \neq \emptyset,\ |K(u)| = 1, \end{cases} \quad \begin{cases} \text{CONFLICTING}, & \mathcal{E}(u) \neq \emptyset,\ K(u) = \emptyset, \\ \text{PARTIAL}, & \mathcal{E}(u) \neq \emptyset,\ K(u) = \{0, 1\}. \end{cases} \tag{1}$$

For sufficient evidence, $K = \{1\}$ and $K = \{0\}$ denote positive and negative polarity, yielding five joint reference categories. An ABSENT state means that no eligible evidence survives binding, not necessarily that the symptom was never mentioned. Table 1 gives a concrete example.

3.3 Reference Construction

Because evidence state has no directly observable gold-standard annotator, we treat reference construction as a measurement problem with competing estimators. On the same vote pool, we compare strict plurality, a one-hot mean baseline, and random-effects latent-class fusion; §4.3 reports their coverage and validation diagnostics. The working model-consensus reference uses judgments from `gemini-2.5-flash`, `qwen3.8-max`, and `claude-sonnet-4-6`, stratified by language. Two DeepSeek models were excluded after failing format-validity gates. The retained pool contains 5,240 valid judgments over 2,744 cells, with uneven per-cell coverage (Table 2). We fuse these judgments using random-effects latent class analysis (RE-LCA) (Qu et al., 1996; Albert & Dodd, 2004), with five latent evidence states, judge-specific emission probabilities, and a cell-level random effect capturing dependence. RE-LCA serves as the working reference because it resolves split votes, not because it is assumed uniquely correct: available anchor comparisons favor plurality on correctness, so aggregation choice remains an explicit sensitivity analysis. Anchor-derived emission

priors and the uncalibrated posterior-abstention threshold are examined in §4.3; Appendix F gives the full specification.

3.4 Evaluation Protocol

The native binary target is $Y = \mathbf{1}[\text{item score} > 0]$ under the source instrument. Because native scores may incorporate information unavailable in the transcript, disagreement between label and evidence does not imply that the clinical assessment itself is incorrect. Our primary metric is the **Unsupported-Correct Rate (UCR)**,

$$\text{UCR} = \frac{U + D}{N - R}, \tag{2}$$

where U counts label-correct predictions whose reference evidence is absent, partial, or conflicting; D counts label-correct predictions for which sufficient evidence supports the opposite direction; N is the number of planned predictions; and R counts nonfailed predictions with unresolved references. UCR is therefore an incidence per planned prediction, not the unsupported proportion among correct predictions. We additionally report **hit caliber** $(U + D)/(C - R_c)$, the unsupported fraction among resolved correct predictions, together with sensitivity bounds, abstention rate, failure rate, and answered coverage. These quantities are reported jointly because abstention or technical failure can reduce N-based UCR without improving support. Because sufficient-negative evidence is structurally harder to establish, §4.1 also reports label-conditional corpus-level gaps.

Table 2: Judge coverage by language stratum. Fractions denote valid judgments over all cells.

Judge	Eng. (n = 2,168)	Chi. (n = 576)
Qwen	2,167/2,168	572/576
Gemini	1,666/2,168	—
Claude	568/2,168	267/576

Models and methods. We evaluate Gemini-2.5-flash-lite, GPT-5-mini, and Qwen-plus on 87 development cells from 76 participants under four prompting strategies, with one prediction per model–method–cell. **B0** directly predicts symptom presence with optional abstention. **B1** additionally requires a decisive patient quote and checks subject, time window, and scope. **B2** pairs forced binary prediction with a sufficient-context gate, abstaining when context is insufficient. **M** jointly elicits evidence frames, evidence state, and polarity, abstaining for non-sufficient states. Results are exploratory; prospective held-out evaluation with frontier models and repeated runs is ongoing.

4 Results

We present findings as a progressive audit in three layers. We begin by establishing the scale of the evidence gap: how much of the benchmark actually contains sufficient textual evidence for the symptom assessments it evaluates (§4.1)? The answer—82.8% of cells lack sufficient evidence—reframes every subsequent accuracy number. We then turn to model evaluation: given this evidence landscape, what does a "correct" prediction actually reflect (§4.2)? We show that models converge on accuracy while diverging on evidential grounding, with 69–97% of correct predictions unsupported regardless of prompting strategy. Finally, we audit the measurement apparatus itself (§4.3): agreement among judges is high but independence is low, perturbation tests fail, and reference choices shift 13% of labels—yet the evidence gap is too large for any of these limitations to explain away, a scale argument that bounds random reference error but not systematic error.

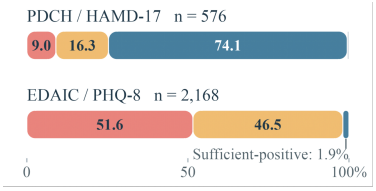

Figure 2: Within-source evidence composition. All source cells in each denominator.

4.1 How Much Evidence Do Depression Dialogues Actually Contain?

The fitted model-consensus reference assigns absent or partial evidence to **82.8%** of cells (95% CI: 82.1–83.5%) across all 2,744 main cells : 42.6% ABSENT (n = 1,170), 40.2% PARTIAL (n = 1,103), and only 17.1% SUFFICIENT (n = 468), all with positive polarity. There are no sufficient-negative or conflicting MAP assignments. The source-level contrast is stark: 74.1% of PDCH cells receive sufficient evidence versus only 1.9% of E-DAIC cells. Source, language, instrument, and preprocessing are confounded, so this contrast is descriptive—but it underscores how heavily the evidence landscape depends on interview format (Figure 2). Figure 3 shows the

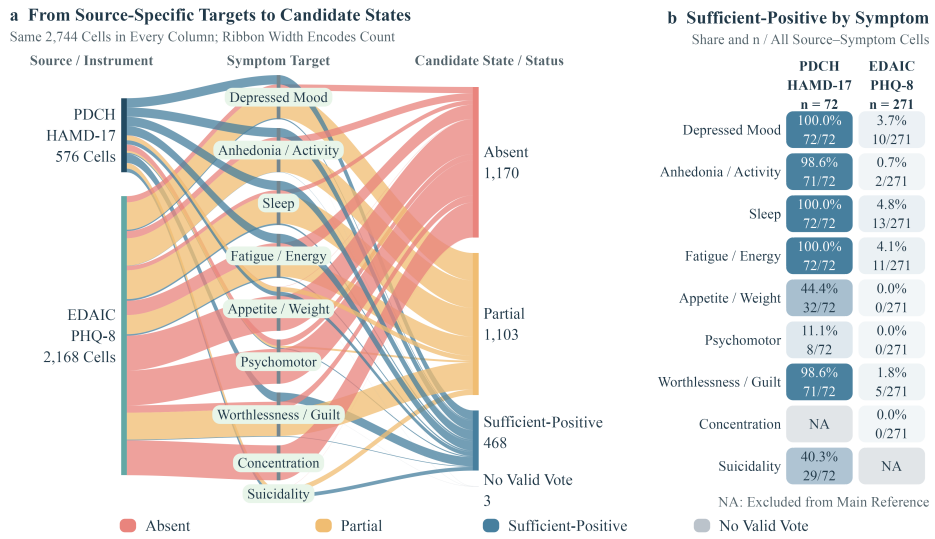

Figure 3: **Evidence landscape. (a)** Core collection; **(b,c)** state proportions overall and by source; **(d)** ribbon widths encode cell counts; **(e)** sufficient-positive shares per source–symptom target (NA = excluded). All states are model-consensus candidates.

full distribution. The candidate reference nonetheless retains the direction of the native labels while remaining far from them: sufficient evidence is assigned to 24.0% of positive-native-label cells versus 8.3% of negative-label cells, and mean native item scores increase monotonically across candidate states (absent 0.77, partial 0.91, sufficient 1.65); 100 sufficient-positive cells (3.6%) carry negative native labels, so divergence is not one-sided. Conditional on the native label, the lack of sufficient evidence is not confined to the negative side, where the negative-evidence contract is structurally harder (§3.2): 76.0% of positive-label cells lack sufficient evidence (95% CI: 73.9–78.1%), against 91.7% of negative-label cells (95% CI: 89.8–93.5%; difference -15.7 points, 95% CI: -19.3 to -11.9). Because the contract makes positive sufficiency easier to reach, the positive-side share is, if anything, conservative; the label–support gap therefore cannot be attributed solely to the asymmetry of negative evidence. Most assessment cells are evaluated against labels for which the dialogue provides no sufficient textual support. This scarcity is not a background condition—it directly determines the ceiling on evidence-grounded evaluation. What does a "correct" prediction on such cells actually reflect?

4.2 Can Models Get the Right Label Without Evidence?

The convergence paradox Under direct prompting (B0), all three models produce exactly $C = 52$ correct predictions out of 87 cells—yet their correct sets differ (intersection 42, union 64). Of the 42 jointly correct cells, only 8 have sufficient evidence aligned with the predicted label; 28 have partial or absent evidence, and 6 have sufficient evidence pointing the *opposite* direction. Models converge on accuracy but diverge on evidential grounding. The same $C = 52$ yields different apparent accuracy rates due to different answered denominators (Qwen 69 cells / 75.4%, Gemini 82 / 63.4%, GPT 83 / 62.7%). N-based UCR is more consistent: .483–.494, with hit caliber at 80.8–82.7% (Figure 4). Approximately four out of five "correct" predictions lack evidence support regardless of which cells a model answers. If direct prompting already produces this pattern, can more sophisticated strategies resolve it?

Can better prompting help? B1 (evidence-first) requires a decisive quotation and binding verification before judging. UCR does not improve (.471–.540): requiring citation cannot overcome evidence scarcity. **B2** (sufficiency gate) achieves the lowest UCR (.207–.402) but at steep coverage cost (34.5–67.8%), and 69.2–77.8% of retained correct predictions still lack support. **M** (structured output) is the most model-dependent: Gemini abstains on 1.1% of cells with 96.9% hit caliber, while GPT abstains on 47.1% with lower UCR (.368). Better prompting trades unsupported correctness for selective abstention without resolving the fundamental evidence scarcity (Table 3). A preregistered magnitude comparison cautions interpretation: hit-caliber estimates (0.692–0.969) bracket the reference-level not-sufficient prevalence (0.828), the pattern expected if model–method conditions differ mainly in

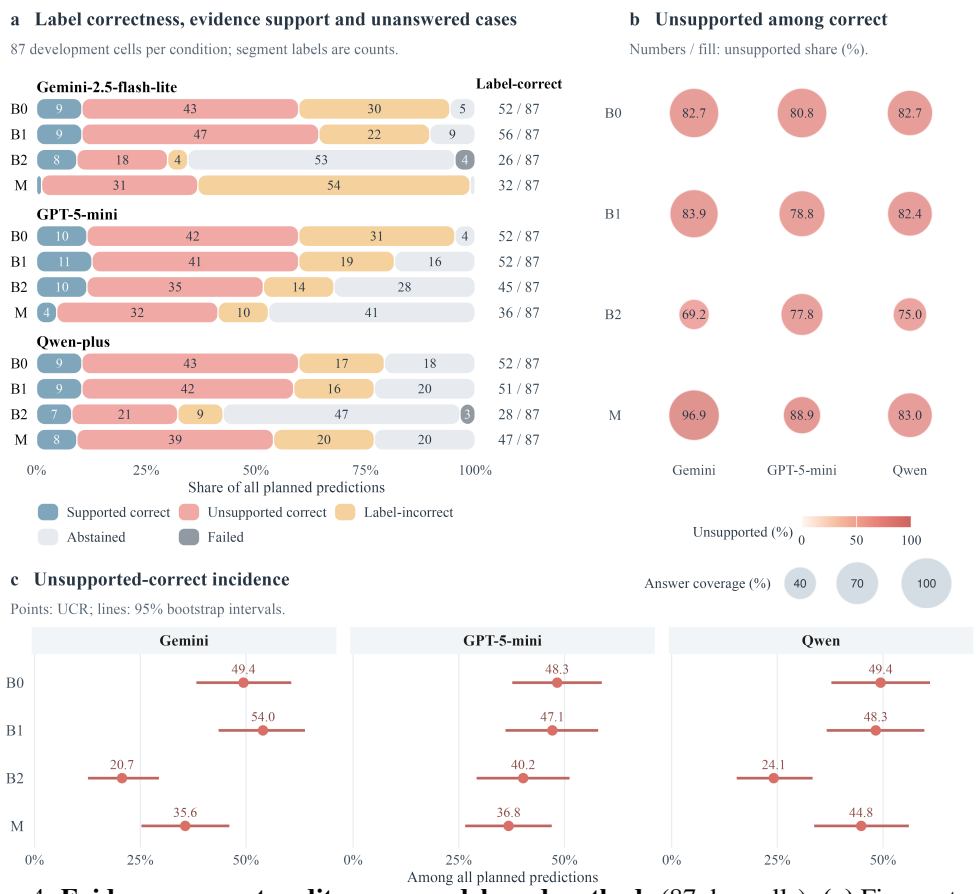

Figure 4: **Evidence-support audit across models and methods** (87 dev cells). **(a)** Five mutually exclusive outcomes; **(b)** UCR with 95% bootstrap intervals; **(c)** hit caliber vs. coverage (circle area = coverage).

Table 3: UCR and companion metrics on 87 development cells ($R = R_c = 0$), together with the agreement and correlation structure of the legacy LLM judge panel. Bracketed values are bootstrap 95% CIs conditional on the provisional reference. Failure is .000 except Gemini-B2 (.046) and Qwen-B2 (.034).

(a) Direct prompting (B0, B1)

Model	M.	UCR [CI]	Hit C.	Abst	Cov
Gem.	B0	.494 [.38,.61]	.827	.057	.943
Gem.	B1	.540 [.44,.64]	.839	.103	.897
GPT	B0	.483 [.38,.59]	.808	.046	.954
GPT	B1	.471 [.36,.58]	.788	.184	.816
Qwen	B0	.494 [.38,.61]	.827	.207	.793
Qwen	B1	.483 [.37,.60]	.824	.230	.770

(b) Selective / structured (B2, M)

Model	M.	UCR [CI]	Hit C.	Abst	Cov
Gem.	B2	.207 [.13,.29]	.692	.609	.345
Gem.	M	.356 [.25,.46]	.969	.011	.989
GPT	B2	.402 [.29,.51]	.778	.322	.678
GPT	M	.368 [.27,.47]	.889	.471	.529
Qwen	B2	.241 [.15,.33]	.750	.540	.425
Qwen	M	.448 [.34,.56]	.830	.230	.770

(c) Legacy judge panel quality

Agreement		Dependence	
Fleiss κ	.785	PR (core)	1.40
Obs. agr.	.866	PR (ext.)	1.31
Unanimous	73.1%	Top eig. share	84–87%
		$E\rho^2(5)$	.949

their exposure to a shared label–support misalignment rather than in evidence use. Tier-level paired contrasts, which this grid reserves for a frontier-model comparison, remain unexecuted.

4.3 How Robust Is the Evidence Measurement Itself?

These findings rest on the model-consensus reference, whose robustness we now examine from three angles: inter-judge independence, perturbation sensitivity, and reference-method agreement.

Agreement versus independence A legacy panel of five LLM protocols annotated 5,283 cells across 587 transcripts. Agreement is high: Fleiss' $\kappa = .785$, observed agreement .866, 73.1% unanimous (Table 3) (c). However, the eigenvalue participation ratio yields only **1.40**—five judges contribute roughly the information of 1.4 independent raters, with 84–87% of eigenvalue mass on the first component. G-theory confirms: protocol variance is 1.0% of total; $E\rho^2 = .949$ is high because judges barely disagree, not because each independently validates the evidence.

Table 4: Perturbation anchors by failure mode. Entries are successes/total, no condition passes the gate.

Mode	Failure mode	Anchors	Claude	Gemini	Qwen
F1	Subject binding	P4-other	2/11	0/11	0/11
F2	Temporal binding	P5, P4-ow, P8	12/64	16/64	0/64
F3	Assertion scope	P4-wd, P3	1/35	10/35	0/35
F4	Withdrawal	P2, P4-wn	4/36	6/36	0/36
F5	Aggregation	P7, P1	6/50	2/50	2/50
F6	State projection	P6a, P6b, P8	17/61	20/61	0/61

Despite low independence at the panel level, protocol choice matters for individual cells: P2 and P5 assign assessability at 3.3–3.4× the odds of P1 and P3 ($\chi^2 = 1{,}051.5$, $p < 10^{-200}$), and 47 of 57 tied cells change assignment under MFRM correction. An MTMM analysis yields convergent-validity ratio 2.39 (same-dimension $r = .738$ vs. different-dimension $r = .309$). On the new three-judge panel, an MTMM decomposition gives convergent 0.514 versus discriminant 0.235 (ratio 2.18; centered 0.360 vs. -0.025), with the language blocks closer than in the legacy panel (English 0.473, Chinese 0.425)—trait structure that is not explained by a judge-style factor, though convergent correlations are not independent validation. Full decompositions are in Appendix B.

Perturbation stress tests Eight perturbation classes spanning six failure modes test whether judges respond predictably to controlled evidence changes (gate: ≥85% accuracy, ≥75% posterior lower bound). No condition passes in either run (Table 4). The highest hit rate is 20/61 (32.8%). Initial failures included prompt–validator mismatches and provider errors; after repair, semantic accuracy remains below threshold.

Where do missing sufficient-negative judgments come from? Because the reference contains no sufficient-negative MAP assignments, we adjudicate the denial-insertion class (P3, expected SUFFICIENT-negative) separately under the repaired instrument. The judge families split: Gemini returns sufficient-negative on 10/16 valid P3 judgments, Claude on 0/5; across all 269 valid anchor votes, 36 are sufficient-negative, against 9 under the initial run. The construct is thus producible by one judge family and absent from another under identical inputs, locating the main cause in judge capability rather than the construct being unreachable. A residual mismatch sits at the CONFLICTING boundary: on contradiction anchors (P7), Gemini resolves the conflict as denial overriding evidence (SUFFICIENT-negative on 10/12 valid judgments) where the contract expects CONFLICTING, and Claude returns CONFLICTING on only 1/13. These counts are small, and Qwen's P3/P7 votes are missing, so the adjudication is directional.

Reference sensitivity RE-LCA fusion covers 99.9% of cells versus 77.3% for plurality, but they disagree on 281/2,120 jointly labeled cells (13.3%). Anchor validation does not favor fusion (67–72% error vs. 55–64% for plurality, McNemar $p < .02$). A one-hot mean baseline is identical to plurality unit by unit (2,744/2,744 cells): with at most three categorical votes, the mean judge degenerates to the modal judge, so reference construction chooses between abstention and model-based resolution of split votes, not between naive and sophisticated averaging. Plurality itself determines six sufficient-negative cells that fusion assigns none, so the zero sufficient-negative count is a property of the fusion fit rather than of every aggregator. A first entropy calibration attempt met the bin-occupancy criterion in three of five bins but fell at the support edge (223/246 voted views carry a single vote; error rates non-monotone in entropy), so no single-point unresolved threshold is adopted. Adaptation of an existing human-annotation subset yields 29 overlapping cells with agreement 27.6–79.3% depending on an unresolved temporal-rule assumption. A separate blind 120-cell human-annotation study (§5) confirms substantial agreement with both RE-LCA and plurality references and preserves the main evidence-gap findings under re-scoring.

Robustness despite measurement limits The measurement apparatus is imperfect—but the evidence gap is too large to be an artifact. Moving cells from absent to sufficient at the disagreement rate (10–13%) would still leave the majority unsupported. UCR values of .207–.540 across 12 conditions, with hit caliber consistently above 69%, represent a pattern too pervasive for any reference choice to explain away. This scale argument bounds random reference error; it does not substitute for reference validity, because reference error could be systematic rather than random (§7). The blinded human validation in §5 provides an independent check: re-scoring against the human reference preserves the main findings.

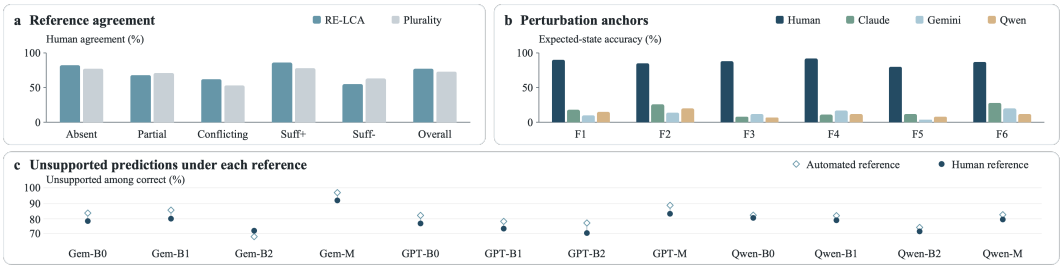

Figure 5: **Human validation. (a)** RE-LCA and plurality show substantial agreement with blinded human consensus, with remaining disagreement concentrated near difficult evidence boundaries. **(b)** Humans reliably recover the intended states of controlled perturbation anchors, whereas LLM judges do not. **(c)** Re-scoring model predictions against the human reference preserves the high unsupported-among-correct rates across model–method conditions.

5 HUMAN VALIDATION

Two blinded annotators independently labeled 120 assessment cells, with disagreements adjudicated by a third rater. RE-LCA showed 76.5% overall exact agreement with human consensus, compared with 73.1% for plurality, with lower agreement concentrated near the CONFLICTING and SUFF- boundaries (Fig. 5a). Humans also recovered the intended perturbation states at 80–92% across six failure modes, whereas individual LLM judges achieved only 4–28% (Fig. 5b). Most importantly, re-scoring the overlapping model audit against the human reference preserved the main finding: unsupported-among-correct averaged 79.1%, versus 81.8% under the automated reference, with a mean absolute change of 3.1 percentage points and a maximum change of 4.9 points across 12 model–method conditions (Fig. 5c).

6 DISCUSSION & CONCLUSION

Our results show that label correctness and textual support are distinct properties. Across EVIDIAL, many assessment cells lack sufficient evidence in the observable dialogue, and many label-correct predictions remain unsupported across models, prompting strategies, and reference choices. This does not reveal how models reason internally; rather, it shows that accuracy alone cannot establish evidence-grounded assessment. The strong variation across sources further indicates that evidence support depends on the *benchmark–instrument–transcript interface*, including how symptoms are elicited, mapped, and preserved in the evaluated text. Benchmark evaluation should therefore treat evidence support as a separate axis alongside label accuracy. This distinction extends beyond depression dialogue whenever evaluation targets contain information not recoverable from the model-visible input. Although EVIDIAL does not establish clinical validity and its references remain subject to measurement uncertainty, it makes a basic evaluation gap explicit: matching the benchmark label and being justified by the available evidence are distinct evaluation dimensions, and reliable behavioral assessment should measure both.

7 LIMITATIONS & FUTURE DIRECTIONS

Construct scope. EVIDIAL measures whether the transcript provides sufficient textual evidence for a symptom assessment, not whether the patient clinically has that symptom. Its references therefore characterize evidence availability rather than diagnostic truth. **Dataset composition.** EDAIC contributes 79% of the main reference cells, and language, source, and clinical instrument are confounded across datasets. Consequently, source-level differences cannot be uniquely attributed to any one of these factors. **Model-audit scope.** The development audit in §4.2 uses 87 cells, lightweight models, and single runs per condition. These experiments establish the presence of an evidence gap but are not intended to support definitive model rankings; broader held-out evaluation with stronger models and repeated runs remains future work. **Reference validity.** The automated reference remains imperfect: coverage is uneven, and some evaluator and judge families overlap. A blinded 120-cell human validation (§5) confirms substantial agreement with the automated reference (76.5% for RE-LCA) and shows that humans reliably recover perturbation anchor states (80–92%) where LLM judges do not; however, this covers a subset of cells and does not constitute a complete audit across all dimensions. Negative evidence is also harder to establish under the current contract, making negative-direction estimates more sensitive to judge behavior. These limitations motivate treating the reference as a calibrated measurement instrument rather than definitive ground truth.

ETHICS STATEMENT

This work uses secondary analysis of existing depression-dialogue corpora. PDCH transcripts are licensed under CC BY-NC 4.0; EDAIC access is request-gated through the original data holders; TalkDep transcripts are publicly available. TalkDep profiles are simulated, whereas the main reference uses clinical interview records. Public availability does not establish uniform redistribution rights, and any release must respect the permissions and access conditions of the original data sources. The study additionally includes a blinded human-validation experiment (§5). Two independent annotators labeled 120 assessment cells under the four-state evidence taxonomy, with disagreements adjudicated by a third rater. These annotations were used to evaluate the agreement of the automated reference with human judgments and to test whether the main audit conclusions were preserved under a human-derived reference. The annotators did not provide clinical diagnoses; their task was restricted to judging textual evidence under the predefined annotation contract. A separate legacy annotation subset covers 99 PDCH interviews in a suicidality-focused task. The available records for this legacy annotation do not sufficiently document institutional approval, compensation, or support arrangements to make claims about those procedures here. The evidence-state annotations produced by EviDial are explicitly not diagnostic labels: they characterize the availability and consistency of textual evidence, not participants' underlying clinical status. In particular, ABSENT or PARTIAL should not be interpreted as evidence that a participant does not have the corresponding symptom or condition; they indicate only that the available transcript does not provide sufficient evidence under the specified evidence contract.

REPRODUCIBILITY STATEMENT

The accompanying archive contains model outputs, usage records, prompts, and analysis scripts. Audited model responses predate the current reference export and were retrospectively rescored. Fixed seeds identify the bootstrap and fusion runs, and input hashes support artifact matching. For the human validation (§5), the paper specifies the sampled assessment cells, blinded independent annotation procedure, third-rater adjudication, agreement computation, and rescoring of the overlapping model audit against the human-derived reference. The associated analysis code reproduces the reported agreement statistics, perturbation-state recovery results, and condition-level changes under human rescoring. The endpoint runner checks only that two supplied freeze strings are nonempty; it does not enforce a complete hash chain or chronological freeze order. Legacy numerical errors and incomplete recovery checks are documented in Appendix C; invalid posterior summaries and absolute information criteria are excluded as evidence. Cross-platform bit-identical reproduction has not been established. Release of raw transcripts and annotation materials derived from restricted corpora remains subject to the permissions of the original data sources. Code, analysis scripts, and reproducibility materials are available in an anonymized repository at `https://anonymous.4open.science/r/evidial-code-F024/README.md`.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and polishing, and to support literature retrieval and discovery of potentially relevant prior work. All AI-assisted text was reviewed and revised by the authors, and all references and factual claims obtained with AI assistance were independently verified against the original sources. LLM judges evaluated as part of the experiments were study subjects rather than sources of the human-validation reference. The authors take full responsibility for the final content of this work.

REFERENCES

Paul S. Albert and Lori E. Dodd. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics*, 60(2):427–435, 2004.

Ron Artstein and Massimo Poesio. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

Robert L. Brennan. *Generalizability Theory*. Springer-Verlag, New York, 2001.

Donald T. Campbell and Donald W. Fiske. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2):81–105, 1959.

Pengfei Cao, Yuanzhe Zhang, Chenxiang Zhang, Wei Chen, Yan Liu, Shuang Xu, et al. A multimodal depression consultation dataset of speech and text with HAMD-17 assessments. *Scientific Data*, 12:1577, 2025. doi: 10.1038/s41597-025-05817-9.

Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310, 2001. doi: 10.1006/jbin.2001.1029.

Junhyuk Choi, Sohhyung Park, Chanhee Cho, Hyeonchu Park, and Bugeun Kim. Diagnosing the reliability of LLM-as-a-judge via item response theory. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2602.00521.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Arun Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. 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)*, 2023.

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

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. The distress analysis interview corpus of human and computer interviews. In *Lrec*, volume 14, pp. 3123–3128. Reykjavik, 2014.

Max Hamilton. The hamilton rating scale for depression. In *Assessment of depression*, pp. 143–152. Springer, 1986.

Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics*, 42(5):839–851, 2009. doi: 10.1016/j.jbi.2009.05.002.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kuper, Idan Szpektor, Noam Slonim, et al. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2022.

Takehiro Ishikawa and Jon Duke. A multi-probe audit of clinical-interview depression detection benchmarks. *arXiv preprint arXiv:2605.23977*, 2026.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2411.06037.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 5684–5696, 2020.

Shreyas KC. Babeljudge: Measuring LLM-as-a-judge reliability across languages and agent trajectories. *arXiv preprint arXiv:2606.22329*, 2026.

Seungone Kim, Juyoung Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Guneet Kohli. Nine judges, two effective votes: Correlated errors undermine LLM evaluation panels. *arXiv preprint arXiv:2605.29800*, 2026.

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. doi: 10.1016/j.jad.2008.06.026.

John Michael Linacre. *Many-faceted Rasch measurement*. PhD thesis, The University of Chicago, 1989.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1906–1919, 2020.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Yongyi Qu, Ming Tan, and Michael H. Kutner. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*, 52(3):797–810, 1996.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, 2023.

Ying Shen, Huiyu Yang, and Lin Lin. Automatic depression detection: An emotional audio-textual corpus and a GRU/BiLSTM-based model. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. doi: 10.1109/ICASSP43922.2022.9746569.

Dan Song, Won-Chan Lee, and Hong Jiao. Exploring LLM autoscoring reliability in large-scale writing assessments using generalizability theory. *arXiv preprint arXiv:2507.19980*, 2025.

Hoyun Song, Migyeong Kang, Jisu Shin, Jihyun Kim, Chanbi Park, et al. Mentalbench: A DSM-grounded benchmark for evaluating psychiatric diagnostic capability of large language models. *arXiv preprint arXiv:2602.12871*, 2026.

Veerendra Kumar Sunkavalli. LLM judges as raters: A pre-registered audit of severity, halo, reliability, and version instability in LLM essay scoring on public corpora. *arXiv preprint arXiv:2608.29517*, 2026.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. Judgebench: A benchmark for evaluating LLM-based judges. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.12784.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

Yang Wang, Jing Huang, Lingyue Du, Yan Guo, Yang Liu, and Ran Wang. Evaluating large language models as raters in large-scale writing assessments: a psychometric framework for reliability and validity. *Computers & Education*, 2025.

Shihao Xu, Tiancheng Zhou, Jiatong Ma, Yanli Ding, Yiming Yan, et al. Lingxidiagbench: A multi-agent framework for benchmarking LLMs in chinese psychiatric consultation and diagnosis. *arXiv preprint arXiv:2602.09379*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 2023.

A Source Corpus Details

Table 5 provides descriptive statistics for the three source corpora, recomputed from the processed unified input.

Table 5: Corpus descriptive statistics. A segment is one stored speaker-turn record, not a full clinician–patient exchange; score ranges are observed rather than theoretical.

Source	Transcripts	Median segments	Segment range	Median total	Total range
PDCH	99	276	159–642	15	1–35
EDAIC	271	60	21–129	6	0–23
TalkDep	12	65.5	34–73	18.5	5–40

PDCH contains real psychiatric diagnostic interviews conducted at a Chinese hospital, with 72 unique participants (some interviewed more than once). Its 37 reconstructed totals are sums of item scores. EDAIC consists of semi-structured distress interviews from the Extended Distress Analysis Interview Corpus, scored on the PHQ-8. TalkDep uses LLM-simulated patient personas following BDI-II profiles. The 370 clinical transcripts represent 343 participants; the other 12 transcripts represent 12 virtual personas.

For comparison, the legacy diagnostic pool contains 587 transcripts: PDCH 99, EDAIC 271, TalkDep 12, D4 60, LingxiDiag 52, MDD5K 51, and EATD 42. All nine dimensions enter that analysis, yielding $587 \times 9 = 5{,}283$ cells. These are distinct denominators from the current main reference and model-output sample.

B Generalizability Theory and MFRM Details

G-theory model. We decompose vote-level observations y_{upk} for participant p, unit (cell) u nested in p, and protocol k:

$$y_{upk} = \mu + \pi_p + \nu_{u:p} + \tau_k + \varepsilon_{upk} \tag{3}$$

where π_p and $\nu_{u:p}$ are mean-zero random effects, τ_k are fixed protocol effects, and ε_{upk} is residual error confounding the unit$\times$protocol interaction with pure error (single observation per cell-protocol combination). Here y_{upk} measures binary assessability, and the participant grouping includes source-recorded virtual or synthetic identities. The legacy sample comprises 5,283 cells from 587 transcripts and seven sources, with 560 source-recorded identities. The 99 PDCH interviews correspond to 72 participants, with 27 participants contributing two interviews.

Variance components are estimated via method-of-moments (Type-I SS) with nested expected-mean-square coefficients. Protocol mean effects are removed before the nested decomposition, and the protocol component is the observed variance across the five protocol means. For measure A (all votes, $n = 26{,}415$): $\hat{\sigma}^2_\pi = 0.054$, $\hat{\sigma}^2_{\nu:p} = 0.131$, $\hat{\sigma}^2_\tau = 0.002$, $\hat{\sigma}^2_\varepsilon = 0.050$.

The relative generalizability coefficient for $k = 5$ fixed protocols:

$$E\rho^2(5) = \frac{\hat{\sigma}^2_\pi + \hat{\sigma}^2_{\nu:p}}{\hat{\sigma}^2_\pi + \hat{\sigma}^2_{\nu:p} + \hat{\sigma}^2_\varepsilon/5} = 0.949 \tag{4}$$

This calculation treats residual contributions as exchangeable under replication of the fixed protocol set. The secondary absolute-decision coefficient also includes the observed protocol component in the error term and equals .946693. Neither coefficient supplies an external criterion for the correctness of a cell's evidence label.

MFRM calibration. The Many-Facet Rasch Model estimates protocol severity parameters via conditional maximum likelihood. For each binary submeasurement, the Rasch specification is $\text{logit}\, P(y_{uk} = 1 \mid \theta_u, \beta_k) = \theta_u - \beta_k$, with $\sum_k \beta_k = 0$. The assessability fit uses all 26,415 votes; the conditional-direction fit uses 16,216 observable votes, with 1,508 all-not-assessable cells contributing no direction observations. Protocol log-odds estimates describe assessability: P2 and P5 have odds multipliers of 3.3–3.4$\times$ relative to a zero-severity protocol; P1 and P3 have multipliers of 0.34–0.38$\times$ for an assessable category. The conditional likelihood ratio test for protocol homogeneity

yields $\chi^2 = 1{,}051.5$ ($df = 4$, $p < 10^{-200}$). The assessability odds multipliers $\exp(-\beta_k)$ for P1–P5 are .3431, 3.2673, .3766, .6905, and 3.4313, respectively. The corresponding conditional-direction multipliers are 1.1539, .9959, 1.3493, .4415, and 1.4606.

Frozen plurality yields 2,378 positive, 869 negative, and 2,036 not-assessable assignments. MFRM yields 2,409, 875, and 1,999, respectively. Of 57 tied cells under plurality, 47 change their assignment under MFRM calibration. These differences quantify disagreement between aggregation procedures. Against a selected 45-cell model re-adjudication, both plurality and MFRM agree on 27 cells; against two model re-readers on an 18-cell set, both agree on 13 cells per reader. Among 22 uniquely joinable cells from a separate annotation pilot, agreement with the mapped pilot plurality is $13/22$ for legacy plurality and $12/22$ for MFRM. These small comparisons involve selection and cross-version mappings and do not establish reference-quality improvement.

C Legacy RE-LCA Analyses and Recovery Checks

The random-effects latent class model assigns each cell u a latent legacy assessment category $Z_u \in \{\text{observable-positive}, \text{observable-negative}, \text{not-assessable}\}$ with a cell-level random effect on protocol emissions. The exploratory models are fitted directly to the legacy three-category votes; the covariate extension changes class priors through additive source and dimension-group terms. MFRM assignments are loaded afterward for comparison. This analysis is separate from construction of the current four-state reference.

The archived implementation has two numerical limitations relevant to interpretation. Its posterior summary function rescales each class separately over integration points without preserving the class-specific scaling factors; the resulting posterior probabilities, entropy, MAP summaries, and ICL values are therefore excluded from our evidence. Its discrete integration also omits a constant grid-width factor when reporting likelihoods, affecting absolute likelihood and information-criterion values. We do not use the archived absolute AIC/BIC values as model-selection evidence. The current reference implementation is evaluated separately; the legacy results cannot validate it by themselves.

Synthetic parameter recovery. The available recovery experiments comprise five scenarios, each using one simulated dataset of 5,283 cells. They assess parameter error against declared tolerances. Three scenarios concern the homogeneous model: realistic signal and zero random-effect variance meet their parameter tolerances; a low-signal scenario fails, with maximum class-prevalence error .1029 against tolerance .05, maximum population-emission error .1241 against .07, and random-effect standard-deviation error .3228 against .30.

Two additional scenarios concern the covariate extension. A crossed source–language design meets its recovery criteria. A scenario preserving the observed confounded source structure fails the source-specific prevalence criterion: maximum error is .0603 against tolerance .03. These checks provide limited evidence about parameter recovery in selected simulated settings. They do not estimate repeated-simulation class-assignment accuracy.

The legacy fitted-model simulation checks also show incomplete fit: the homogeneous specification fails the recorded stratified criterion in 15 of 16 strata; the source-plus-dimension-group extension fails in 11 of 16; and an exploratory source-plus-nine-dimensions extension fails in three of 16. These results limit substantive interpretation of the legacy latent-class models and their use for reference validation.

D MTMM and Language Comparisons

Multitrait–Multimethod analysis. Following Campbell & Fiske (1959), we compute mean same-trait, different-method correlations and mean different-trait, same-method correlations for legacy binary assessability. Results: same-trait correlation $r = 0.738$, different-trait correlation $r = 0.309$, ratio $= 2.39$. After participant-mean centering: the corresponding correlations are $r = 0.620$ and $r = -0.069$. Centering changes the correlation structure, and the negative value should be interpreted accordingly. These descriptive correlations do not independently confirm construct validity.

The matrix uses nine symptom dimensions as traits and five protocols as methods. Repeated observations for the same identity, dimension, and protocol are averaged, giving 560 rows and 45 variables.

The mean same-trait, different-method correlation averages 90 pairs; the different-trait, same-method mean averages 180 pairs. Within the Chinese-language block, the uncentered correlations are .8404 and .5118 across 277 identities; within the English-language block, they are .5479 and .0538 across 283 identities. The Chinese block combines PDCH, D4, EATD, LingxiDiag, and MDD5K; the English block combines EDAIC and TalkDep.

Spectral summary. For correlation-matrix eigenvalues $\lambda_1, \ldots, \lambda_5$, the participation ratio is $(\sum_k \lambda_k)^2 / \sum_k \lambda_k^2$. Its interpretation concerns the raw vote vectors. For example, five raters that independently flip a balanced binary true label with probability .05 each have marginal pairwise vote correlation .81. Their population participation ratio is $5/(1 + 4 \times .81^2) = 1.380$. Thus a low participation ratio can arise from shared signal even with conditionally independent errors. Assessing error dependence requires additional assumptions or an external reference.

Language comparison. The legacy language analysis was a descriptive parameter screen using separately fitted binary assessability models and 25 parametric bootstrap datasets per block. Source, instrument, language, and sample provenance are confounded. Moreover, one flagged contrast involved random-effect standard deviations whose reported bootstrap standard errors were zero under a discrete parameter search. We therefore make no formal measurement-invariance claim from that screen and do not interpret its estimated assessability prevalence as symptom-positive prevalence or as a causal language effect.

E Perturbation Class Definitions

The eight perturbation classes and their expected final states are given in Table 6. These are the construction targets used by the executed scoring scripts, not independently verified labels for the original interviews.

Table 6: Perturbation classes with expected evidence-state targets. P4 inserts distractors while preserving the original supporting evidence; its four variants test invariance under each binding operation.

Class	Manipulation	Expected final state	Failure mode
P1	Deletion (evidence removed)	ABSENT	F5: Aggregation
P2	Withdrawal (later retraction)	ABSENT	F4: Withdrawal
P3	Denial insertion	SUFFICIENT -neg	F3: Assertion scope
P4-ow	Out-of-window distractor	SUFFICIENT -pos	F2: Temporal
P4-op	Other-person attribution	SUFFICIENT -pos	F1: Subject
P4-wd	Weak denial	SUFFICIENT -pos	F3: Assertion scope
P4-wn	Withdrawn statement	SUFFICIENT -pos	F4: Withdrawal
P5	Temporal boundary test	ABSENT	F2: Temporal
P6a	Half-support reduction	PARTIAL	F6: State projection
P6b	Partial evidence only	PARTIAL	F6: State projection
P7	Contradiction insertion	CONFLICTING	F5: Aggregation
P8	Window-incomplete denial	PARTIAL	F2/F6: Temporal/Projection

The initial run contains 319 views; the update contains 343. The initial scoring script treats P6b as descriptive, while the updated script includes it in the gated set. The reported F1–F6 groups aggregate these variants and are not independent because P8 appears twice. Scoring tests agreement with the expected final state rather than an estimated original-to-perturbed causal effect. Human assessment of perturbation anchors was completed as part of the blinded validation study (§5), with humans recovering intended states at 80–92% across failure modes. Generator-family exclusivity ensures the model generating perturbations is from a different family than the judges evaluating them.

F Reference Fusion Specification

Let Y_{uj} be the categorical vote of judge j on cell u, and let $\mathcal{O}_u$ contain only judges with valid votes on that cell. Write $\pi_k(\boldsymbol{x}_u)$ for the five-state softmax prevalence model and $e_j(y \mid k, b)$ for the categorical

emission probability. The random effect b_u is shared by judgments on the same cell; effects for different cells are integrated separately. Using a zero-mean Gaussian density ϕ_σ restricted to the numerical integration interval, the observed-vote likelihood is

$$\mathcal{L}(\theta) = \prod_{u:\,|\mathcal{O}_u|>0} \sum_{k=1}^{5} \pi_k(\boldsymbol{x}_u) \int_{-6}^{6} \left[\prod_{j \in \mathcal{O}_u} e_j(Y_{uj} \mid k, b) \right] \frac{\phi_\sigma(b)}{\int_{-6}^{6} \phi_\sigma(t)\, dt}\, db. \tag{5}$$

The design vector contains an intercept, a joint source/language indicator, and a dimension-group indicator, without an interaction term. Computation replaces each integral with a sum over $b_g = -6 + 0.15g$, $g = 0, \ldots, 80$, using normalized Gaussian weights

$$w_g(\sigma) = \frac{\exp[-b_g^2/(2\sigma^2)]}{\sum_{h=0}^{80} \exp[-b_h^2/(2\sigma^2)]}. \tag{6}$$

Thus, b_u is represented by a finite-grid approximation to $\mathcal{N}(0, \sigma^2)$, not a participant-level intercept. The fit additionally incorporates the emission pseudo-counts described in §3.3. Posterior state probabilities sum jointly over the grid and are normalized across states.

The production likelihood sum omits a constant grid-width factor, so its stored log likelihood differs from the normalized-grid expression by $-n \log(0.15)$ for n voted cells. This constant does not change posterior assignments or comparisons on the same grid and sample. Absolute stored likelihood or information-criterion values are not reported as evidence here. The current posterior routine preserves the relative scaling across states, unlike the legacy routine discussed in Appendix C.