

An Honesty-Aware Multi-Agent Framework for Synthetic Psychiatric Intake Data Generation

Xinyuan Zhang*, Chang Dao*, Zijian Wang, Yechao Pang and Juexiao Zhou

The Chinese University of Hong Kong, Shenzhen

Corresponding author: juexiao.zhou@gmail.com

Abstract

Psychiatric intake combines self-report scales with clinician-rated interviewing, but public full-trace data are scarce and most synthetic resources assume truthful reporting. We present an honesty-aware multi-agent framework for psychiatric intake simulation that explicitly models honest reporting, concealment, and malingering under asymmetric information. Starting from structured patient profiles, five role-specific agents generate self-report scale outputs, strategy-guided clinician-rated interviews, honesty inference from the interview transcript alone, and honesty-aware diagnostic fusion over four common psychiatric domains: major depressive disorder (MDD), anxiety disorder (AD), bipolar disorder (BD), and attention-deficit/hyperactivity disorder (ADHD). Using normalized clinical and healthy source profiles, we construct a dataset of 2,008 complete intake sessions with 58,918 dialogue turns, linking profiles, scales, strategy traces, inferred honesty, and final diagnostic reports. Experiments show that assigned reporting conditions are expressed in generated patient behavior and can be recovered from the interview transcript alone, with condition-level inference accuracy between 60.1% and 94.8%. Under hidden-information constraints, the best LLM achieves 85.4% exact recovery of synthetic reference domains with 0.99 severity MAE and zero detected leakage. Honesty-aware fusion reduces deviation from benchmark reference labels relative to self-report alone, and strategy-guided interviewing attains higher accuracy with fewer evaluator questions than item-by-item questioning. These results support honesty-aware synthetic intake data as a controlled research benchmark, not a clinically validated pre-assessment system.

1 Introduction

Rising demand for mental-health services has motivated large language model (LLM)-based applications in risk and depression detection (Kumar et al., 2025; Cao et al., 2025; Weber et al., 2025), questionnaire-based screening (Ravenda et al., 2025; Rosenman et al., 2024a), synthetic counseling and clinical dialogues (Yao et al., 2022a; Wang et al., 2024a; Chen et al., 2023; Lee et al., 2024), and digital patients for diagnostic interactions (Wang et al., 2024c; Luo et al., 2025; Tu et al., 2025; Louie et al., 2024). Psychiatric intake is a particularly relevant setting: it is standardized and protocol-driven, combining demographic and chief-complaint collection, self-report scales, and semi-structured interviews that inform clinician-rated scales (McQuiston and Zinns, 2019; Sheehan et al., 1998). This structure provides a controlled setting for studying LLM-agent interviewing.

Yet publicly available full psychiatric intake records remain scarce due to privacy and ethical constraints (Guo et al., 2024). Existing resources often cover only parts of the intake pipeline, such as isolated questionnaires, counseling-style dialogues, or role-played patient interactions. More importantly, most synthetic psychiatric dialogue datasets assume truthful self-report. This assumption limits evaluation in settings where patients may strategically distort their reports through symptom concealment or malingering, both of which have been extensively studied in psychometrics and clinical assessment (Barbosa-Torres et al., 2023; Drob et al., 2009; Mittenberg et al., 2002; Bril-Barniv et al., 2017).

This creates a gap for synthetic psychiatric intake data that jointly models patient profiles, self-report scales, semi-structured interviews, clinician-rated scales, honesty conditions, and diagnostic summaries. Such data enable controlled

*Equal contribution.

evaluation of whether a simulated assessment system can recover assigned reporting conditions, adapt its questioning strategy, and integrate conflicting evidence under asymmetric information.

We address this gap with an honesty-aware multi-agent framework for controlled psychiatric intake simulation. Our system orchestrates five agents (Assessor, Patient, Evaluator, Strategy, and Diagnostician) to generate complete intake sessions from structured patient profiles. The Patient Agent is the only agent with access to hidden diagnosis, severity, personality, and honesty conditions, while the Strategy Agent must infer honesty from visible dialogue evidence alone. The framework therefore creates an asymmetric-information setting for studying concealment, malingering, and honesty-aware diagnostic fusion.

Our contributions are:

- We introduce the first honesty-aware psychiatric intake simulation framework that models both symptom **concealment and malingering** in scale-based assessment and clinician-rated interviews.
- We construct a controlled synthetic benchmark of **2,008 complete psychiatric intake sessions**, linking patient profiles, scale outputs, dialogues, strategy traces, honesty inference, and diagnostic reports.
- We show that, within the synthetic benchmark, strategy-guided interviewing supports honesty inference and diagnostic fusion with far fewer evaluator questions than item-by-item scale administration, providing a testbed for interaction-efficient interviewing.

2 Related Work

2.1 Medical dialogue systems and psychiatric assessment

Medical dialogue research has progressed from broad consultation corpora to clinically structured diagnostic data. MedDialog dataset provides large-scale Chinese and English medical consultations (Zeng et al., 2020), while D⁴ dataset focuses on depression-oriented doctor–patient conversations grounded in diagnostic criteria and psychiatrist annotation (Yao et al., 2022b). More recent work moves toward synthetic and LLM-assisted clinical dialogue generation, including NoteChat for note-conditioned patient–physician conversations

(Wang et al., 2024a) and MDD-5k for long mental-disorder diagnostic dialogues synthesized from anonymized cases (Yin et al., 2025). LLMs have also been used to interpret psychiatric interviews and support conversational diagnosis (Rosenman et al., 2024b; Tu et al., 2025), while mental-health dialogue systems often focus on emotional support or counseling (Liu et al., 2021; Chen et al., 2023; Qiu et al., 2024). In contrast, our work centers on psychiatric assessment dialogue simulation under controlled reporting conditions rather than response generation, counseling, or diagnosis prediction alone.

2.2 Multi-agent LLM frameworks

Multi-agent LLM systems use role specialization and coordination to decompose complex tasks. CAMEL introduced role-playing communicative agents (Li et al., 2023), AutoGen provided a flexible framework for multi-agent conversation and tool use (Wu et al., 2024), and MetaGPT organized collaboration through standardized role-based procedures (Hong et al., 2024). Related work also applies multi-agent interaction to evaluation and reasoning, such as ChatEval’s debate-based text evaluation (Chan et al., 2024). In mental-health settings, MAGneT decomposes counseling-session generation into specialized counseling-technique agents (Mandal et al., 2025), and AutoCBT proposes a cognitive behavioral therapy-oriented autonomous multi-agent framework (Xu et al., 2025). These studies show the value of specialization, but they do not model a full psychiatric assessment workflow with distinct patient, assessor, strategy, evaluator, and diagnostician roles.

2.3 LLMs as simulators of human and patient behavior

A growing body of work studies whether LLMs can simulate human behavior and role-consistent interaction. Generative Agents introduced memory, reflection, and planning for believable social behavior (Park et al., 2023), while other work treats LLMs as simulated participants or benchmarks their role-playing and social intelligence abilities (Argyle et al., 2023; Aher et al., 2023; Wang et al., 2024b; Zhou et al., 2024). In mental-health contexts, PATIENT-Ψ and Roleplay-Doh are especially relevant because they simulate patient-like behavior for training and practice (Wang et al., 2024c; Louie et al., 2024). Clinical studies of malingering further show that distorted

symptom reporting can affect interview validity (Barbosa-Torres et al., 2023). Our work builds on this line by treating patient simulation as a controlled psychiatric-assessment problem, explicitly varying honest reporting, concealment, and malingering within one framework.

3 Method

3.1 Task and Framework Overview

Figure 1 presents the overall framework. Given a structured profile, the framework generates self-report scales, a strategy-guided clinician-rated interview, honesty inference, and fused diagnostic outputs. To support reproducibility, we provide a public repository containing the implementation, prompts, workflow configurations, evaluation scripts, toy examples, and aggregate results and anonymized human-evaluation materials: <https://github.com/changdao565/honesty-aware-psychiatric-intake>. The complete sessions, processed profiles, case-level manifests, and provider traces are not part of the public release.

3.2 Patient Profile Construction

The input to our framework is a structured patient profile. We normalize heterogeneous source records into a unified schema before agent-based generation.

Source profile datasets. We construct profiles from two sources. PsyCoProfile (Wan et al., 2026) provides clinical records covering frequent combinations of four common and clinically important psychiatric domains: MDD, AD, BD, and ADHD. NIMH Healthy Research Volunteer Dataset (NIMH) (Nugent et al., 2022) provides healthy records. Because neither source contains the honesty annotations required by our task, we construct variants corresponding to honest reporting, concealment, and malingering from the normalized profiles.

Profile schema and information partition. Each normalized profile contains demographics, chief complaint, clinical history, optional life events, personality descriptors, benchmark-reference diagnosis and severity labels, and honesty information. We partition these fields into a *visible profile* and a *hidden patient state*. The visible profile includes demographics, chief complaint, and clinical history, and is available to assessment-side agents. The hidden state includes private symp-

Module	Visible Profile	Hidden State	Dialogue Evidence	Self-report Output	Clinician-rated Output	Honesty Inference
Assessor	Y	—	—	—	—	—
Patient	Y	Y	Y	Y	—	—
Strategy	—	—	Y	—	—	—
Evaluator	—	—	Y	—	Y	—
Diagnostician	—	—	—	Y	Y	Y

Table 1: Input information access across modules.

toms, benchmark-reference diagnosis and severity labels, personality descriptors, honesty value, and honesty rationale, and is only available to the Patient Agent. This information boundary creates the asymmetry required by our honesty-aware setting: the Patient Agent acts according to hidden state, while the Evaluator, Strategy Agent, and Diagnostician rely only on visible dialogue evidence and derived assessment outputs. Table 1 summarizes this access boundary.

Personality annotation. Each profile includes Big Five personality descriptors. Following work that operationalizes LLM personality through trait descriptors and intensity levels (Serapio-García et al., 2025), we use them as natural-language controls for response style, emotional expressiveness, cooperativeness, and cross-turn consistency, rather than as clinical labels.

Severity annotation. Each clinical diagnosis is assigned a 1–10 cross-diagnostic benchmark severity score, reflecting symptom intensity and functional impairment. This score is a framework reference rather than an independent clinical determination. The detailed rubric is provided in Appendix A. Human validation on 60 sampled profiles using a 1–5 point Likert scale showed that the assigned severity labels were generally consistent with the profile descriptions (mean 4.109/5; 83.7% scores ≥ 4 ; Appendix I.2). The task assessed profile–label consistency rather than independent diagnosis.

Honesty annotation. Each profile contains a hidden honesty value $h \in [-1, 1]$ and an honesty rationale, where negative values indicate concealment, zero indicates approximately honest reporting, and positive values indicate malingering. Here, h is a profile-level dominant reporting condition; it does not encode domain-specific mixtures or turn-level switching. Since the source datasets do not provide honesty annotations, we construct non-honest variants using literature-guided motivation and manifestation themes. Motivation themes justify minimal profile modifications and honesty rationales, while manifestation themes form the feature bank for Strategy Agent hon-

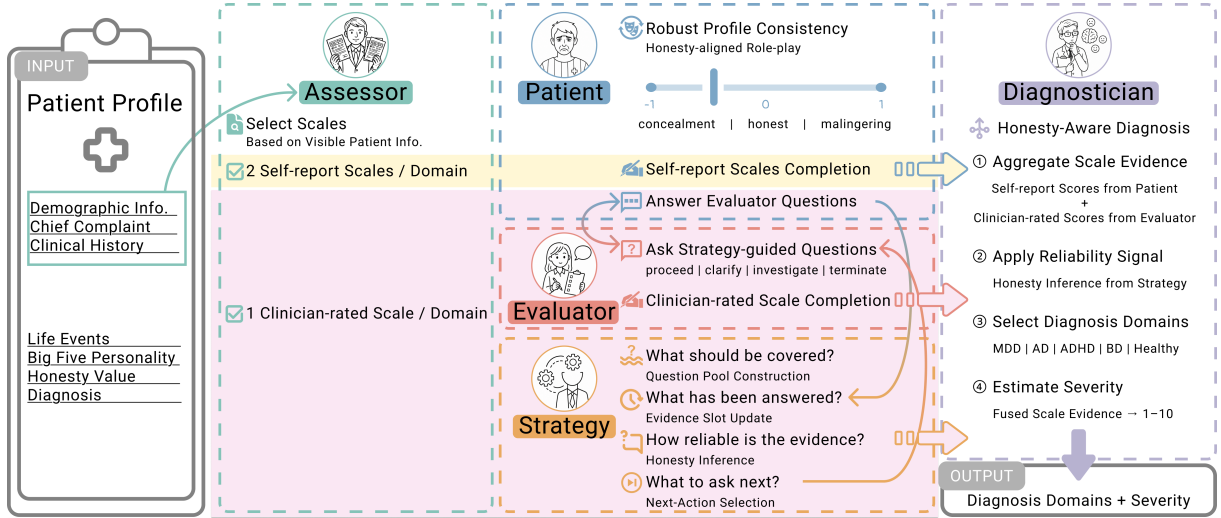


Figure 1: Overview of the honesty-aware multi-agent framework for synthetic psychiatric intake data generation.

esty inference. The resources contain 5/6 motivation and 10/18 manifestation themes for concealment/malingering, respectively; they serve different functions but remain conceptually related. Details appear in Appendix B.

3.3 Five-Agent Framework

As shown in Figure 1, the framework contains five role-specific agents. Patient and Evaluator use LLMs for natural-language interaction, while Assessor, Strategy, and Diagnostician use deterministic procedures for controllability and reproducibility.

Assessor Agent. The Assessor Agent constructs a fixed four-domain assessment battery over MDD, AD, BD, and ADHD, so scale selection does not leak hidden diagnosis labels. It ranks candidate scales from Appendix K using intake-safe visible fields and outputs a structured `ScaleSelection` object.

Patient Agent. The Patient Agent receives the visible profile and hidden patient state, completes self-report scales, and answers interview questions while preserving profile consistency. Self-report scores are adjusted by the hidden honesty value, and interview answers are generated from the hidden profile, dialogue history, personality descriptors, and a private honesty strategy. Details are provided in Appendix D.

Strategy Agent. The Strategy Agent controls the clinician-rated interview from visible dialogue and assessment state. It updates evidence slots, selects the next action, and infers an honesty signal $h_{\text{strategy}} \in [-1, 1]$ from concealment

and malingering features. Its inputs include selected scales and question metadata, evidence slots, asked-question identifiers, and stopping criteria, but exclude hidden h , rationales, profile modifications, diagnosis/severity labels, and Patient prompts. Concealment and malingering cues are scored separately; both confirmed directions yield mixed, and at most one non-accusatory rephrase is permitted. Detailed decision rules, evidence-slot schema, and feature definitions are provided in Appendix D.

Evaluator Agent. The Evaluator Agent conducts the clinician-rated interview and produces clinician-rated scale scores. It asks targeted follow-up questions when evidence is missing, vague, or suspicious, then maps transcript evidence to clinician-rated totals, item evidence, and evidence summaries. Details are summarized in Appendix D.

Diagnostician Agent. The Diagnostician Agent performs Honesty-Aware Diagnostic Fusion from visible self-report outputs, clinician-rated outputs, and the Strategy Agent’s inferred honesty signal. It does not access the hidden patient profile. Full domain prediction, reliability-gating, and severity-estimation details are provided in Appendix E.

3.4 Generated Dataset

Each generated case corresponds to one complete psychiatric intake session derived from a patient profile. All cases, including healthy cases, complete the same four-domain assessment battery, so the presence or absence of a selected domain does not reveal the hidden diagnosis. A case contains

Statistic	Value
Generated cases	2,008
Clinical / healthy cases	1,004 / 1,004
Dialogue turns	58,918
Avg. turns per dialogue	29.34
Avg. questions per dialogue	14.17
Assessment domains per case	4
Self-report scales per case	8
Clinician-rated scales per case	4

Table 2: Overview of the generated dataset.

the selected scales, self-report results, evaluator-patient dialogue, strategy trace, evidence slots, clinician-rated results, inferred honesty, and final diagnostic output. Representative qualitative case studies are presented in Appendix J. Table 2 summarizes the core dataset statistics, while detailed schema fields and annotation definitions are provided in Appendix F.

Generation combines LLM-based dialogue generation with deterministic control and scoring: patient answers and evaluator questions are model-generated. Details on the LLMs, role-level decoding settings, and retry policies are provided in Appendix C.2. We use a grouped 70/15/15 train/development/test split, with 1,408, 300, and 300 cases respectively. The grouping unit is the pair of source dataset and source record ID, ensuring that variants derived from the same source profile do not appear across different splits. Detailed distributional statistics are provided in Appendix G.

4 Experiments and Results

Our experiments address four research questions.

RQ1: Can assigned reporting conditions be encoded in patient behavior and recovered from the interview transcript?

RQ2: How robust is the framework across LLMs under hidden-information constraints?

RQ3: Can honesty-aware diagnostic fusion reduce deviation from synthetic benchmark references under assigned concealment and malingering conditions?

RQ4: Which components matter most for diagnostic accuracy, honesty robustness, and efficiency?

Unless otherwise noted, we evaluate on 2,008 simulated intake sessions and supplement the main experiments with automatic dialogue-quality analysis and human evaluation. The hidden diagnosis, severity, and honesty targets are framework-constructed references rather than independent clinical determinations.

4.1 Honesty Conditioning and Inference

RQ1 examines whether an assigned reporting condition is encoded in profile annotations, expressed in generated patient behavior, and recovered from the interview transcript.

Metrics. To answer RQ1, we assess three aspects. Let S_{true} , S_{disc} , S_{under} , S_{rep} , and S_{over} denote the sets of reference, disclosed, under-reported, reported, and over-reported symptoms, respectively. Let $\tilde{s}_{\text{rep}}$ be the normalized generated self-report score and $\tilde{s}_{\text{honest}}$ be the paired honest-profile proxy. We measure: (1) *annotation-level alignment*, the percentage of profiles whose honesty value, rationale, and modification summary match the expected direction; (2) *patient behavior alignment*, using $\text{Disclosure} = |S_{\text{disc}}|/|S_{\text{true}}|$, $\text{Under-report} = |S_{\text{under}}|/|S_{\text{true}}|$, $\text{Over-report} = |S_{\text{over}}|/|S_{\text{rep}}|$, and signed self-report deviation $\Delta_{\text{SR}} = 100(\tilde{s}_{\text{rep}} - \tilde{s}_{\text{honest}})$; and (3) *strategy honesty inference*, the accuracy N_{match}/N of matching the hidden assigned direction, where N_{match} is the number of matched cases.

Annotation and behavior. Table 3 shows that patient behavior aligns with assigned reporting conditions. Annotation alignment is high across conditions (95.2%–100.0%). Concealment reduces disclosure from 62.3% to 38.4% and increases under-reporting from 58.3% to 84.1%, while malingering increases over-reporting from 21.6% to 82.9% in clinical cases and remains high in healthy cases (85.7%–100.0%).

Self-report deviation. Figure 2 shows the expected signed deviation pattern: negative for concealment, near zero for honest baselines, and positive for malingering. Deviations range from −36.9 percentage points for severe clinical concealment to 16.4 percentage points for severe clinical malingering, and reach 31.3 percentage points for severe healthy malingering.

Strategy inference. Strategy inference recovers the main assigned directions, with condition-level accuracy ranging from 60.1% to 94.8% in Table 3. Figure 3 shows 93.7% accuracy for hidden honest cases, 71.4% for concealment, and 72.1% for malingering. Cross-direction errors are rare: 0.4% concealment-to-malingering and 0.7% malingering-to-concealment.

4.2 Robustness Across LLMs

RQ2 investigates diagnostic robustness across three LLMs: DeepSeek V4, GPT-4.1, and GLM-

Condition	N	Annotation alignment (%)	Disclosure (%)	Under-reporting (%)	Over-reporting (%)	Signed self-report deviation (pp)	Inference accuracy (%)
Clinical honest	502	100.0%	74.7%	0.0%	0.0%	0.0 [0.0, 0.0]	94.8%
Clinical mild concealment	84	100.0%	62.3%	58.3%	0.0%	-9.4 [-11.3, -7.5]	76.2%
Clinical moderate concealment	80	100.0%	50.1%	77.5%	0.0%	-24.9 [-29.1, -20.9]	76.2%
Clinical severe concealment	88	98.7%	38.4%	84.1%	0.0%	-36.9 [-41.6, -31.9]	62.5%
Clinical mild malingering	88	98.9%	77.6%	0.0%	21.6%	2.7 [1.6, 3.9]	76.1%
Clinical moderate malingering	86	100.0%	90.4%	0.0%	66.3%	11.5 [9.4, 13.6]	82.6%
Clinical severe malingering	76	98.7%	88.6%	0.0%	82.9%	16.4 [14.0, 18.9]	72.4%
Healthy honest	502	100.0%	100.0%	0.0%	0.0%	0.0 [0.0, 0.0]	92.6%
Healthy mild malingering	168	100.0%	90.5%	0.0%	85.7%	9.5 [8.7, 10.3]	60.1%
Healthy moderate malingering	167	98.2%	71.0%	0.0%	100.0%	29.0 [28.0, 30.1]	74.3%
Healthy severe malingering	167	95.2%	68.7%	0.0%	100.0%	31.3 [30.5, 32.1]	74.3%

Table 3: Honesty-conditioned behavior results for RQ1. pp = percentage points. Brackets indicate 95% confidence intervals.

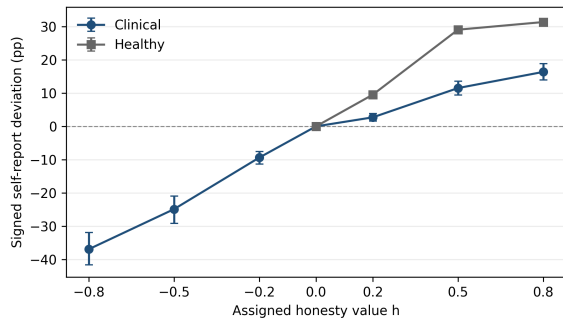


Figure 2: Signed self-report deviation by assigned honesty value. Values are in percentage points; error bars indicate 95% confidence intervals.

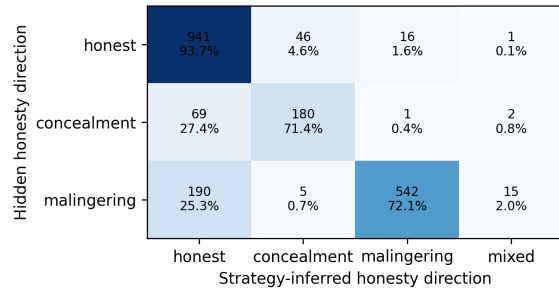


Figure 3: Hidden vs. Strategy-inferred honesty directions. Cell percentages are row-normalized.

5-FP8.

Metrics. We measure three aspects: (1) *domain recovery*, using Exact, Coverage, Extra, and healthy false positive rate (H-FPR), where Exact requires the predicted domain set to match the synthetic reference set, Coverage requires all reference domains to be recovered, Extra counts additional predicted domains, and H-FPR measures healthy-reference cases predicted as clinical; (2) *severity quality*, using Sev. MAE, Signed Err., Joint@1, and Vec. MAE, where Sev. MAE is computed over correctly recovered domains, Signed Err. is predicted minus benchmark-reference severity, Joint@1 requires exact domains and per-domain

severity error no larger than 1, and Vec. MAE uses a fixed diagnosis-severity vector with absent domains set to 0; and (3) *efficiency and leakage*, using runtime, token usage, and detected hidden-information leakage.

Overall performance. Table 4 shows that all three LLMs complete 2,008 cases with zero detected leakage. DeepSeek V4 performs best overall, achieving the highest Exact, Coverage, and Joint@1 scores and the lowest Vec. MAE. GLM-5-FP8 generally ranks second, while GPT-4.1 shows lower domain recovery and a higher healthy false-positive rate. All three LLMs obtain severity MAE close to 1 point on the 1–10 scale, with slightly negative signed errors, indicating mild underestimation on correctly recovered domains.

Cross-model variation is concentrated in domain recovery rather than severity calibration. Sev. MAE remains close to 1 point across LLMs, whereas Exact, Coverage, and H-FPR vary more, suggesting that backend differences mainly affect how much diagnostically useful evidence is exposed during multi-turn interviewing rather than how severity is mapped once a domain is recovered. Since only the patient and evaluator dialogue turns are model-generated, small differences in answer style and follow-up wording can alter the transcript evidence seen by the downstream honesty-aware fusion rules, producing larger changes in missed and extra domains.

Honesty-conditioned results. Figures 4 and 5 show complementary honesty-conditioned patterns. Signed severity error follows the expected direction: concealment leads to underestimation, honest cases stay near zero, and malingering leads to overestimation. Joint@1 is lowest under moderate and severe concealment, suggesting that under-reported symptoms make joint domain-and-severity recovery harder. DeepSeek V4 remains strongest across most honesty groups.

Model	N	Exact	Cover.	Extra	H-FPR	Sev. MAE	Signed Err.	Joint@1	Vec. MAE	Leak
DeepSeek V4	2,008	1,714/2,008 (85.4%)	832/1,004 (82.9%)	137/2,008 (6.8%)	48/1,004 (4.8%)	0.99	-0.20	1,409/2,008 (70.2%)	0.46	0
GPT-4.1	2,008	1,471/2,008 (73.3%)	683/1,004 (68.0%)	240/2,008 (12.0%)	175/1,004 (17.4%)	1.00	-0.37	1,226/2,008 (61.1%)	0.74	0
GLM-5-FP8	2,008	1,544/2,008 (76.9%)	677/1,004 (67.4%)	173/2,008 (8.6%)	83/1,004 (8.3%)	1.04	-0.33	1,278/2,008 (63.6%)	0.62	0

Table 4: Overall performance across LLMs. Sev. MAE is diagnosis-conditioned severity MAE; Signed Err. is predicted minus benchmark-reference severity.

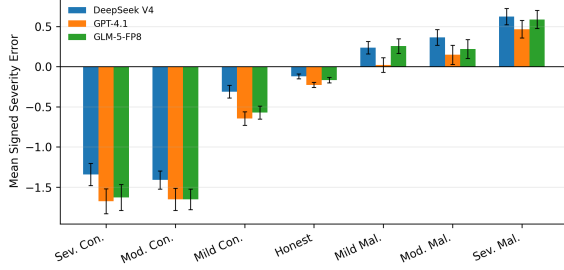


Figure 4: Signed severity error by honesty condition. Negative values indicate underestimation and positive values indicate overestimation; error bars indicate uncertainty estimates.

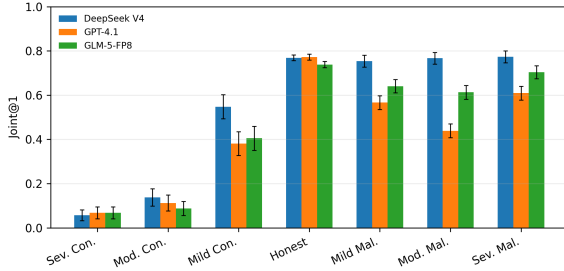


Figure 5: Joint@1 by honesty condition.

Model	Question/Case	Time/Case (s)	Token/Case (K)	Token (M)
DeepSeek V4	14.17	271.47	53.10	106.63
GPT-4.1	12.86	38.65	37.81	75.92
GLM-5-FP8	13.99	296.37	52.60	105.62

Table 5: Runtime and token usage across LLMs.

Runtime and token usage. Table 5 shows an accuracy–efficiency trade-off. GPT-4.1 is fastest and uses the fewest tokens, but has lower diagnostic performance. DeepSeek V4 performs best but requires more runtime and tokens, while GLM-5-FP8 has similar cost with lower accuracy.

4.3 Assessment Fusion across Honesty Conditions

RQ3 examines how assigned reporting conditions affect self-report, clinician-rated assessment, and final diagnosis, and whether diagnostic fusion reduces deviation from synthetic benchmark references.

Metrics. We report 4 metrics. (1) *SRD* is self-report deviation, computed

Condition	N	SRD	CRD	FD	Imp.
Clinical concealment	252	1.65	2.56	1.12	0.54
Clinical malingering	250	2.54	2.23	1.05	1.49
Clinical honest	502	1.64	2.22	0.89	0.75
Healthy honest	502	0.00	0.38	0.04	-0.04
Healthy malingering	502	2.09	0.52	0.30	1.79

Table 6: Stage-wise deviation from synthetic benchmark references and final correction. SRD = self-report deviation; CRD = clinician-rated deviation; FD = final diagnosis deviation; Imp. = SRD – FD.

as $SRD = |\text{self-report severity} - \text{benchmark-reference severity}|$. (2) *CRD* is clinician-rated deviation, computed as $CRD = |\text{clinician-rated severity} - \text{benchmark-reference severity}|$. (3) *FD* is final diagnosis deviation, computed as the mean absolute difference between the final-diagnosis and benchmark-reference severity vectors over MDD, AD, BD, and ADHD, with absent domains assigned baseline severity 1. (4) *Imp.* measures final correction over self-report, computed as $Imp. = SRD - FD$; positive values indicate that the final diagnosis is closer to the benchmark reference than self-report. All metrics are reported in points on the 1–10 severity scale.

Final correction. Table 6 shows that FD is lower than SRD in all clinical groups and in healthy malingering, with the largest drops in healthy malingering (2.09 to 0.30) and clinical malingering (2.54 to 1.05). In the healthy-honest floor case, fusion increases deviation from 0.00 to 0.04.

Clinician-rated assessment. CRD is not uniformly lower than SRD: it is higher in clinical concealment and clinical honest cases, but lower in clinical and healthy malingering. Thus, the final improvement comes from diagnostic fusion rather than simply replacing self-report with clinician-rated scoring.

Condition differences. Imp. is highest for healthy malingering (1.79) and clinical malingering (1.49), but smaller for clinical concealment (0.54), suggesting stronger correction for exaggerated than under-disclosed self-report.

4.4 Ablation Study

RQ4 examines which components of the strategy-guided interview contribute to diagnostic accuracy, honesty robustness, and interaction efficiency.

Settings. All comparisons use the same stratified 200-case subset, with 40 cases from each source-honesty condition. We compare Full Strategy against two component ablations and three Evaluator-mode variants: *Item-only* removes strategy-guided evidence tracking and asks scale items in order; *No-suspicion* disables feature-bank suspicion and honesty-aware re-questioning; and evaluator variants use *Fixed-scale*, *Adaptive*, or *Suspicion-aware*. The three Evaluator-mode variants retain the same Strategy backbone and differ only in Evaluator mode/state prompting.

We report diagnostic accuracy, over-/under-predicted domains, severity MAE, and interaction cost using average evaluator questions and mean token usage in Table 7.

Strategy control. Removing strategy-guided evidence tracking reduces Exact from 71.0% to 48.0% and increases Extra domains from 39 to 146. Item-only also requires far more questions and tokens, showing that scale-order questioning is less accurate and less efficient. It is an exhaustive negative control rather than a short adaptive baseline.

Suspicion mechanism. No-suspicion modestly lowers Exact from 71.0% to 68.0% and increases Extra domains from 39 to 50. With 13.49 questions, it is the shorter adaptive comparator and closes much of the Item-only gap. Disabling suspicion also removes non-honest detection: Full Strategy detects 60.0% of concealment and 45.0% of malingering cases, while No-suspicion detects neither. The false suspicion rate remains low for both settings. Detailed honesty-robustness results are provided in Appendix H.2.

Evaluator variants. Full Strategy obtains the best Coverage and MAE among evaluator variants, while tying Suspicion-aware on Exact. Because all four rows retain the Strategy loop, these differences reflect Evaluator mode/state-prompt sensitivity rather than component effects.

4.5 Automatic Dialogue Quality Analysis

We further report corpus-level automatic quality metrics for the 2,008 generated sessions and compare them with PsyCoTalk, the source dialogue dataset corresponding to PsyCoProfile. Our gen-

erated dialogues are shorter in interaction length (14.17 vs. 45.88 rounds) but have longer evaluator/patient turns (102.6/207.4 vs. 34.7/43.4 characters), while showing lower entity coverage (0.166 vs. 0.888), lower BERTScore F1 (0.529 vs. 0.638), and higher PPL (34.01 vs. 9.31). These metrics indicate a trade-off and do not support greater naturalness or clinical richness than PsyCoTalk. Full automatic metric results are provided in Appendix H.3.

4.6 Human Evaluation

We conduct a crowdsourced human evaluation on 150 generated sessions using the Credamo platform¹, with 60 honest, 30 concealment, and 60 malingering cases and at least three qualified ratings per case. For honesty-specific items, raters saw the assigned target category; this was not a blind honesty-inference task. Details are provided in Appendix I.

Table 8 shows positive ratings overall, with an average score of 4.060 and 83.5% positive ratings. Dialogue quality and diagnosis/evidence quality exceed 4.1, while honesty behavior is lower at 3.905. These ratings assess plausibility, not true honesty, independent diagnosis, clinical utility, or preference for fused summaries.

5 Discussion

Intended use and users. The immediate users are NLP and clinical-AI researchers, benchmark developers, and educators studying simulated interviews under asymmetric information. “Lower burden” means fewer Evaluator questions and tokens within the simulation, not reduced patient or clinical burden; pre-assessment and training uses require independent validation.

Beyond theory-of-mind scores. Prior work has reported strong theory-of-mind performance for GPT-4 (Strachan et al., 2024). Our controlled benchmark tests a different ability: gathering evidence across turns, respecting hidden-information boundaries, interpreting structured scales, and avoiding overcalling disorders under distorted synthetic self-report. In this setting, DeepSeek V4 outperforms GPT-4.1 on benchmark-reference labels, with higher Exact accuracy, higher Joint@1, and a lower healthy-reference false-positive rate. This suggests that honesty-aware intake is not only a

¹<https://www.credamo.world/>

Group	Setting	Exact	Coverage	Extra	Missing	MAE	Questions	Tokens
Reference	Full Strategy	71.0%	85.5%	39	30	0.851	14.12	47,528
Component ablation	Item-only	48.0%↓	83.0%↓	146↑	34↑	0.880↑	68.92↑	150,163↑
Component ablation	No-suspicion	68.0%↓	83.5%↓	50↑	33↑	0.888↑	13.49↓	45,358↓
Mode sensitivity	Fixed-scale	67.0%↓	81.0%↓	41↑	39↑	0.864↑	13.97↓	47,110↓
Mode sensitivity	Adaptive	67.0%↓	80.0%↓	46↑	42↑	0.931↑	14.11↓	47,905↑
Mode sensitivity	Suspicion-aware	71.0%	83.0%↓	41↑	34↑	0.918↑	14.04↓	47,282↓

Table 7: RQ4 results on the stratified 200-case subset. Full Strategy is the reference; arrows indicate changes relative to it. Item-only and No-suspicion are component ablations; Fixed-scale, Adaptive, and Suspicion-aware are Evaluator-mode variants that retain the Strategy backbone. Extra/Missing denote total over-/under-predicted domains. Questions and Tokens are per-case averages.

Dimension	Mean [95% CI]	W1	Pos.
Dialogue quality	4.131 [4.092, 4.169]	.862	.864
Honesty behavior	3.905 [3.861, 3.949]	.765	.767
Diagnosis/evidence	4.118 [4.077, 4.157]	.851	.861
Overall	4.060 [4.027, 4.093]	.830	.835

Table 8: Human evaluation results. W1 = within-one-point agreement; Pos. = ratings ≥ 4 .

mental-state reasoning task, but also a reliability and evidence-integration task.

Fusion rather than replacement. Clinician-rated assessment is not a simple replacement for self-report. In some settings, self-report remains closer to the synthetic reference because it is generated nearer to latent profile severity, while clinician-rated scores are inferred from finite dialogue evidence. The healthy-honest floor case is a small exception (0.00 to 0.04). The system improves because it fuses self-report, interview evidence, honesty inference, and diagnostic reasoning.

Interaction strategy matters. The ablation results show that strategy-guided interviewing improves both accuracy and efficiency. Removing strategy-guided evidence tracking reduces Exact accuracy, increases extra predicted domains, and requires nearly five times more questions than the full system. Item-only is an exhaustive negative control, whereas No-suspicion is the shorter adaptive comparator.

6 Conclusion

This paper introduced an honesty-aware multi-agent framework for synthetic psychiatric intake simulation and a 2,008-case benchmark. By separating visible profile information from hidden benchmark-reference diagnosis, severity, personality, and assigned reporting conditions, the framework supports controlled study of honest reporting, concealment, and malingering in one assessment pipeline.

Experiments show that the framework expresses

assigned reporting conditions in patient behavior, supports transcript-based honesty inference, and achieves strong benchmark-reference domain and severity recovery without detected hidden-information leakage. The results also show that clinician-rated assessment complements rather than replaces self-report, and that strategy-guided interviewing improves benchmark accuracy and interaction efficiency. Accordingly, the contribution is a controlled research testbed for interviewing and evidence fusion.

Limitations

This work has several limitations. First, the dataset is fully synthetic, and the honesty labels are literature-guided constructions rather than gold annotations from real psychiatric intake encounters. Diagnosis and severity labels are likewise benchmark references rather than independent clinical determinations. The scalar h encodes a dominant case-level condition, not domain-specific mixtures or turn-level switching; Strategy’s mixed output denotes competing transcript evidence, not a mixed generator class. This is necessary for privacy-preserving controlled experiments, but it may also favor the behavioral patterns encoded by our profile normalization, prompting setup, and honesty feature bank. Construction and inference use distinct motivation and manifestation resources, but both are literature-guided, so construct-level dependence remains. Real patients may combine concealment, exaggeration, poor insight, stigma management, and culturally shaped disclosure in ways that are less cleanly separable than our hidden honesty variable.

Second, the current benchmark has limited clinical scope. It covers four psychiatric domains and is derived from normalized profiles from two source datasets, so it does not yet capture the full diversity of comorbidity, longitudinal course, crisis presentation, medication effects, or demo-

graphic and cultural variation seen in practice. Our qualitative failures also show that some settings remain difficult even in the synthetic regime, especially healthy malingering false positives and bipolar recovery under conflicting evidence. Independent authored or expert role-played transcripts are needed to assess external validity.

Third, several evaluation components are framework-specific rather than direct clinical endpoints. Hidden severity is assigned with a cross-diagnostic rubric, and clinician-rated outputs are reconstructed from finite dialogue evidence through deterministic mappings. Human evaluation is also limited in scale. HE1 assesses plausibility and HE2 profile-label consistency; neither validates true honesty, independent diagnosis, clinical utility, or preference for fused summaries. For these reasons, the results should be interpreted as evidence that the framework is useful for simulation and benchmarking, not as proof of deployment readiness or diagnostic safety in real clinical settings.

The public repository provides code, prompts, configurations, toy examples, tests, and aggregate results; complete sessions and provider traces are not released.

Ethical Considerations

This synthetic research benchmark is not a clinical assessment or deception-detection system and must not be used for diagnosis, screening, triage, treatment planning, crisis response, or decisions about an individual's honesty. Reporting discrepancies in real care may reflect fear, trauma, culture, poor insight, or access barriers rather than intentional deception.

A lexical audit found synthetic self-harm- or suicide-related language in 193/2,008 sessions (9.61%) and substance-use-related language in 16 sessions. The framework has no validated crisis-response protocol; Appendix Table 27 specifies prohibited clinical and high-impact uses.

Acknowledgments

This work is supported by The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), under Award No UDF01004172.

References

Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. [Using large language models to simulate](#)

[multiple humans and replicate human subject studies](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. [Out of one, many: Using language models to simulate human samples](#). *Political Analysis*, 31(3):337351.

Carlos Barbosa-Torres, Natalia Bueso-Izquierdo, and Sixto Cubo-Delgado. 2023. [Analysis of malingered psychological symptoms in a clinical sample for early detection in initial interviews](#). *European Archives of Psychiatry and Clinical Neuroscience*, 273(2):427–438.

Shani Bril-Barniv, Galia S. Moran, Adi Naaman, David Roe, and Orit Karnieli-Miller. 2017. [A qualitative study examining experiences and dilemmas in concealment and disclosure of people living with serious mental illness](#). *Qualitative Health Research*, 27(4):573–583. PMID: 28682733.

Pengfei Cao, Yuanzhe Zhang, Chenxiang Zhang, Wei Chen, Yan Liu, Shuang Xu, Miao Xu, Wenqing Jin, Jinjie Xu, Dan Wang, Wei Wang, Xue Wang, Wen Wang, Yanping Ren, Jun Zhao, Rena Li, and Kang Liu. 2025. [A multimodal depression consultation dataset of speech and text with hamd-17 assessments](#). *Scientific Data*, 12(1):1577.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. [ChatEval: Towards better LLM-based evaluators through multi-agent debate](#). In *Proceedings of the Twelfth International Conference on Learning Representations*.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. [SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183, Singapore. Association for Computational Linguistics.

Sanford L Drob, Kevin B Meehan, and Shari E Waxman. 2009. [Clinical and conceptual problems in the attribution of malingering in forensic evaluations](#). *The journal of the American Academy of Psychiatry and the Law*, 37(1):98106.

Zhijun Guo, Alvina Lai, Johan H Thygesen, Joseph Farrington, Thomas Keen, and Kezhi Li. 2024. [Large language models for mental health applications: Systematic review](#). *JMIR Ment Health*, 11:e57400.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. [MetaGPT: Meta](#)

- programming for a multi-agent collaborative framework. In *Proceedings of the Twelfth International Conference on Learning Representations*.
- Akshi Kumar, Aditi Sharma, and Saurabh Raj Sangwan. 2025. [Dynamenta: Dynamic prompt engineering and weighted transformer architecture for mental health classification using social media data](#). *IEEE Transactions on Computational Social Systems*, pages 1–11.
- Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, Kyong-Mee Chung, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. [Cactus: Towards psychological counseling conversations using cognitive behavioral theory](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14245–14274, Miami, Florida, USA. Association for Computational Linguistics.
- Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. [CAMEL: Communicative agents for “mind” exploration of large language model society](#). In *Advances in Neural Information Processing Systems*, volume 36.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. [Towards emotional support dialog systems](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3469–3483. Association for Computational Linguistics.
- Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. [Roleplay-doh: Enabling domain-experts to create LLM-simulated patients via eliciting and adhering to principles](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10570–10603, Miami, Florida, USA. Association for Computational Linguistics.
- Ming-Jie Luo, Shaowei Bi, Jianyu Pang, Lixue Liu, Ching-Kit Tsui, Yunxi Lai, Wenben Chen, Yahan Yang, Kezheng Xu, Lanqin Zhao, Ling Jin, Duoru Lin, Xiaohang Wu, Jingjing Chen, Rongxin Chen, Zhenzhen Liu, Yuxian Zou, Yangfan Yang, Yiqing Li, and Haotian Lin. 2025. [A large language model digital patient system enhances ophthalmology history taking skills](#). *NPJ Digital Medicine*, 8(1):502.
- Aishik Mandal, Tanmoy Chakraborty, and Iryna Gurevych. 2025. [MAGNeT: Coordinated multi-agent generation of synthetic multi-turn mental health counseling sessions](#). *Preprint*, arXiv:2509.04183.
- Hunter L. McQuiston and Rachel Zinns. 2019. [Workloads in clinical psychiatry: Another way](#). *Psychiatric Services*, 70(10):963–966.
- Wiley Mittenberg, Christine Patton, Elizabeth M. Canyock, and Daniel C. Condit. 2002. [Base rates of malingering and symptom exaggeration](#). *Journal of Clinical and Experimental Neuropsychology*, 24(8):1094–1102. PMID: 12650234.
- Allison C. Nugent, Adam G. Thomas, Margaret Mahoney, Alison Gibbons, Jarrod T. Smith, Antoinette J. Charles, Jacob S. Shaw, Jeffrey D. Stout, Anna M. Namyst, Arshitha Basavaraj, Eric Earl, Travis Riddle, Joseph Snow, Shruti Japee, Adriana J. Pavletic, Stephen Sinclair, Vinai Roopchansingh, Peter A. Bandettini, and Joyce Chung. 2022. [The nimh intramural healthy volunteer dataset: A comprehensive meg, mri, and behavioral resource](#). *Scientific Data*, 9(1):518.
- Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. [Generative agents: Interactive simulacra of human behavior](#). In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22. Association for Computing Machinery.
- Huachuan Qiu, Peiyan Zhao, Anqi Zhang, Weiyang Hu, Zeyin Zhang, Jingshi Wang, and Xinyu He. 2024. [SMILE: Single-turn to multi-turn inclusive language expansion via ChatGPT for mental health support](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 590–636, Miami, Florida, USA. Association for Computational Linguistics.
- Federico Ravenda, Seyed Ali Bahrainian, Andrea Raballo, Antonietta Mira, and Noriko Kando. 2025. [Are LLMs effective psychological assessors? leveraging adaptive RAG for interpretable mental health screening through psychometric practice](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8975–8991, Vienna, Austria. Association for Computational Linguistics.
- Gony Rosenman, Talma Hendler, and Lior Wolf. 2024a. [Llm questionnaire completion for automatic psychiatric assessment](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 403–415. Association for Computational Linguistics.
- Gony Rosenman, Talma Hendler, and Lior Wolf. 2024b. [LLM questionnaire completion for automatic psychiatric assessment](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 403–415, Miami, Florida, USA. Association for Computational Linguistics.
- Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2025. [A psychometric framework for evaluating and shaping personality traits in large language models](#). *Nature Machine Intelligence*, pages 1–15.

- David V Sheehan, Yves Lecrubier, K Harnett Sheehan, Patricia Amorim, Juris Janavs, Emmanuelle Weiller, Thierry Hergueta, Roxy Baker, Geoffrey C Dunbar, and 1 others. 1998. The mini-international neuropsychiatric interview (mini): the development and validation of a structured diagnostic psychiatric interview for dsm-iv and icd-10. *J clin psychiatry*, 59(Suppl 20):22–33.
- James W. A. Strachan, Dalila Albergio, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. 2024. [Testing theory of mind in large language models and humans](#). *Nature Human Behaviour*, 8(7):1285–1295.
- Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, and 7 others. 2025. [Towards conversational diagnostic artificial intelligence](#). *Nature*, 642(8067):442–450.
- Tianxi Wan, Jiaming Luo, Siyuan Chen, Kunyao Lan, chenjianhua, Haiyang Geng, and Mengyue Wu. 2026. [From medical records to diagnostic dialogues: A clinical-grounded approach and dataset for psychiatric comorbidity](#). In *The Fourteenth International Conference on Learning Representations*.
- Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2024a. [NoteChat: A dataset of synthetic patient-physician conversations conditioned on clinical notes](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15183–15201, Bangkok, Thailand. Association for Computational Linguistics.
- Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024b. [RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.
- Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate V Hardy, Hong Shen, Fei Fang, and Zhiyu Chen. 2024c. [PATIENT- \$\psi\$: Using large language models to simulate patients for training mental health professionals](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12772–12797, Miami, Florida, USA. Association for Computational Linguistics.
- Samantha Weber, Nicolas Deperrois, Robert Heun, Laura Frühschütz, Anna Monn, Stephanie Homan, Andrea Häfliger, Erich Seifritz, Tobias Kowatsch, MULTICAST consortium, Lena Jäger, Katharina Schultebrasucks, Sapir Gershov, Jacopo Mocellin, Birgit Kleim, and Sebastian Olbrich. 2025. [Using a fine-tuned large language model for symptom-based depression evaluation](#). *npj Digital Medicine*, 8(1):598.
- Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2024. [AutoGen: Enabling next-gen LLM applications via multi-agent conversations](#). In *Proceedings of the First Conference on Language Modeling*.
- Ancheng Xu, Di Yang, Renhao Li, Jingwei Zhu, Minghuan Tan, Min Yang, Wanxin Qiu, Mingchen Ma, Haihong Wu, Bingyu Li, Feng Sha, Chengming Li, Xiping Hu, Qiang Qu, Derek F. Wong, and Ruifeng Xu. 2025. [AutoCBT: An autonomous multi-agent framework for cognitive behavioral therapy in psychological counseling](#). *Preprint*, arXiv:2501.09426.
- Binwei Yao, Chao Shi, Likai Zou, Lingfeng Dai, Mengyue Wu, Lu Chen, Zhen Wang, and Kai Yu. 2022a. [D4: A chinese dialogue dataset for depression-diagnosis-oriented chat](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2438–2459. Association for Computational Linguistics.
- Binwei Yao, Chao Shi, Likai Zou, Lingfeng Dai, Mengyue Wu, Lu Chen, Zhen Wang, and Kai Yu. 2022b. [D4: a Chinese dialogue dataset for depression-diagnosis-oriented chat](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2438–2459, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Congchi Yin, Feng Li, Shu Zhang, Zike Wang, Jun Shao, Piji Li, Jianhua Chen, and Xun Jiang. 2025. [Mdd-5k: A new diagnostic conversation dataset for mental disorders synthesized via neuro-symbolic llm agents](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):2571525723.
- Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. [MedDialog: Large-scale medical dialogue datasets](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250. Association for Computational Linguistics.
- Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. [SOTOPIA: Interactive evaluation for social intelligence in language agents](#). In

A Severity Rubric

We use a cross-diagnostic psychiatric severity rubric with scores from 1 to 10. Severity is judged by symptom intensity, persistence, functional impairment, behavioral dysregulation, and level of clinical concern. Table 9 gives the annotation anchors used for hidden severity labels and for severity-error evaluation.

Score	Label	Description
1	No evident psychiatric symptoms	No clinically meaningful symptoms are documented. Mental state appears within normal limits.
2	Minimal symptoms	Very mild or occasional symptoms may be present, but they do not cause distress or functional impairment.
3	Mild symptoms	Symptoms are noticeable but mild, causing slight distress or minor functional impact in isolated areas.
4	Mild-to-moderate symptoms	Symptoms are present with some consistency, leading to mild functional impairment in one or two domains.
5	Moderate symptoms	Symptoms are clearly present and cause moderate distress or impairment in daily functioning, but basic self-care is maintained.
6	Moderate-to-severe symptoms	Symptoms are frequent and intense, causing significant functional impairment across multiple domains. Some risk behaviors or deteriorating relationships may be present.
7	Severe symptoms	Symptoms are severe and pervasive, markedly interfering with work, relationships, and self-care. Clinical intervention is strongly warranted.
8	Very severe symptoms	Severe and persistent symptoms with major functional collapse. High level of clinical concern, possible safety risks, or frequent crisis episodes.
9	Extremely severe symptoms	Symptoms are extreme and near-continuous, with profound disability, serious safety concerns, and urgent need for intensive intervention.
10	Most extreme severity	Catastrophic severity. Complete functional breakdown, imminent danger to self or others, or psychotic-level disruption requiring emergency or inpatient care.

Table 9: Cross-diagnostic psychiatric severity rubric used for hidden severity annotation and severity evaluation.

B Honesty Annotation Rubric

Each profile contains a hidden honesty value $h \in [-1, 1]$. Negative values indicate concealment, zero indicates no clear distortion or approximately honest reporting, and positive values indicate malingering. This scalar specifies a profile-level dominant generation condition for controlled matched-profile comparisons. It does not encode simultaneous domain-specific directions or turn-level switching. The hidden value is available only to the Patient Agent; Strategy computes a separate inference from visible interview evidence.

The criteria were developed through PRISMA-style retrieval and thematic analysis over PubMed,

Europe PMC, and Crossref. The searches retrieved 992 malingering records, of which 62 informed motivation analysis and 134 informed manifestation analysis, and 1,206 concealment records, of which 38 and 53 informed the corresponding analyses. The resulting resources contain 6 malingering and 5 concealment motivation themes for profile construction, and 18 malingering and 10 concealment manifestation themes for Strategy inference. They serve different operational roles but address related constructs within the same literature-guided framework and are therefore not fully independent. Tables 10 and 11 document this separation and the retrieval counts. Table 12 gives the scalar anchors, and Table 13 summarizes the construction-side motivation categories. These literature-derived themes are benchmark design resources, not consensus diagnostic criteria or an expert-validated deception instrument.

Resource	Framework role	Literature basis	Output
Motivation knowledge base	Construct plausible profile variants and private rationales	PubMed, Europe PMC, and Crossref retrieval followed by thematic analysis	5 concealment and 6 malingering motivation themes
Manifestation feature bank	Detect observable transcript behavior from visible evidence	Separate thematic analysis over the same three literature sources	10 concealment and 18 malingering manifestation themes

Table 10: Operational separation between resources used for non-honest profile construction and Strategy-Agent inference.

Condition	Retrieved	Motiv. incl.	Manifest. incl.	Themes (M/F)
Concealment	1,206	38	53	5 / 10
Malingering	992	62	134	6 / 18

Table 11: Retrieval-and-thematic-analysis summary for the honesty resources. M/F denotes motivation/manifestation themes; this is a summary rather than a full PRISMA flow diagram.

Anchor	Interpretation
-1.0	Concealment
0.0	No clear distortion / approximately honest
1.0	Malingering

Table 12: Honesty-value anchors.

Rule	Direction	Motivation category and description
C01	Concealment	Stigma avoidance: minimizing, omitting, or underreporting symptoms to avoid shame, rejection, discrimination, or being judged as unstable or dangerous.
C02	Concealment	Autonomy preservation: minimizing symptoms to avoid coercive care, involuntary hospitalization, privacy loss, monitoring, medication changes, or family notification.
C03	Concealment	Consequence avoidance: underreporting symptoms to avoid social, occupational, educational, military, licensing, custody, insurance, immigration, or legal consequences.
C04	Concealment	Identity management: concealing symptoms to protect an identity of being normal, capable, safe, independent, or not dependent on psychiatric services.
C05	Concealment	Denial or poor insight: apparent concealment may reflect limited insight, normalization, defensive coping, social desirability, or culturally shaped reluctance.
M01	Malingering	External legal gain: exaggerating or feigning symptoms for legal or forensic advantage.
M02	Malingering	External financial gain: exaggerating symptoms for compensation, disability benefits, sick leave, damages, or other material benefits.
M03	Malingering	Resource access: exaggerating or feigning symptoms to obtain hospitalization, shelter, food, medication, protection, safety, or services.
M04	Malingering	Duty or environment avoidance: exaggerating symptoms to avoid work, school, military service, jail conditions, sanctions, or interpersonal obligations.
M05	Malingering	Instrumental distress communication: amplifying symptoms as a maladaptive way to communicate distress or be taken seriously.
M06	Malingering	Factitious-like motivation: feigning appears motivated by occupying the patient role rather than a clear external reward and should be interpreted cautiously.

Table 13: Motivation categories used for hidden honesty annotation.

C LLM-Prompted Components

In the formal workflow, LLM-generated free text that enters the interview is used for patient-facing answers and evaluator question naturalization. The Patient self-report prompt path remains wired, but the exported structured self-report scores are constructed deterministically from profile and scale logic. Accordingly, this appendix summarizes only the Patient and Evaluator prompt templates. Deterministic scale selection, Strategy decisions, clinician-rated scoring, and formal final diagnosis integration are described in their corresponding method sections rather than as LLM prompts.

C.1 Patient Agent Prompt

The Patient Agent prompt controls natural-language patient answers during the clinician-rated interview. It receives the full private synthetic profile and hidden honesty setting, but it must express them only through ordinary patient language and must not reveal hidden labels or internal parameters.

Prompt for Patient Agent Interview Response

You are the assigned patient in a structured psychiatric interview.
Always answer in the first person and stay in character.
Do not reveal that the interaction is simulated.

Input:

- Private synthetic profile: {private_profile}
- Dialogue history visible to the patient: {dialogue_history}
- Current evaluator question: {evaluator_question}
- Private honesty directive: {honesty_directive}

Rules:

1. Use only information consistent with the assigned profile.
2. Express symptoms in ordinary patient language, not hidden labels.
3. If honest, answer according to the profile facts.
4. If concealment, soften, minimize, or avoid sensitive details.
5. If malingering, exaggerate only in the profile-indicated direction.
6. Never reveal hidden diagnosis, hidden severity, honesty value, private annotations, or internal strategy parameters.
7. Do not ask the interviewer questions except for clarification.

Output Format:

Patient response: {one concise natural-language answer}

Structured self-report scores are generated deterministically from profile and scale logic, although the Patient prompt path remains wired for a visible patient-style symptom summary. The prompted Patient Agent is therefore used for interview language, not for unconstrained free-form scale completion.

C.2 Evaluator Agent Prompt

The Evaluator Agent prompt converts structured Strategy guidance into the next patient-facing clinical question. It does not independently choose hidden labels, read self-report outputs, or access private patient state.

Prompt for Evaluator Agent Follow-up Question

You are the Evaluator Agent conducting a clinical interview.
Ask exactly one concise, professional, patient-facing question.

Input:

- Public patient view: {public_patient_view}
- Selected clinician-rated scales: {selected_scales}
- Dialogue history: {dialogue_history}
- Strategy decision: {strategy_decision}
- Target dimensions: {target_dimensions}
- Visible reliability inference: {reliability_inference}

Task:

1. Convert the Strategy decision into natural clinical wording.
2. Ask about missing, vague, or low-confidence evidence.
3. Avoid verbatim scale-item wording when possible.
4. If clarification is requested, ask a focused follow-up.
5. If consistency check is requested, ask neutrally about timeline, functional impairment, preserved functioning, or context.

Forbidden:

- Do not accuse the patient of deception.
- Do not expose Strategy identifiers, hidden profile fields, hidden benchmark-reference diagnosis/severity, or hidden honesty value.

Output Format:

Evaluator question: {one conversational clinician utterance}

Generation configuration. Table 14 summarizes the formal configuration used for the primary 2,008-case run. Sanitized DeepSeek and GPT-4.1

configuration files are provided in the public repository and report the provider interface, model identifier, role-level temperatures and token caps, timeout, and reasoning controls.

Configuration field	Formal setting
Language and runtime	English; CAMEL-based multi-agent runtime; mock mode disabled.
Model request identifiers	Primary run: deepseek-v4-pro; comparison run: gpt-4.1. Both use OpenAI-compatible request interfaces.
Evaluator mode	Full strategy-guided evaluator.
Safety turn limit	64 turns per session.
Suspicion gate	Cumulative suspicion threshold 0.24; feature-confidence threshold 0.80; at most one suspicion rephrase.
Consistency probe style	Validity-probe rephrasing over timeline, preserved functioning, and symptom–function consistency.
Reasoning policy	Deterministic patient-turn strategy, evaluator strategy reasoning, evaluator rating summary, and diagnostician summary.
Patient self-report path	Patient prompt path remains wired, but structured self-report scores are exported from deterministic profile-and-scale logic.
Information boundaries	Honesty value hidden from Evaluator and Diagnostician; benchmark-reference diagnosis hidden from Diagnostician; output leakage scanning enabled.
Temperatures	Assessor 0.1; Patient 0.6; Evaluator 0.4; Diagnostician 0.1.
Response limits	Assessor 500; Patient 550; Evaluator 300; Diagnostician 900 generated tokens (default fallback 800); timeout 90 seconds.
Other decoding controls	Neither top_p nor a decoding seed was explicitly set in the formal configuration.

Table 14: Formal generation configuration for the primary 2,008-case run.

D Strategy Agent Details

The Strategy Agent is a control component for the interview rather than a patient-facing speaker. It maintains structured state over what evidence has been collected, what remains uncertain, and whether the interview transcript contains reliability concerns. This design keeps the questioning policy explicit while leaving natural-language question wording to the Evaluator Agent.

Inputs and outputs. The Strategy Agent does not communicate with the patient. It receives selected scales, visible dialogue history, question-pool metadata, evidence slots, asked question identifiers, and stopping criteria. It does not receive

hidden h , the private honesty rationale, the profile modification summary, hidden diagnosis/severity labels, or the Patient’s private generation strategy. Its outputs are structured strategy decisions, updated evidence slots, and an honesty inference derived from visible transcript evidence. Table 15 summarizes these inputs and outputs.

Component	Role in Strategy Agent
Selected scales	Define the assessment domains and rater-administered dimensions requiring interview evidence.
Dialogue history	Visible evaluator and patient turns used to update evidence slots and infer honesty/reliability.
Question pool	Structured candidate questions with target dimensions, question type, cost, and information value.
Evidence slots	Structured state for symptom presence, frequency, duration, severity, impairment, confidence, and status.
Strategy decision	Next action: ask the next question, clarify, perform a consistency check, stop, or use fallback.
Honesty inference	Feature-bank-derived inference over visible patient answers, including concealment, malingering, suspicion, and confidence scores.

Table 15: Strategy Agent inputs and outputs.

Evidence slot schema. Evidence slots are updated after evaluator–patient exchanges and provide the intermediate state used for clinician-rated assessment. Each slot stores the current estimate for one target dimension, its supporting dialogue evidence, provenance, confidence, and sufficiency status. Table 16 lists the fields maintained for each slot.

Field	Meaning
scale_item	Scale item or scale-dimension key, usually formed from the clinician-rated scale identifier and a target dimension.
dimension	Clinical dimension tracked by the interview.
presence	Boolean symptom-presence estimate when supported by dialogue evidence.
frequency	Frequency evidence extracted from patient answers.
duration	Duration or course evidence extracted from patient answers.
severity	Severity wording or intensity evidence extracted from patient answers.
impairment	Functional impairment evidence.
evidence_quotes	Patient-answer snippets supporting the slot.
source_question_ids	Question identifiers linked to the evidence.
last_updated_turn_id	Dialogue turn that most recently updated the slot.
confidence	Numeric confidence score from 0 to 1.
status	One of unknown, insufficient, sufficient, or conflicting.

Table 16: Evidence slot schema maintained by the Strategy Agent.

Next-action decision rules. The Strategy Agent uses an internal decision vocabulary that is translated into the Evaluator Agent’s operational action vocabulary. Consistency checks are implemented as non-accusatory probes for unclear or conflicting evidence rather than as direct accusations. Table 17 shows the decision-to-action mapping.

Strategy decision	Evaluator action	Trigger and purpose
ask_next	proceed	Selects the highest-value non-redundant question from the question pool for missing or low-coverage evidence.
clarify	clarify	Requests focused clarification when targeted evidence is vague or low-confidence. In the formal workflow, reliability-probe behavior requires confirmed feature-bank suspicion.
consistency_check	investigate	Requests a non-accusatory consistency or validity probe when conflicting evidence or confirmed feature-bank suspicion is present.
stop	terminate	Stops when evidence is sufficient for all assessment domains or when no allowed non-redundant question remains.
fallback	proceed or investigate	Uses targeted fallback wording when no eligible question-pool item remains. If suspicion is confirmed, fallback is phrased as a consistency check.

Table 17: Mapping between Strategy Agent decisions and Evaluator actions.

Tables 18 and 19 list the feature-bank categories used by the Strategy Agent before computing the honesty scores described below.

ID	Concealment feature description	Tags
C01	Minimize or downplay core symptoms; frame distress as minor or manageable.	minimization
C02	Selective denial with leakage cues; deny problems while mentioning mild symptom fragments.	denial with leakage
C03	Avoid details about timeline, frequency, triggers, or severity.	evasiveness
C04	Positive self-presentation and impression management.	impression management
C05	Underreport risk-related content with short denials.	risk underreporting
C06	Normalize distress as stress or fatigue and distance from mental-health labels.	normalization
C07	Camouflaging or masking: appearing normal or competent while privately distressed.	camouflaging, masking
C08	Rehearsed social scripts or effortful passing.	social script, effortful passing
C09	Selective disclosure or intentionally hidden symptoms/diagnosis.	selective disclosure
C10	Stigma-driven hypervigilant self-monitoring.	self-stigma, self-monitoring

Table 18: Concealment-oriented feature-bank categories used for Strategy Agent honesty inference.

ID	Malingering feature description	Tags
M01	Over-endorsement across many symptom domains.	overendorsement
M02	Extreme absolute language such as always, completely, or cannot do anything.	extreme language
M03	Catastrophizing or dramatization without situational support.	catastrophizing
M04	Diffuse impairment expanding from focal issue to global failure.	diffuse impairment
M05	“More is worse” pattern, appending more symptoms to signal severity.	more-is-worse
M06	Function-symptom mismatch between claimed inability and sustained activity.	mismatch
M07	No fluctuation or missing natural variability.	no fluctuation
M08	Context reactivity around assessment, proof, ratings, or materials.	context reactivity
M09	Atypical or improbable symptom combinations.	fabrication-like
M10	Weak trauma narrative logic with timeline gaps or contradictions.	PTSD, fabrication-like
M11	Template-like flashback or nightmare description.	scripted narrative
M12	Jargon-heavy diagnostic language with few lived examples.	jargon-heavy
M13	Cross-source inconsistency between severe self-report and observed functioning.	inconsistency
M14	Sudden severe presentation without history.	severity-time mismatch
M15	Treatment-process gap despite severe suffering claims.	treatment-process gap
M16	Secondary-context emphasis around compensation, litigation, services, or documents.	secondary context
M17	Broadly elevated depression and anxiety with poor differentiation.	broad high affect
M18	Cognitive mismatch between severe claimed impairment and organized narrative.	cognitive mismatch

Table 19: Malingering-oriented feature-bank categories used for Strategy Agent honesty inference.

Honesty inference scoring. The Strategy Agent infers honesty only from evidence in the interview transcript. Let w_i denote the feature weight of matched feature i . Concealment and malingering scores are computed separately and clipped to $[0, 1]$:

$$C = \min(1, \sum_{i \in \mathcal{C}} w_i), \quad M = \min(1, \sum_{i \in \mathcal{M}} w_i).$$

The suspicion score and signed honesty score are:

$$S = \min(1, \max(C, M) + 0.25 \min(C, M)),$$

$$h_{\text{infer}} = \text{clip}(M - C, -1, 1).$$

Scoring element	Formal implementation
Input Matching	Visible evaluator–patient dialogue only. Deterministic rule/regular-expression matching to manifestation categories; no hidden label or LLM feature extraction.
Default feature	Weight 0.12 and confidence 0.55.
High-confidence feature	Weight 0.24 and confidence 0.85; the action-level high-confidence gate is 0.80.
Score aggregation	Concealment and malingering scores are accumulated separately and capped at 1; $S = \min(1, \max(C, M) + 0.25 \min(C, M))$.
Category gate	Direction-specific evidence must accompany the 0.24 score threshold; both confirmed directions yield mixed, and neither yields honest.
Interview action	At most one non-accusatory clarification or consistency rephrase is permitted in the formal configuration.

Table 20: Strategy-Agent suspicion scoring and action mapping.

Negative values indicate concealment-like evidence and positive values indicate malingering-like evidence. Because C and M are accumulated separately, the categorical inference can be honest, concealment, malingering, or mixed; confirmed inference requires thresholded evidence rather than one weak cue. The formal run contains 18 Strategy-inferred mixed cases although the hidden generator has no mixed class. Here mixed denotes competing observable reliability cues, not simultaneous behavior represented by the scalar generator.

Patient Agent details. In the formal workflow, the Patient Agent produces a structured self-report artifact rather than unconstrained free-form scale completion. For each selected self-report scale, the artifact contains item-level scores, evidence quotes, confidence values, and scale-level totals. These values are generated by deriving a profile-conditioned baseline symptom ratio and adjusting it with the hidden honesty value before conversion to scale responses. Honest cases remain near the baseline, concealment reduces endorsement in target domains, and malingering increases endorsement in claimed or exaggerated domains.

For clinician-rated interviewing, the Patient Agent uses LLM-based natural-language generation with a deterministic private strategy constructed from the hidden profile, honesty rationale, and modification summary. Honest patients follow the profile facts, concealment patients minimize or avoid sensitive details, and malingering patients exaggerate in the profile-indicated direction while remaining consistent with the assigned role.

Evaluator Agent details. When the Strategy Agent marks evidence as missing, vague, or suspicious, the Evaluator turns the target dimension into a focused follow-up question rather than continuing a fixed script. Under the validity-probe setting, these follow-ups target temporal consistency, functional impairment, preserved functioning, and contextual stability so that the interview gathers evidence useful for clinician-rated scoring.

After the interview, the Evaluator completes clinician-rated scales deterministically from evidence in the interview transcript. The resulting structured ClinicianRatedOutput contains clinician-rated totals, item evidence, and evidence summaries derived from the selected scale definitions and dialogue-grounded evidence.

Evaluator-mode sensitivity settings. In Table 7, Full, Fixed-scale, Adaptive, and Suspicion-aware all retain the same hybrid Strategy loop for question selection, evidence-slot updates, manifestation-feature matching, suspicion scoring, re-questioning, and stopping. The launcher varies only the Evaluator mode/state-prompt setting. Fixed-scale and Adaptive set the copied suspicion value in Evaluator state to zero, whereas Suspicion-aware and Full retain it; all modes still receive the structured Strategy decision in the question prompt. These rows therefore test mode/state-prompt sensitivity rather than removal of Strategy components. Item-only disables Strategy, while No-suspicion retains adaptive evidence tracking but disables the manifestation feature bank and suspicion probes. Suspicion can be activated by a high-confidence feature match or cumulative evidence above the 0.24 threshold, and at most one suspicion-driven rephrase is allowed. Table 21 makes the comparison groups explicit.

Setting	Group	Strategy	Evidence	Feature bank	Re-question	Eval. susp.
Item-only	Component ablation	No	No	No	No	None
No-suspicion	Component ablation	Yes	Yes	No	No	0
Full Strategy	Reference	Yes	Yes	Yes	Yes	Copied
Fixed-scale	Mode sensitivity	Yes	Yes	Yes	Yes	0
Adaptive	Mode sensitivity	Yes	Yes	Yes	Yes	0
Suspicion-aware	Mode sensitivity	Yes	Yes	Yes	Yes	Copied

Table 21: Component matrix for the ablation and Evaluator-mode comparisons. “Eval. susp.” is the suspicion value copied into Evaluator state; all mode-sensitivity rows still receive the structured Strategy decision.

E Diagnostic Fusion Details

Domain prediction. The four assessment domains are treated as candidate diagnosis domains rather than predicted labels. For each domain, the Diagnostician aggregates support from self-report ratios and item endorsement together with clinician-rated totals, item evidence, and evidence summaries. A domain is emitted only when the combined support exceeds the formal threshold $\tau = 2.0$, positive visible evidence exists, and no domain-specific veto condition is triggered.

Reliability gating. The Strategy Agent’s inferred honesty signal is used as a reliability gate rather than as a direct diagnosis label. Honesty-aware gating is activated only when both suspicion and confidence are sufficiently high. Under inferred malingering, self-report-only positive evidence is discounted and clinician-rated corroboration is required. Under inferred concealment, low or minimized self-report does not by itself veto a domain, allowing clinician-rated evidence to rescue under-reported symptoms. If core clinician-rated evidence is negative and no reliable self-report positive evidence exists, the output is *healthy*.

Severity and output. Severity is estimated only for predicted domains and is derived from visible scale scores rather than hidden benchmark-reference severity. The Diagnostician maps self-report ratios and clinician-rated totals to severity candidates, averages the two sources when both are available, and clamps the final value to the 1–10 range. The resulting structured `DiagnosisReport` contains predicted diagnosis domains, severity, evidence quality, limitations, and the Strategy-derived honesty inference.

F Dataset Details

F.1 Dataset Instance Schema

The full experimental session artifact comprises a visible patient view, assessment outputs, dialogue

traces, strategy state, and the final diagnostic report; Table 23 documents this schema. The public repository provides code, prompts, sanitized formal and ablation configurations, deterministic toy/mock examples, aggregate results, checksums, anonymized human-evaluation materials, and documentation. Complete sessions, processed profiles, case-level manifests, and provider traces are not publicly released; Table 22 summarizes the available verification scope.

Artifact class	Public	Verification scope
Implementation and batch scripts	Yes	Inspect the five-agent workflow and execution entry points.
Prompts and configurations	Yes	Inspect role instructions, formal settings, and ablation settings.
Toy/mock examples, schemas, and tests	Yes	Run deterministic smoke tests and validate output structure.
Metric scripts, aggregates, and checksums	Yes	Inspect metric implementations and verify reported tables from aggregate artifacts.
Human-evaluation materials	Yes	Inspect anonymized questionnaires, scope, and aggregate results.
Complete sessions, processed profiles, and case-level manifests	No	Not released in the current public repository.
Raw provider traces, hidden reasoning, source records, and credentials	No	Excluded from the public release.

Table 22: Contents and reproducibility scope of the public camera-ready repository.

Component	Meaning
public_patient_view	Profile fields visible to assessment-side agents, including demographics, chief complaint, past medical history, and family history.
selected_scales	Self-report and clinician-rated scales selected by the Assessor Agent.
self_report	Self-report scale outputs completed by the Patient Agent.
dialogue	Multi-turn Evaluator–Patient interview, with speaker, turn identifier, text, and metadata.
strategy_trace	Strategy Agent decision sequence, selected questions, target dimensions, reasons, and honesty inference.
evidence_slots	Structured interview evidence for clinician-rated assessment dimensions.
clinician_rated	Clinician-rated scale outputs generated from the interview.
diagnosis_report	Diagnostician output containing predicted diagnoses, severity values, evidence quality, and limitations.
metadata	Run-level metadata, configuration, model, quality information, and file bookkeeping.

Table 23: Top-level dataset instance schema.

F.2 Annotation Schema

The full internal session artifacts store profile labels, scale outputs, dialogue metadata, honesty labels, and diagnosis labels as structured fields. Table 24 summarizes this schema; these case-level annotations are not included in the public repository.

Annotation type	Fields
Profile labels	Source dataset, source condition, hidden diagnosis, hidden severity, and hidden honesty value.
Scale labels	Selected scales, self-report scores, and clinician-rated scores.
Dialogue labels	Speaker, turn ID, text, and strategy metadata.
Honesty labels	Hidden honesty, inferred honesty, suspicion score, concealment score, malingering score, confidence, and matched features.
Diagnosis labels	Predicted diagnosis domains, predicted severity, evidence quality, and limitations.

Table 24: Annotation schema of the full experimental session artifacts.

F.3 Unavailable or Non-flattened Labels

The dataset does not include turn-level patient response strategy labels. It does not include fixed response type labels. Question type is partially recoverable from strategy metadata and the question pool, but is not provided as a flattened turn-level label. Human expert validation labels are not part of the dataset annotation schema.

F.4 Additional Dataset Statistics

Tables 25 and 26 report compact distributional statistics for the formal 2008-case run.

Label	Hidden	Inferred
Honest	1004	1200
Concealment	252	231
Malingering	752	559
Mixed	–	18

Table 25: Additional honesty statistics for the formal 2008-case run. Concealment and malingering hidden counts aggregate mild, moderate, and severe variants.

Predicted diagnosis label	Count
Depression	951
Anxiety	724
ADHD	432
Bipolar	207
Healthy	483
Undetermined visible evidence	476

Table 26: Predicted diagnosis-label counts. Counts are multi-label and therefore do not sum to the number of cases.

F.5 Sensitive Content and Prohibited Uses

A lexical audit of the full internal artifacts found synthetic self-harm- or suicide-related language in 193 of 2,008 sessions (9.61%) and substance-use-related language in 16 sessions. No per-session clinical safety label is provided. The Evaluator prompt requests gentle, direct wording for sensitive questions, but the workflow does not implement a validated crisis-escalation or referral protocol. These materials therefore require the use restrictions in Table 27.

Not intended / prohibited use	Boundary
Real-patient diagnosis, screening, triage, or treatment planning	Not for clinical decision-making.
Self-harm/suicide-risk assessment or crisis response	Not a crisis-response system.
Real-patient deception, honesty, concealment, or malingering detection	Not for patient-level honesty labeling.
Insurance, legal, employment, immigration, benefits, or school-discipline decisions	Not for high-impact decisions about individuals.
Replacement for licensed-clinician judgment	Benchmark labels are not medical determinations.

Table 27: Prohibited-use boundaries for the framework and released artifacts.

G Train/Development/Test Split Details

G.1 Split Files

Table 28 summarizes the split artifacts used for the experiments and the resulting case counts. These case-level split files are not part of the public camera-ready repository. The split is defined at the grouped source-profile level rather than by randomly splitting sessions.

(a) Split artifacts

Artifact	Content
Split assignments	Per-case split assignment in JSONL and CSV formats.
Split manifest	Manifest with split field.
Training manifest	Training cases.
Development manifest	Development cases.
Test manifest	Test cases.
Split summary	Split method, counts, leakage check, and distributions.

(b) Split counts

Split	Cases	Percent
Train	1408	70.12%
Development	300	14.94%
Test	300	14.94%

Table 28: Train/development/test split artifacts and case counts.

G.2 Split Unit and Leakage Prevention

The split unit is the pair of source dataset and source record ID. Cases are not randomly split at the session level. Instead, variants derived from the same source profile are assigned to the same split. This grouping prevents honest and non-honest variants of the same source record from crossing split boundaries. The split summary reports zero group cross-split violations.

G.3 Stratification Fields

The grouped split is stratified first by source dataset and then balanced over non-honest honesty

value buckets, diagnosis combinations, and severity bins. The split contains 1004 source-profile groups, each with two generated variants.

G.4 Per-Split Distributions

Table 29 reports the main fields used to inspect balance across the train, development, and test splits.

Field	Value	Train	Dev.	Test	All
Source condition	Clinical	704	150	150	1004
	Healthy honest	352	75	75	502
	Healthy malingering	352	75	75	502
Severity bin	Low 1–3	704	150	150	1004
	Mid 4–6	192	60	48	300
	High 7–10	512	90	102	704
Hidden honesty	Honest	704	150	150	1004
	Mild concealment	60	12	12	84
	Moderate concealment	56	13	11	80
	Severe concealment	61	14	13	88
	Mild malingering	180	38	38	256
	Moderate malingering	177	36	40	253
	Severe malingering	170	37	36	243
Diagnosis combination	ADHD + Anxiety	38	36	36	110
	ADHD + Anxiety + Depression	160	12	12	184
	ADHD + Depression	100	26	24	150
	Anxiety + Bipolar + Depression	78	34	34	146
	Anxiety + Depression	282	0	0	282
	Bipolar + Depression	46	42	44	132
	Healthy	704	150	150	1004

Table 29: Per-split distributions for the grouped train/development/test split.

H Supplementary Experimental Analyses

H.1 Assessment Fusion across Honesty Conditions

This section supplements the main assessment-fusion results by reporting assessment-stage deviation across diagnosis domain, severity bin, and exact honesty value. Tables 30, 31, 32, and 34 use the same breakdown metrics as the main assessment-fusion table: SRD is self-report deviation, CRD is clinician-rated deviation, FD is vector-penalized final diagnosis deviation, and Gap is signed self-report minus clinician-rated deviation. Table 33 is reported separately because it summarizes paired statistical comparisons rather than condition-level breakdowns.

Condition	Dom.	N	SRD	CRD	FD	Gap
Clinical concealment	AD	252	1.25	1.95	1.32	0.61
Clinical concealment	ADHD	252	0.39	–	0.51	–
Clinical concealment	BD	252	0.38	1.80	1.21	0.24
Clinical concealment	MDD	252	1.83	2.49	1.42	1.75
Clinical honest	AD	502	1.30	1.80	0.83	1.85
Clinical honest	ADHD	502	0.41	–	0.47	–
Clinical honest	BD	502	0.38	2.01	1.21	0.01
Clinical honest	MDD	502	1.71	2.43	1.05	3.21
Clinical malingering	AD	250	2.16	1.68	1.07	2.76
Clinical malingering	ADHD	250	1.18	–	0.74	–
Clinical malingering	BD	250	0.66	2.00	1.30	0.57
Clinical malingering	MDD	250	2.21	2.44	1.06	3.61

Table 30: Assessment-fusion deviation by diagnosis domain for clinical cases.

Condition	Dom.	N	SRD	CRD	FD	Gap
Healthy honest	AD	502	0.00	0.42	0.02	-0.42
Healthy honest	ADHD	502	0.00	–	0.02	–
Healthy honest	BD	502	0.00	0.11	0.00	-0.11
Healthy honest	MDD	502	0.00	0.62	0.10	-0.62
Healthy malingering	AD	502	4.02	0.58	0.46	3.43
Healthy malingering	ADHD	502	0.25	–	0.17	–
Healthy malingering	BD	502	0.00	0.13	0.01	-0.13
Healthy malingering	MDD	502	4.11	0.83	0.57	3.28

Table 31: Assessment-fusion deviation by diagnosis domain for healthy cases.

Bin	N	SRD	CRD	FD	Gap	Imp-F
1–3	1004	1.05	0.45	0.17	0.90	0.88
4–6	544	1.69	1.89	0.89	2.95	0.80
7–10	460	2.08	2.80	1.10	4.32	0.97

Table 32: Assessment-fusion deviation by severity bin.

Comparison	Diff.	95% CI	<i>p</i>	N
SRD vs. CRD	0.08	[0.01, 0.14]	0.0186	2008
SRD vs. FD	0.88	[0.83, 0.93]	< 0.001	2008
CRD vs. FD	0.80	[0.75, 0.85]	< 0.001	2008

Table 33: Paired assessment-stage comparisons for assessment fusion. Positive mean differences indicate improvement in the later-stage estimate.

Honesty	N	SRD	CRD	FD	Gap	Imp-C	Imp-F
Mild concealment	84	1.10	2.47	0.74	3.22	-1.37	0.36
Moderate concealment	80	1.93	2.44	1.27	1.49	-0.51	0.66
Severe concealment	88	1.93	2.77	1.34	1.63	-0.83	0.60
Honest	1004	0.82	1.30	0.46	1.71	-0.48	0.36
Mild malingering	256	1.21	1.04	0.45	1.76	0.16	0.76
Moderate malingering	253	2.69	1.19	0.62	3.49	1.49	2.06
Severe malingering	243	2.87	1.02	0.58	3.75	1.85	2.29

Table 34: Assessment-fusion deviation by exact honesty condition.

H.2 Ablation Study: Honesty Robustness

Table 35 reports the honesty-inference portion of the ablation study. Disabling suspicion removes non-honest detection, while the full strategy detects both concealment and malingering with a low false-suspicion rate.

Setting	Conceal.	Maling.	False Susp.
Full Strategy	60.0%	45.0%	1.2%
No-suspicion	0.0%	0.0%	0.0%

Table 35: Honesty-robustness portion of the ablation study. Conceal. and Maling. denote detection rates for concealment and malingering cases; False Susp. denotes false suspicion on hidden honest cases.

Metric	Our Generated	PsyCoTalk
Dialogues	2,008	3,000
Avg. rounds/dialog	14.17	45.88
Evaluator/doctor chars	102.6	34.7
Patient chars	207.4	43.4
Clinical entity coverage	0.166	0.888
BERTScore F1	0.529	0.638
PPL	34.01	9.31

Table 36: Automatic dialogue-quality comparison with PsyCoTalk. Character counts are averaged per turn; lower perplexity (PPL) indicates better language-model fluency.

H.3 Automatic Dialogue Quality Metrics

Table 36 gives the compact comparison formerly reported in the main text. Table 37 reports additional corpus-level automatic descriptors for the 2,008 generated dialogues. These metrics are descriptive checks of dialogue length, lexical diversity, semantic overlap, and language-model fluency; they are not used as clinical correctness labels.

Taken together, the automatic metrics indicate a trade-off rather than uniform quality superiority. The generated sessions are shorter and organized around controlled evidence gathering, but are less similar to the source dialogue corpus on entity coverage, semantic-overlap, and fluency proxies.

Metric	Value
Dialogues	2,008
Average rounds per dialogue	14.17
Average evaluator turn length	102.6 chars
Average patient turn length	207.4 chars
Distinct-1 / Distinct-2, all utterances	0.000009 / 0.000175
Clinical entity coverage, macro average	0.1655
self-BLEU-2 / self-ROUGE-L, 5k sample	0.7925 / 0.4541
BOW average / extrema cosine, 10k sample	0.9267 / 0.8063
BERTScore F1, 2k sample	0.5288
Perplexity	34.01

Table 37: Additional automatic dialogue-quality descriptors for the formal 2,008-case generated-dialogue corpus.

I Human Evaluation Details

I.1 Human Evaluation 1: Dialogue-Level Evaluation

We conducted the first human evaluation on the Credamo platform and recruited participants with psychology, psychiatry, clinical medicine, nursing, counseling, or related health-care backgrounds. Participants were required to provide informed consent and complete a paid survey. Before the formal task, participants were presented with an anonymized study overview, an estimated completion time of approximately 30 minutes, confidentiality assurances, voluntary-participation instructions, and task requirements. The evaluation was administered in Chinese; no research-team names, contact addresses, payment amounts, participant names, or location identifiers are reported in this paper.

The evaluation consisted of two parts. First, participants completed background and eligibility questions, including highest education level, professional background, specialty or free-text occupation description, years of clinical or related experience, and four embedded domain-screening questions covering MDD, anxiety disorder, ADHD, and bipolar disorder. Second, participants completed the formal evaluation task. Each participant rated six anonymized cases, and each case contained 17 five-point Likert-scale questions covering dialogue quality, patient communication state, honesty-related behavior, and diagnosis/evidence plausibility. For the honesty-specific items, raters were shown the assigned synthetic target category and judged whether the visible be-

havior matched it; HE1 was therefore not a blind honesty-inference task and did not ask raters to estimate the numeric value of h . The dialogue-quality prompts covered clinical logic, factual reliability, fluency, and naturalness. All ratings used a 1–5 scale, where higher scores indicated greater plausibility or quality. These judgments do not establish a participant’s true honesty status, independent clinical diagnosis, clinical utility, or preference for fused over self-report-only or clinician-rated-only summaries. They were used only for aggregate human evaluation and are not part of the public case-level annotation release.

Eval.	Shown and judged	Sample	Interpretive boundary
HE1	Dialogue, assigned target category, and final-output materials; plausibility and consistency judgments	150 sessions; 133 retained participants; 13,566 item ratings	Not blind to target category; not validation of true honesty, independent diagnosis, clinical utility, or fused-summary preference.
HE2	Profile plus displayed diagnosis/severity labels; profile-label consistency	60 profiles; 39 participants; 933 ratings	Not an independent diagnostic or blind severity-assignment task.

Table 38: What the two human evaluations validate and do not validate.

Applicable item/condition	Cases	Ratings	Mean [95% CI]	≥ 4
Q7 / honest	60	320	4.194 [4.100, 4.286]	85.9%
Q8 / concealment	30	161	3.910 [3.806, 4.013]	80.7%
Q9 / malingering	60	317	3.926 [3.835, 4.016]	79.2%

Table 39: Condition-appropriate HE1 reanalysis. Confidence intervals use 5,000 case-level bootstrap resamples. This is a post-hoc descriptive analysis; raters saw the assigned target category.

I.2 Human Evaluation 2: Severity-Label Validation

We used a second Credamo human-evaluation task to validate whether the hidden diagnosis-severity labels assigned to synthetic patient profiles were supported by the profile descriptions. The same anonymization and consent principles were used, including an approximately 30-minute expected time commitment, professional-background screening, and embedded domain-screening questions. Participants read anonymized patient profiles and rated the consistency of the provided severity labels for four diagnostic dimensions: MDD, anxiety disorder, bipolar disorder, and ADHD. This task evaluated label support rather than asking respondents to produce an independent clinical diagnosis or treatment recommendation.

The exported human-evaluation file contained 39 submitted responses. All respondents answered

the four embedded domain-screening questions correctly, so all submitted responses were retained for aggregate validation. Each respondent rated six patient profiles, with four diagnostic-domain severity labels per profile. The validation set covered 60 sampled profiles and yielded 933 non-missing 1–5 consistency ratings. Across all ratings, the mean score was 4.109/5, and 83.7% of ratings were at least 4.

Domain	Ratings	Mean	≥ 4
MDD	234	4.184	91.5%
Anxiety	233	3.987	81.5%
Bipolar	233	4.167	82.0%
ADHD	233	4.099	79.8%
All	933	4.109	83.7%

Table 40: Human Evaluation 2 severity-label validation results by diagnostic domain.

I.3 Participants and Supplemental Survey Statistics

For Human Evaluation 1, we recruited 150 participants via the Credamo platform. After excluding 17 low-quality responses, the final analysis included 133 valid participants. All retained participants correctly answered the four embedded domain-screening questions before entering the formal evaluation task.

The final Human Evaluation 1 sample included participants from psychiatry, psychological counseling, health-related student backgrounds, nursing, clinical medicine, rehabilitation, pharmacy, radiology, social work, and other related medical or mental-health backgrounds. In the retained sample, 45 participants were psychiatrists, 41 were psychological counselors, 34 were students in related fields, and 13 reported other related professional backgrounds. The average completion time was 24.29 minutes (SD = 8.22). Among participants with parseable experience-year responses, the average professional experience was 7.33 years (SD = 8.13).

Human Evaluation 2 respondents included 25 psychiatrists, 6 psychological counselors, 3 health-related students, and 5 participants from other related clinical or health-care backgrounds. Their average completion time was 20.41 minutes (SD = 7.03), and their average reported professional experience was 8.79 years (SD = 6.58). Table 41 summarizes key supplemental human-evaluation statistics not reported in the main text.

Statistic	HE1	HE2
Submitted responses	150	39
Retained responses	133	39
Primary rating target	Dialogues	Profile labels
Rated cases/profiles	150	60
Total item ratings	13,566	933
Mean completion time (min.)	24.29	20.41
Completion-time SD (min.)	8.22	7.03
Mean experience (years)	7.33	8.79
Experience SD (years)	8.13	6.58

Table 41: Supplemental statistics for the two human-evaluation tasks. HE1 is the dialogue-level evaluation; HE2 is the profile severity-label validation.

Age, gender, and city-level location were not available in the de-identified human-evaluation exports and are therefore not reported. Although the sample was recruited from Chinese participants, the task specifically targeted participants with relevant psychology, psychiatry, clinical, nursing, or health-care backgrounds rather than the general population.

J Case Studies

We report four representative cases from the same formal 2008-case full-strategy run used in the quantitative analyses. The examples are intentionally compact: they support the quantitative findings rather than serving as a separate qualitative study. To avoid exposing unnecessary clinical narrative, we summarize structured labels, dialogue-level signals, predictions, and anonymized artifact identifiers instead of reproducing raw dialogue. For double-blind review, artifacts are referenced by anonymized case IDs rather than raw session filenames. The case IDs denote the qualitative role of each example: Case-HM-FP is a false positive under healthy malingering, Case-CR is a concealment-recovery case, Case-ASC is an ambiguous subclinical case, and Case-BF is a bipolar failure case.

J.1 False Positive under Healthy Malingering

Patient profile. The evaluation-only hidden profile is healthy, with severe malingering assigned as the private honesty condition ($h = 0.8$).

Key dialogue signals. The Strategy Agent asked 13 questions, issued one consistency check, and inferred malingering from the interview transcript (inferred score 0.24; suspicion score 0.39). Matched feature-bank cues included extreme language and secondary-context emphasis, with a

weaker impression-management cue in the opposite direction.

Model prediction. Despite the hidden healthy label, the final prediction contained depression severity 6, anxiety severity 5, and bipolar severity 1. The evaluation record marks this as a healthy false positive, with three extra domains and zero leakage warnings.

Why it matters. This case is a useful failure mode for the paper: reliability inference correctly detected malingering-like evidence, but the final scale-based diagnostic integration still over-called clinical domains. It supports reporting false-positive behavior under healthy malingering rather than presenting the system as uniformly robust.

Artifact. Case-HM-FP.

J.2 Concealment Recovery

Patient profile. The evaluation-only hidden labels indicate severe concealment ($h = -0.8$) with depression, anxiety, and bipolar domains.

Key dialogue signals. The Strategy Agent asked 14 questions, issued one consistency check, and inferred concealment with a signed score of -0.72 and suspicion score 0.87. Matched evidence included selective disclosure, anticipated stigma, denial-with-leakage, normalization, and evasiveness features.

Model prediction. The final output predicted depression severity 7, anxiety severity 6, and bipolar severity 6. This exactly recovered the hidden domain set, achieved severity-tolerant joint accuracy at 1, and produced zero leakage warnings.

Why it matters. This case supports the main honesty-aware claim: even when the patient minimizes or selectively discloses symptoms, the multi-agent pipeline can still recover clinically relevant domains through structured questioning and scale integration.

Artifact. Case-CR.

J.3 Ambiguous Subclinical Case

Patient profile. The hidden profile is healthy, with mild malingering assigned as the private honesty condition ($h = 0.2$).

Key dialogue signals. The Strategy Agent asked 13 questions and inferred mixed honesty rather than forcing a single direction. The feature scores were close but not identical: concealment 0.36, malingering 0.48, suspicion 0.57. Matched cues included both malingering-like evidence and minimization/normalization-style evidence.

Model prediction. The final output was an undetermined visible-evidence label. In the evaluation vector, no clinical domain was assigned, matching the hidden healthy profile; leakage warnings were zero.

Why it matters. This case illustrates why uncertainty is useful. Mixed reliability evidence did not automatically become a clinical diagnosis, and the system remained conservative when visible evidence was subclinical or internally mixed.

Artifact. Case-ASC.

J.4 Bipolar Failure Case

Patient profile. The evaluation-only hidden labels indicate severe malingering ($h = 0.8$) with depression, anxiety, and bipolar domains.

Key dialogue signals. The Strategy Agent asked 14 questions, issued one consistency check, and inferred malingering with score 0.60 and suspicion score 0.75. The diagnostic evidence quality recorded bipolar conflict and low-confidence clinician-rated scale estimates.

Model prediction. The final prediction retained only depression severity 6. In the evaluation vector, anxiety and bipolar were missing, giving a false domain set and a joint vector MAE of 3.5; leakage warnings were zero.

Why it matters. This case makes the limitations concrete. Bipolar recovery remains a hard setting when scale evidence conflicts and visible presentation is distorted, so the appendix does not only cherry-pick successful honesty-aware recovery cases.

Artifact. Case-BF.

J.5 Worked Example: Concealment Recovery

We include one worked example to make the case-study format more concrete without exposing raw session files or personally identifying information.

This example corresponds to Case-CR. It is selected because it shows the intended division of labor across agents: the patient can under-disclose, the strategy module can notice the pattern from visible evidence, the evaluator can ask neutral follow-ups, and the final diagnostic integration can still recover the major clinical domains.

Example Input: Case-CR

```
Input:
- Public patient view:
  demographics: young adult male writer
  chief complaint: one year of emotional instability,
  nervousness, irritability, somatic anxiety, social
  anxiety,
  low self-esteem, self-blame, and loneliness
  history: prior mood/anxiety treatment, self-harm-related
  emergency care without psychiatric hospitalization,
  hypertension, and ongoing psychotropic medication
  family history: mood-disorder history
- Private patient state:
  hidden domains: depression; anxiety; bipolar disorder
  hidden honesty: severe concealment, h = -0.8
  behavior directive: guarded disclosure; symptom
  minimization;
  selective detail avoidance; first-person role consistency
- Assessment setup:
  candidate domains: depression; anxiety; bipolar disorder;
  ADHD
  self-report scales: HADS, KADS, GAD-7, SPIN-17,
  MDQ, BEDS, ASRS-v1.1, WFIRS-S
  clinician-rated scales: HAM-D, HAM-A, BDRS, ADHD-RS-IV
  boundary: hidden labels and honesty values visible only to
  the Patient Agent
- Self-report input:
  private profile facts, hidden honesty directive, selected
  self-report scale definitions
  expected artifact fields: item scores, evidence quotes,
  confidence values, scale totals, and severity labels
- Clinician-rated input:
  selected clinician-rated scale definitions, interview
  transcript,
  evidence slots, and Strategy guidance
  excluded inputs: hidden labels, hidden honesty value, and
  private patient state
```

Example Output: Case-CR

```
Output:
- Interaction summary:
  Assessor selects a four-domain assessment battery covering
  depression,
  anxiety, bipolar disorder, and ADHD. The patient answers
  in
  ordinary first-person language but minimizes sensitive
  details.
  The evaluator asks 14 questions and one neutral
  consistency
  check; evidence slots become sufficient for depression,
  anxiety, and bipolar dimensions.
- Self-report output:
  HADS 15/15, severe; KADS 21/21, severe
  GAD-7 16/24, marked; SPIN-17 6/9, marked
  MDQ 15/24, moderate; BEDS 6/9, marked
  ASRS-v1.1 0/18, minimal/negative; WFIRS-S 0/21,
  minimal/negative
  summary: guarded or minimized endorsement for mood,
  anxiety, and bipolar-domain symptoms
- Clinician-rated output:
  HAM-D 25/52, severe depression; HAM-A 16/56,
  moderate anxiety
  BDRS 30/60; ADHD-RS-IV 18; all rater scales complete
  suspicion score at rating: 0.87
  summary source: visible interview evidence only
- Strategy inference:
  category: concealment
  signed honesty score: -0.72
  suspicion score: 0.87
  concealment score: 0.84
  malingering score: 0.12
  matched features: selective disclosure; anticipated
  stigma;
  denial with leakage; normalization; evasiveness
- Final diagnosis:
  depression severity 7
  anxiety severity 6
  bipolar severity 6
  no ADHD diagnosis
- Evaluation outcome:
  hidden domain set recovered
  severity-tolerant joint accuracy at 1
  zero leakage warnings
Why useful:
  The case shows how structured questioning and reliability
  inference can recover major clinical domains even when the
  patient under-discloses symptoms. The final diagnosis uses
  visible self-report output, clinician-rated evidence,
  selected
  scale metadata, and Strategy feature-bank inference rather
  than hidden labels.
Artifact: Case-CR
```

K Psychometric Scale Repository

Table 42 details the psychiatric instruments employed in our system's repository, categorized by diagnostic domain and administration type.

Abbr.	Full Name	N	Scoring / Cut-off
Depression			
<i>Clinician-rated scales</i>			
HAM-D	Hamilton Rating Scale for Depression	21	0–7 Normal; > 19 Severe
MADRS	Montgomery–Åsberg Depression Rating Scale	10	0–6 Normal; > 34 Severe
<i>Self-report scales</i>			
BDI	Beck Depression Inventory	21	0–10 Normal; > 30 Severe
BHS	Beck Hopelessness Scale	20	Cut-off ≥ 9
CDI	Children’s Depression Inventory	27	> 12 Mild; > 38 Severe
CES-D	Center for Epidemiological Studies–Depression Scale	20	Cut-off ≥ 16
CES-DC	CES-D for Children	20	Cut-off ≥ 16
CESD-R	CES-D Revised	20	> 10 Mod.; > 22 High
DASS-21	Depression Anxiety Stress Scales–21	21	Severity ratings per subscale
EPDS	Edinburgh Postnatal Depression Scale	10	> 9 Possible; > 13 Probable
GDS-15	Geriatric Depression Scale–15	15	≥ 5 Mild; ≥ 10 Severe
HADS	Hospital Anxiety and Depression Scale	14	≥ 8 Borderline; ≥ 11 Clinical
KADS	Kutcher Adolescent Depression Scale	11	> 10 Mild; > 20 Mod./Sev.
MDI	Major Depression Inventory	12	ICD-10 Algorithm
MFQ-Self	Mood and Feelings Questionnaire–Self Report	13	> 7 Mild; > 15 Mod./Sev.
PHQ-9	Patient Health Questionnaire–9	9	5/10/15/20 severity thresholds
SDS	Zung Self-Rating Depression Scale	20	≥ 50 Mild; ≥ 60 Mod.; ≥ 70 Sev.
Anxiety			
<i>Clinician-rated scale</i>			
HAM-A	Hamilton Anxiety Rating Scale	14	> 14 Mod.; > 23 Sev.
<i>Self-report scales</i>			
BAI	Beck Anxiety Inventory	21	0–21 Low; 22–35 Mod.; 36+ Sev.
GAD-7	Generalized Anxiety Disorder–7	8	5/10/15 severity thresholds
PSWQ	Penn State Worry Questionnaire	16	> 39 Mod.; > 59 Sev.
SPIN-17	Social Phobia Inventory	17	> 19 Mild; > 30 Mod.; > 40 Sev.
STAI-State	State–Trait Anxiety Inventory, State Scale	20	> 39 Mod.; > 59 High
Bipolar			
<i>Clinician-rated scales</i>			
BDRS	Bipolar Depression Rating Scale	20	Higher scores indicate greater severity
YMRS	Young Mania Rating Scale	11	8–12 Hypomania; ≥ 20 Mania
<i>Self-report scales</i>			
BEDS	Barcelona Bipolar Eating Disorder Scale	10	Cut-off ≥ 13
BSDS	Bipolar Spectrum Diagnostic Scale	20	≥ 13 Mod.; ≥ 20 High
HCL-32	Hypomania Checklist–32	32	Cut-off ≥ 14
MDQ	Mood Disorder Questionnaire	15	Algorithmic positive screen
ADHD			
<i>Clinician-rated scale</i>			
ADHD-RS-IV	ADHD Rating Scale–IV	18	0–54 Total; two 0–27 subscales
<i>Self-report scales</i>			
ASRS-v1.1	Adult ADHD Self-Report Scale v1.1, Part A Screener	6	4+ positive screen
WFIRS-S	Weiss Functional Impairment Rating Scale–Self Report	69	Domain-wise impairment; no universal cut-off

Table 42: Psychometric scales used in the formal Assessor Agent configuration. Each formal session uses eight self-report scales and four rater-administered scales selected from this 32-scale repository.