BENCHMARKING INTERACTIVE CLINICAL DIAGNOSIS UNDER PATIENT STIGMA AND SELECTIVE INFORMATION DISCLOSURE

Anonymous authors
Paper under double-blind review

ABSTRACT

The focus of medical artificial intelligence (AI) is shifting from static knowledge reasoning to interactive clinical diagnosis. However, existing frameworks assume unconditionally cooperative patients who transparently disclose clinical facts, contradicting real-world consultations where disease stigma, privacy concerns, and fragile interpersonal trust induce selective information disclosure. We propose S^2InQBench, the first benchmark that formalizes interactive diagnosis as an imperfect-information, general-sum game to evaluate medical AI under disease-stigma awareness. By mining public medical corpora, we first normalize disease stigma by 185 canonical Clinical Stigma Units (CSUs) with scaled relative sensitivity. Then, we construct a neuro-symbolic agentic patient that mandatorily invokes a dedicated tool before responding to determine its disclosure posture towards doctor-queried information. This tool dynamically tracks the patient's trust in the doctor via an evolving Beta belief distribution and stochastically samples disclosure strategies based on the posterior tail probability of trust overcoming stigma sensitivity, thereby mirroring human bounded rationality and emotional hesitation over rigid cognitive principles. Evaluations across 13 frontier LLMs reveal that ideal-patient evaluations overestimate the clinical robustness of current AI, evidenced by degraded inquiry efficiency, trust-building, and diagnostic accuracy under selective-disclosure postures. Dedicated agentic scaffolding improves sensitive-evidence elicitation, yet diagnostic accuracy remains limited, underscoring the need for trust-aware, reliable clinical AI.

1 INTRODUCTION

Clinical medicine is fundamentally an iterative process of active information elicitation, wherein clinicians presented with an incomplete initial presentation must proactively formulate targeted inquiries, navigate diagnostic uncertainty, and adaptively synthesize evolving patient clues (Wong et al., 2026; Zhou et al., 2026). Reflecting this diagnostic reality, the focus of medical Large Language Models (LLMs) is shifting from passive knowledge retrieval evaluated on static examinations (Jin et al., 2021; Liu et al., 2023) toward interactive, agentic clinical inquiry (Ding et al., 2026; Jia et al., 2025; Rose et al., 2025). Notably, pioneering systems such as AMIE (Tu et al., 2024; Saab et al., 2025; Brodeur et al., 2026) have demonstrated the viability of conversational diagnostic consultations, spurring a wave of interactive benchmarks that evaluate models within simulated Objective Structured Clinical Examinations (OSCE) (Yao et al., 2026) and multi-turn outpatient environments (Schmidgall et al., 2024; Li et al., 2024; Liu et al., 2024; Qiao et al., 2026).

To enable evaluation without human-in-the-loop overhead, existing benchmarks deploy LLM-driven patient simulators to mimic outpatient encounters. Methodologically, these simulators range from fact-retrieval prompts (Li et al., 2024; Schmidgall et al., 2024) and state-tracking memory banks (Liao et al., 2024) to fine-grained Atomic Information Unit (AIU) disclosure controllers (Gong et al., 2025; Qiao et al., 2026). Despite their engineering sophistication, **all existing benchmarks rest upon a fatally idealized premise: the assumption of the idealized, unconditionally cooperative patient.** Under this paradigm, simulated patients are treated as passive, transparent database oracles that immediately and honestly disclose their internal ground-truth clinical facts whenever prompted (Schmidgall et al., 2024; Gong et al., 2025). This sterile "high-signal-to-noise" assumption funda-

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

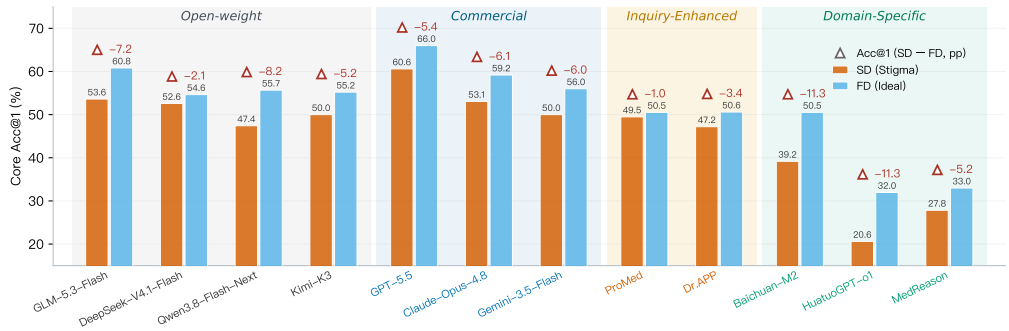


Figure 1: Top-1 diagnostic accuracy on S^{2}InQBench-Core under the idealized stigma-driven selective disclosure and ideal full-disclosure conditions. Performance degradation across all models.

mentally contradicts real-world clinical practice (Brodeur et al., 2026). In clinical reality, diagnostic inquiries do not occur in a vacuum; medical encounters are deeply laden with psychological vulnerability, emotional apprehension, and defensive information management. Consequently, current benchmarks reward aggressive, mechanical interrogations while completely overlooking a physician's bedside manner, creating a dangerous *illusion of clinical competence* where models achieve inflated diagnostic accuracies in benchmarks but degrades substantially when facing hesitant, real-world human patients.

This critical disparity is governed by **disease stigma and active privacy management**—a ubiquitous phenomenon formalized in the Health Stigma and Discrimination Framework (HSDF) (Stangl et al., 2019; Earnshaw & Chaudoir, 2009) and the Cognitive-Affective Model of Concealable Stigma (Pachankis, 2007; Chaudoir & Fisher, 2010). In medical consultations, health conditions are rarely perceived as neutral physiological parameters. When inquiries encroach upon sensitive lifestyles, sexual health, or anatomical taboos, patients experience acute self-stigma, moral apprehension, and fear of healthcare discrimination (Corrigan, 2004; Link & Phelan, 2001). More importantly, clinical disclosure is not a passive reflex, but a dynamic, trust-gated psychological decision procedure. When confronted by a cold, judgmental, or prematurely intrusive clinician, patients instinctively engage in selective disclosure—actively withholding sensitive etiologies, minimizing severity, evading questions, or outright fabricating benign alibis (Pachankis, 2007). Eliciting such critical, high-stigma evidence demands strategic clinical inquiry, such as establishing a therapeutic alliance through validated empathy, normalizing taboo behaviors, and ensuring confidentiality. Therefore, an ecologically valid benchmark must evaluate whether a medical AI system can navigate this delicate boundary between diagnostic information-seeking and interpersonal trust-building.

To bridge this foundational gap, we propose S^{2}InQBench, an interactive clinical diagnosis benchmark that evaluates AI doctors under realistic patient disease stigma and selective information disclosure. We formalize clinical consultation as an **imperfect-information, general-sum game**, which explicitly models the strategic tension between the physician's pursuit of diagnostic certainty and the patient's psychological defense of privacy. To support ecologically valid evaluation, we first normalize disease stigma using **185 distinct Clinical Stigma Units (CSUs)** mined from extensive public medical corpora, and scale their relative sensitivity into a unified metric space $\mathcal{S} \in [0.05, 0.95]$ by employing a **Bradley-Terry pairwise comparison model** grounded on psychometric rubrics. A neuro-symbolic patient agent is then introduced to conduct autonomous evaluations: using a dedicated tool, it parses doctor utterances into clinical communicative acts and updates an evolving Beta distribution to model interpersonal trust. By incorporating prior-regressive memory decay and quadratic trespassing penalties, this tool computes the posterior probability that the patient will overcome stigma to determine its response strategy: Full Disclosure (FD), Partial Omission (PO), Evasion/Redirection (ER), or Distortion/Fabrication (DF). Finally, we instantiate the patient profile using 347 public cases with rich clinical conditions involving the discovered canonical CSUs and establish a comprehensive evaluation suite that assesses diagnostic accuracy, stigmatized evidence recall, conversational efficiency, and longitudinal rapport-building.

We conduct extensive evaluations across 13 LLMs, including 4 domain- and 2 inquiry-enhanced models. The results reveal illusions of clinical competence in current medical AI. First, replacing the ideal full-disclosure patient with our stigma-aware patient, whose selective disclosure reflects

realistic communication barriers, shows that ideal-patient evaluations overestimate the robustness of current inquiry policies: communication becomes less efficient, trust-building weakens, and consequently diagnostic performance degrades (Figure 1). Second, evidence acquisition and diagnosis are not strictly causal: models may diagnose correctly without complete evidence (hallucination), yet still fail after acquiring the full minimum necessary information (ceiling). Third, integrating a dedicated agent scaffold (e.g., `MEDDxAgent` (Rose et al., 2025)) improves sensitive-evidence elicitation, but does not reliably improve the eventual diagnostic accuracy. Our contributions are:

- We formalize stigma-aware clinical inquiry as an imperfect-information, general-sum game and introduce Clinical Stigma Units (CSUs) as quantifiable units of sensitive clinical information.
- We propose S^2InQBench, an interactive diagnosis benchmark that instantiates a neural-symbolic patient with 347 clinical cases to evaluate AI doctors against our formulated task. Human-validity analyses underscore this agent in reproducing key human sensitivity and disclosure dynamics.
- Our evaluations reveal 1) the overestimation of clinical communication capacity by ideal patients, 2) the ungrounded and capped diagnostic performance of current LLMs, and 3) the limitation of agent scaffolds in coordinating sensitive-evidence elicitation and diagnostic accuracy. These findings, coupled with the evaluation suite and environment, provide a foundation for developing and training methods that jointly optimize clinical inquiry, trust building, and diagnostic performance.

2 Clinical Stigma Mining

To ground interactive diagnosis in clinical reality, we construct a taboo repository that induces defensive concealment. This section formalizes the Clinical Stigma Unit (CSU), outlines our extraction-filtering-grounding pipeline, and details the scaling and auditing of latent stigma sensitivities.

2.1 Clinical Stigma Unit

Grounded in the Health Stigma and Discrimination Framework (HSDF) (Stangl et al., 2019; Earnshaw & Chaudoir, 2009) and the Cognitive-Affective Model of Concealable Stigma (Pachankis, 2007; Chaudoir & Fisher, 2010), disease stigma is not an isolated physiological label; rather, it is a dynamic, multi-dimensional psychosocial construct. Patients actively conceal health information when an inquiry encroaches upon moral transgressions, identity devaluation, or structural vulnerabilities (Link & Phelan, 2001; Corrigan, 2004). To bridge the gap between abstract clinical psychology principles and computational benchmark design, we introduce the **Clinical Stigma Unit (CSU)**.

Definition 1 (Clinical Stigma Unit)*: A CSU is the minimal clinical semantic and contextual unit within a medical encounter that systematically elicits acute self-stigma, anticipated healthcare discrimination, bodily shame, or defensive concealment from the patient.*

We operationalize CSUs across four hierarchical linguistic granularities: (1) **Canonical Clinical Entities**, proprietary diagnostic terms, psychiatric disorders, or controlled substances carrying established societal prejudice; (2) **Sensitive Behavioral Events**, verb-object event structures or traumatic actions undertaken by the patient that breach behavioral norms; (3) **Psychosocial Contexts**, socio-environmental and historical realities that expose the patient to secondary legal, professional, or interpersonal penalties upon disclosure; and (4) **Context-Attributed Symptoms**, physiologically benign symptoms whose diagnostic significance depends on a concealed etiology or a taboo anatomical context. We provide comprehensive theoretical definitions, psychological mechanisms, and clinical exemplars for each category in Appendix B.

2.2 Extraction and Canonicalization Pipeline

Sourced from public clinical questions and real multi-turn consultations (see § 4), we introduce a workflow that constructs a comprehensive yet non-redundant CSU repository in three steps:

Multi-Granularity Extraction. Given a raw clinical narrative or dialogue history, an LLM is instructed to parse clinical disclosures guided by four psychometric stigma dimensions: *Moral Blame & Stereotype*, *Anticipated Healthcare Discrimination*, *Bodily & Anatomical Taboo*, and *Secondary Disclosure Risk*. For each candidate span, the model extracts a structured 5-tuple:

$\mathcal{E} = \langle \tau_{\text{raw}}, c_{\text{norm}}, \mathcal{D}_{\text{psych}}, \mathcal{G}_{\text{ling}}, \mathcal{R}_{\text{stigma}} \rangle$, where τ_{raw} denotes the raw text span, c_{norm} is the normalized candidate concept, $\mathcal{D}_{\text{psych}}$ represents the stigma dimension, $\mathcal{G}_{\text{ling}}$ is the linguistic granularity, and $\mathcal{R}_{\text{stigma}}$ provides the clinical rationale explaining the patient's concealment motivation.

Filtering and Clustering. To eliminate clinical noise and false positives, we first filter malformed outputs and implement a clinical whitelist dictionary comprising neutral medical terms (e.g., *essential hypertension, uncomplicated viral pharyngitis, closed radius fracture*). Concepts devoid of taboo modifiers or sensitive etiologies are eliminated. Then, for each retained candidate, we synthesize a rich descriptor $\mathbf{d} = [c_{\text{norm}} \,\|\, \mathcal{D}_{\text{psych}} \,\|\, \mathcal{R}_{\text{stigma}}]$ and project it into a dense latent representation $\mathbf{z} = \text{Encoder}(\mathbf{d}) \in \mathbb{R}^d$ using a domain-specific biomedical embedding model (BioLinkBERT-Large; Yasunaga et al., 2022). After that, we apply Agglomerative Hierarchical Clustering (AHC) using cosine distance with a tight linkage distance threshold (Ward, 1963). This unsupervised aggregation groups disparate colloquial expressions of identical clinical taboos (e.g., "methamphetamine injection", "ice bingeing", and "crystal meth use disorder") into unified semantic clusters without requiring manual clustering heuristics.

Medoid Selection and Ontology Grounding. For each cluster $\mathcal{C}_k$, we select the medoid concept c_k^* that maximizes the average intra-cluster semantic similarity and exhibits high clinical frequency: $c_k^* = \arg\max_{c_i \in \mathcal{C}_k} \sum_{c_j \in \mathcal{C}_k} \frac{\mathbf{z}_i \cdot \mathbf{z}_j}{\|\mathbf{z}_i\|_2 \|\mathbf{z}_j\|_2}$. Each medoid is subsequently mapped and grounded into standardized medical ontologies, including the UMLS Metathesaurus, ICD-10-CM, and SNOMED CT. Concepts falling in the extreme long tail of clinical occurrence are pruned, yielding a consolidated repository of $K = 185$ canonical CSUs spanning the full spectrum of outpatient care. Figure 2(a) demonstrates their composition across data sources.

2.3 Quantification of Latent Sensitivity via Bradley-Terry Modeling

Direct scalar scoring of clinical taboos inevitably incurs calibration drift, central tendency compression, and ranking intransitivity. We therefore formulate sensitivity estimation as a probabilistic pairwise comparison task under the Bradley-Terry model (Bradley & Terry, 1952), which derives interval-scale metric distances from discrete relative decisions.

Firstly, each CSU $i \in \{1, \ldots, K\}$ is parameterized by an unobserved latent stigma intensity $\mu_i \in \mathbb{R}$. The probability that CSU i induces greater defensive reticence than CSU j ($i \succ j$) follows the logistic distribution:

$$P(i \succ j) = \sigma(\mu_i - \mu_j) = \frac{1}{1 + \exp(-(\mu_i - \mu_j))}. \tag{1}$$

Then, to avoid uninformative exhaustive comparisons, we curate ∼2,400 high-information comparison pairs stratified into: (i) *intra-dimensional deep pairs* (40%) to refine local orderings; (ii) *cross-dimensional proximal pairs* (40%) to calibrate cross-category scales; and (iii) *global golden anchor pairs* (20%) across five benchmark conditions to stabilize absolute coordinates. Pairwise preferences $y_{ij} \in \{0, 1\}$ are generated by a rubric-instantiated rater ensemble and subsequently audited and adjudicated by clinical consensus; both the rubric and the automatic scoring pipeline are documented in Appendices B.3 and B.4. Given the comparison set $\mathcal{D} = \{(i_k, j_k, y_{i_k j_k})\}_{k=1}^M$, we solve for the latent parameters $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_K]^T$ by minimizing the L_2-regularized negative log-likelihood:

$$\mathcal{L}(\boldsymbol{\mu}) = -\sum_{(i,j) \in \mathcal{D}} \Big[y_{ij} \ln \sigma(\mu_i - \mu_j) + (1 - y_{ij}) \ln \big(1 - \sigma(\mu_i - \mu_j)\big) \Big] + \frac{\lambda}{2} \|\boldsymbol{\mu}\|_2^2. \tag{2}$$

As $\mathcal{L}(\boldsymbol{\mu})$ is strictly convex, it admits a unique global optimum $\boldsymbol{\mu}^*$, computed iteratively via the Newton-Raphson method (gradient and Hessian derivations are provided in Appendix C).

Finally, we apply a min-max linear rescaling onto the bounded interval $[\mathcal{S}_{\min}, \mathcal{S}_{\max}] = [0.05, 0.95]$:

$$\mathcal{S}_i = \mathcal{S}_{\min} + (\mathcal{S}_{\max} - \mathcal{S}_{\min}) \cdot \frac{\mu_i^* - \min(\boldsymbol{\mu}^*)}{\max(\boldsymbol{\mu}^*) - \min(\boldsymbol{\mu}^*)}. \tag{3}$$

Setting $\mathcal{S}_{\min} = 0.05$ preserves baseline bodily privacy for routine physiological queries, while $\mathcal{S}_{\max} = 0.95$ prevents absorbing probability barriers in downstream Beta trust dynamics, reflecting that any taboo can be disclosed once therapeutic rapport is fully established.

2.4 Psychometric Auditing and Clinical Validation

To ensure the empirical validity of our calibrated sensitivities, we conduct rigorous psychometric auditing and expert validation. As reported in Table 9, *Circular Triad* analysis yields a transitivity violation rate of under 2.3%, satisfying Weak Stochastic Transitivity (WST), while cross-annotator agreement demonstrates high reliability (Kendall's $W = 0.892$; ICC$(2, k) = 0.914$, $p < 0.001$). Furthermore, in a blind comparison of 20 stratified CSUs against gold-standard ratings from three board-certified clinical psychologists, our Bradley-Terry calibration achieves strong concordance (Pearson $r = 0.924$, Spearman $\rho = 0.931$), substantially outperforming direct zero-shot prompting with GPT-4o ($r = 0.741, \rho = 0.718$), which suffers from severe variance compression. Refer to Appendix D for detailed experimental setups, evaluation rubrics, and granular per-sample ratings.

3 Benchmarking Clinical Inquiry under Patient Stigma

In this section, we benchmark interactive clinical diagnosis as an imperfect-information, general-sum game (§ 3.1) to evaluate whether an intelligent medical system can elicit sensitive health conditions and formulate safe diagnoses when encountering stigma-sensitive patients (§ 3.2).

3.1 Task Formulation

Unlike conventional benchmarks that treat medical history taking as passive database retrieval, where a perfectly cooperative patient retrieves the doctor-required information and then honestly replies, we conceptualize the outpatient consultation between the evaluated Doctor Agent ($\mathcal{D}$) and Patient Agent ($\mathcal{P}$) as an **imperfect-information, general-sum interaction**. This formulation captures the fundamental tension in real-world clinical encounters: the physician seeks diagnostic certainty and therapeutic safety under conversational efficiency, whereas the patient balances the desire for symptom relief against the psychological defense of personal privacy and disease stigma. When interpersonal trust fails to overcome perceived stigma, the patient strategically withholds, deflects, or distorts clinical evidence, creating an interactive testbed in which conversational bedside manner directly determines diagnostic outcome.

To operationalize this interaction with clinical fidelity, we ground the patient profile $\mathcal{M}$ in the standardized SOAP (Subjective, Objective, Assessment, Plan) paradigm (Weed, 1968), enforcing strict information boundaries across three chronological phases:

- **Intake (Turn 0)**: The environment initializes $\mathcal{D}$ with an incomplete public context $\mathcal{C}_0$ containing basic demographics and a spontaneous chief complaint. The patient's *Subjective* history $\mathcal{U}_S$ (symptom trajectory, pertinent negatives, comorbidities, and psychosocial history) and *Objective* data $\mathcal{U}_O$ (physical findings, vitals, laboratory tests, and imaging) remain private to $\mathcal{P}$, while the ground-truth assessment and management plan $\langle A^*, P^* \rangle$ are blinded to both agents.
- **Dynamic Inquiry**: At each turn $1 \dots T_{\max}$, $\mathcal{D}$ formulates natural language inquiries $u_t^{\mathcal{D}}$ to elicit missing evidence. In response, $\mathcal{P}$ dynamically updates its internal trust state, determines a disclosure posture, and returns a conversational reply $r_t^{\mathcal{P}}$ that contains controlled Atomic Information Units (AIUs). The interaction is bounded by a turn budget $T_{\max} = |\text{MNI}| + 3$, where MNI is the minimum necessary information set for establishing A^*.
- **Diagnostic Closure**: The encounter terminates when $\mathcal{D}$ submits its structured assessment A and treatment plan P, or upon budget exhaustion. A closed-loop audit then evaluates both conversational process quality (e.g., MNI coverage, dialogue parsimony) and clinical outcome correctness (e.g., diagnostic accuracy, recall of critical stigmatized evidence, and contraindication avoidance).

3.2 Patient Agent: Neuro-Symbolic Cognitive Architecture

To prevent unconstrained hallucinations and ensure that dialogue behavior strictly adheres to clinical psychological principles, our patient agent is architected as a tool-augmented agentic system, coupling an off-the-shelf LLM backbone with a dedicated neuro-symbolic cognitive tool $\mathcal{T}_{\text{cog}}$:

$$\mathcal{P} = \langle \mathcal{M}, \mathcal{B}_t, \mathcal{T}_{\text{cog}}, \mathcal{G}_{\text{LLM}} \rangle, \tag{4}$$

where $\mathcal{M} = \langle \mathcal{U}_S, \mathcal{U}_O \rangle$ denotes the full clinical profile (encompassing all private subjective and objective facts), and $\mathcal{B}_t = \text{Beta}(\alpha_t, \beta_t)$ represents the internal trust belief state. The formulation of

this agentic patient fundamentally departs S^{2}InQBench from prior programmatic and prompt-only patient simulators (Schmidgall et al., 2024; Gong et al., 2025).

Upon receiving the doctor's inquiry $u_t^{\mathcal{D}}$, the patient agent does not directly synthesize a verbal response. Instead, it calls the tool $\mathcal{T}_{\text{cog}}$ to evaluate the inquiry against its internal trust state and determine a disclosure posture G_t, which subsequently governs surface-level text generation:

$$(G_t, \mathcal{B}_t) \leftarrow \mathcal{T}_{\text{cog}}(u_t^{\mathcal{D}}, \mathcal{B}_{t-1}), \Rightarrow r_t^{\mathcal{P}} \sim \mathcal{G}_{\text{LLM}}(\cdot \mid \mathcal{M}, H_{t-1}, u_t^{\mathcal{D}}, G_t), \tag{5}$$

where $H_{t-1} = \{(u_k^{\mathcal{D}}, r_k^{\mathcal{P}})\}_{k=1}^{t-1}$ denotes the dialogue history and $r_t^{\mathcal{P}}$ is the realized response. Internally, $\mathcal{T}_{\text{cog}}$ bridges semantic perception with Bayesian decision-making: it employs a natural language understanding (NLU) module f_{NLU} to extract clinical communicative acts and topic stigma sensitivity ($\mathbf{a}_t, \mathcal{S}_t \leftarrow f_{\text{NLU}}(u_t^{\mathcal{D}})$), updates the trust belief state ($\mathcal{B}_t \leftarrow \text{Update}(\mathcal{B}_{t-1}, \mathbf{a}_t, \mathcal{S}_t)$), and samples a symbolic disclosure posture $G_t \sim \pi_{\text{disc}}(\cdot \mid \mathcal{B}_t, \mathcal{S}_t)$. The communicative acts $\mathbf{a}_t$ parsed by f_{NLU} are grounded in the Roter Interaction Analysis System (RIAS) (Roter & Larson, 2002) and the Empathic Communication Coding System (ECCS) (Bylund & Makoul, 2002). These acts are partitioned into trust-building categories $\mathcal{A}^+$ (empathic validation a^+_{emp}, confidentiality assurance a^+_{conf}, normalization a^+_{norm}, and collaborative partnership a^+_{collab}) and trust-damaging categories $\mathcal{A}^-$ (premature intrusive probing a^-_{intru}, moral confrontation a^-_{judg}, and dismissive mechanical phrasing a^-_{dism}), each weighted by its established psychological impact (specifications listed in Appendix G).

We model patient interpersonal trust not as a deterministic scalar, but as a continuous posterior belief distribution $T_t \sim \text{Beta}(\alpha_t, \beta_t)$, initialized from an uninformative prior $\alpha_0 = \beta_0 = 1.0$ (i.e., $\text{Uniform}(0, 1)$ with prior expectation $\bar{T}_0 = 0.5$ and variance $\text{Var}(T_0) = 1/12$). Concurrently, the queried topic is mapped to our calibrated CSU repository (§ 2), assigning a sensitivity parameter $\mathcal{S}_t \in [0.05, 0.95]$, or reverting to a baseline bodily privacy threshold $\mathcal{S}_{\text{base}} = 0.05$ for routine physiological queries. Belief parameters update dynamically through asymmetric evidence increments, where positive evidence is driven by therapeutic rapport and negative evidence incorporates both communicative missteps and an unearned intimacy penalty:

$$\begin{aligned} \Delta\alpha_t &= \eta^+ \sum_{a^+ \in \mathcal{A}^+} w(a^+) p(a^+ \mid H_{t-1}, u_t^{\mathcal{D}}), \\ \Delta\beta_t &= \eta^- \left[\sum_{a^- \in \mathcal{A}^-} w(a^-) p(a^- \mid H_{t-1}, u_t^{\mathcal{D}}) + \gamma\, \Phi_{\text{trespass}}(\mathcal{S}_t, \bar{T}_{t-1}) \right]. \end{aligned} \tag{6}$$

To embody the psychological reality of *negativity bias*: trust is fragile to destroy yet arduous to build, we enforce an asymmetric update sensitivity $\eta^- = 3\eta^+$. The boundary trespassing penalty is formulated as:

$$\Phi_{\text{trespass}}(\mathcal{S}_t, \bar{T}_{t-1}) = \left[\text{ReLU}\left(\mathcal{S}_t - \bar{T}_{t-1}\right)\right]^2, \tag{7}$$

which activates quadratically whenever the topic's stigma $\mathcal{S}_t$ exceeds the prior expected trust $\bar{T}_{t-1} = \alpha_{t-1}/(\alpha_{t-1} + \beta_{t-1})$, ensuring that premature probing into sensitive etiologies triggers defensive resistance even if phrased politely. To model long-sequence emotional attenuation and prevent singular variance collapse, parameters smoothly regress toward the uninformative prior:

$$\begin{aligned} \alpha_t &= \lambda_{\text{decay}} \alpha_{t-1} + (1 - \lambda_{\text{decay}})\alpha_0 + \Delta\alpha_t, \\ \beta_t &= \lambda_{\text{decay}} \beta_{t-1} + (1 - \lambda_{\text{decay}})\beta_0 + \Delta\beta_t, \end{aligned} \tag{8}$$

where $\lambda_{\text{decay}} \in (0.95, 1.0]$ denotes the retention factor.

To govern communicative disclosure, rather than enforcing a naive, deterministic hard comparison between trust and sensitivity (e.g., disclosing if and only if $\bar{T}_t \geq \mathcal{S}_t$), we formulate disclosure as a stochastic decision governed by the posterior tail probability of overcoming stigma. This is because hard thresholding fundamentally presumes an idealized, hyper-rational decision-maker, contradicting the clinical reality that human disclosure is always characterized by bounded rationality, emotional ambivalence, and psychological stochasticity. We operationalize this mechanism by computing the tail probability that the patient's latent trust strictly surpasses the topic's stigma resistance:

$$P_{\text{trust}}(t) = P(T_t > \mathcal{S}_t) = 1 - I_{\mathcal{S}_t}(\alpha_t, \beta_t) = 1 - \frac{\text{B}(\mathcal{S}_t; \alpha_t, \beta_t)}{\text{B}(\alpha_t, \beta_t)}, \tag{9}$$

where $I_x(\alpha, \beta)$ denotes the regularized incomplete Beta function. This probabilistic treatment naturally captures the governing role of *epistemic uncertainty*: during early turns when observations

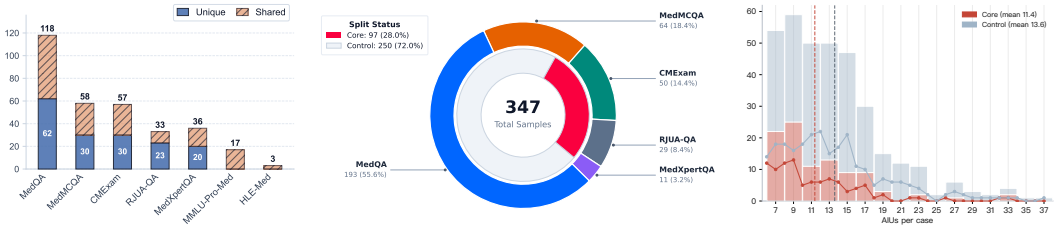


Figure 2: Composition of the benchmark corpus: (a) distribution of canonical CSUs across clinical sources, (b) source provenance of patient profiles, and (c) distribution of AIUs per case.

are sparse, the broad variance $\mathrm{Var}(T_t)$ disperses probability density, substantially suppressing the tail probability $P_{\text{trust}}(t)$ even if the point expectation exceeds sensitivity ($\bar{T}_t > \mathcal{S}_t$). This faithfully replicates the clinical hesitation of a patient who appreciates a clinician's preliminary warmth but remains unwilling to risk exposing guarded taboos. As communicative evidence accumulates over the encounter, posterior variance shrinks, crystallizing $P_{\text{trust}}(t)$ into decisive behavioral certainty.

We operationalize the disclosure policy π_{disc} as a categorical distribution parameterized by the tail probability $P_{\text{trust}}(t)$ and the macro cooperation propensity $\sigma_0(\bar{T}_t) = \mathrm{sigmoid}(5 \cdot (\bar{T}_t - 0.5))$:

$$\begin{aligned} &\pi_{\text{disc}}(G_t = \text{FD} \mid \mathcal{B}_t, \mathcal{S}_t) \propto P_{\text{trust}}(t) \cdot \sigma_0(\bar{T}_t), \\ &\pi_{\text{disc}}(G_t = i \mid \mathcal{B}_t, \mathcal{S}_t) \propto \big(1 - P_{\text{trust}}(t)\big) \cdot w_i \cdot \begin{cases} \sigma_0(\bar{T}_t), & i = \text{PO}, \\ 1 - \sigma_0(\bar{T}_t), & i \in \{\text{ER}, \text{DF}\}, \end{cases} \end{aligned} \tag{10}$$

where $i \in \{\text{PO}, \text{ER}, \text{DF}\}$ indexes the three defensive modes, and probabilities are normalized over all four postures. Full Disclosure (FD) transparently yields the clinical facts; Partial Omission (PO) shares surface complaints while deliberately suppressing taboo causal antecedents; Evasion and Redirection (ER) verbally deflects intrusive questions toward benign topics; and Distortion and Fabrication (DF) actively constructs plausible, false alibis. If the required facts are absent in the patient profile, the turn is skipped. Otherwise, the sampled discrete token G_t is injected as an immutable symbolic directive into the generator LLM's prompt, ensuring that surface-text realizations faithfully reflect the patient's latent defensive posture without unconstrained behavioral drift.

4 Data Collection and Processing

Data Collection. Both our CSU repository and patient profiles are drawn from public corpora. Medical question-answering datasets MedQA (Jin et al., 2021), CMExam (Liu et al., 2023), MedMCQA (Pal et al., 2022), MedXpertQA (Zuo et al., 2025), MMLU-Pro-Med (Wang et al., 2024), and HLE-Med (Phan et al., 2025) provide broad departmental coverage, multilingual questions, expert-level cases, and authoritative gold diagnoses. RJUA-QA (Lyu et al., 2023) collects real-world Chinese doctor-patient consultations that contributes authentic multi-turn discourse patterns.

Data Processing. We filter the collected data by task type using GPT-4o, retaining 19,236 clinically diagnosable samples for structured processing. These samples are first converted into SOAP-formatted profiles composed of AIUs. Then, they undergo multi-stage processing: stigma-coverage filtering, cross-model solvability checking, block-level stigma-criticality testing, near-duplicate-and-leakage cleaning, and stratified human audit. Finally, 347 interactive clinical cases are retained to instantiate S^2InQBench, including 97 *stigma-core* cases, where high-sensitivity AIUs are causally necessary to derive the gold diagnosis, and 250 *control* cases, where comparable sensitive facts are present but diagnostically redundant. Figure 2 summarizes the source distributions of the 347 retained patient profiles and their AIU counts in panels (b) and (c), respectively. Appendix F provides the complete processing protocol and stepped results.

5 Evaluation and Analyses

Evaluation protocol and metrics. We report three metric families, with full definitions in Appendix H.3: process efficiency (average consultation length, #Turn), diagnostic outcome (Acc@1,

Table 1: Evaluation results on the S^2InQBench core and control subsets. Aquila-Med results are marked with * since it suffers from severe catastrophic forgetting and degrades conversational ability.

	Core				Control			
Model	Acc@1 (%)	Cov_{MNI} (%)	Recall_C (%)	Trust-AUC	Acc@1 (%)	Cov_{MNI} (%)	Recall_C (%)	Trust-AUC
General LLMs								
GPT-5.5 (OpenAI, 2026)	**60.6**	**70.8**	60.7	0.557	54.0	66.9	38.2	**0.555**
Claude Opus 4.8 (Anthropic, 2026)	53.1	70.2	**61.2**	0.549	48.0	**82.7**	54.4	0.500
Gemini 3.5 Flash (Google DeepMind, 2026)	50.0	70.5	40.6	**0.561**	50.0	78.5	**59.0**	0.500
GLM-5.3-Flash (Z.ai, 2026)	53.6	67.3	54.9	0.547	**59.6**	70.7	44.6	0.548
DeepSeek-V4.1-Flash (DeepSeek-AI et al., 2026)	52.6	57.1	46.7	0.521	56.2	67.6	48.9	0.500
Qwen3.8-Flash-Next (Qwen Team, 2026)	47.4	60.7	46.3	0.507	53.2	63.4	39.4	0.542
Kimi-K3 (Kimi Team, 2026)	50.0	60.6	48.3	0.527	55.6	66.3	44.0	0.527
Medical LLMs								
Baichuan-M2 (Dou et al., 2025)	39.2	57.0	44.0	0.482	39.2	61.5	35.0	0.500
MedReason (Wu et al., 2025)	27.8	54.7	42.8	0.465	28.0	59.6	36.6	0.488
Aquila-Med (Zhao et al., 2024)	13.7*	17.5*	15.5*	0.479*	16.0*	18.7*	17.2*	0.493*
HuatuoGPT-o1-8B (Chen et al., 2024)	20.6	55.9	45.9	0.503	28.8	62.6	36.7	0.508
Inquiry-Enhanced Medical LLMs								
ProMed (Ding et al., 2026)	49.5	56.5	42.9	0.499	47.2	60.6	34.6	0.504
Dr.APP (Zhu et al., 2025)	47.2	55.5	45.2	0.500	38.2	52.1	29.9	0.497

Acc@3, minimum-necessary-information coverage Cov_{MNI}, and validated high-stigma evidence recall Recall_C), and longitudinal trust (Trust-AUC). We use `GLM-5.3-Flash` as the patient backbone throughout; Appendix H.4 shows that replacing it with `GPT-5.5` preserves model rankings ($\rho = 1.00$) and changes Acc@1 by at most 0.9 pp.

Cross-model comparison. Table 1 reveals a trade-off between medical knowledge, conversational competence, and diagnostic robustness. GPT-5.5 achieves the highest core Acc@1 (60.6%), whereas Claude Opus 4.8 and Gemini 3.5 Flash lead stigma-sensitive recall (61.2%) and Trust-AUC (0.561), respectively. Among open general-purpose models, GLM-5.3-Flash offers the strongest overall balance and the best control Acc@1 (59.6%); DeepSeek-V4.1-Flash is more economical but gives up coverage and accuracy. The efficiency frontier in Figure 4 shows why brevity is not a proxy for quality: Dr.APP finishes in about 2.9 turns but has weak coverage and accuracy, while GPT-5.5 reaches higher accuracy in roughly seven turns. GLM-5.3-Flash is the only model that remains competitive across both quality and turn cost. Domain-specific and inquiry-enhanced medical models improve on some medical baselines but do not match the strongest general LLMs, showing that specialization or additional inquiry structure alone does not ensure stigma-sensitive diagnosis.

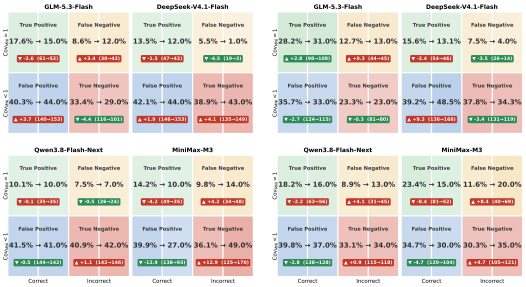


Figure 3: Diagnostic confusion matrices under stigma-aware (left) and ideal-patient (right) conditions; arrows show vanilla → +`MEDDxAgent`.

Finding 1: Stigma is a systematic source of diagnostic error. The ideal-patient ablation in Table 2 shows that the penalty is systematic rather than model-specific. On the stigma-core split, replacing selective disclosure with full disclosure improves Acc@1 for every model by 1.0–11.3 percentage points and increases MNI coverage by up to 15.2 pp. Effects on the control split are smaller and mixed, confirming that the degradation is concentrated where sensitive evidence is diagnostically load-bearing. Selective disclosure therefore harms both access to necessary evidence and its downstream synthesis, so ideal-patient scores overstate clinical robustness.

Finding 2: Clinical diagnosis exhibits both ungrounded inference and a synthesis ceiling. A grounded diagnosis should be supported by the evidence elicited during consultation, but Table 1 and Figure 3 reveal two violations. Models sometimes predict the correct label with $\text{Cov}_{\text{MNI}} < 1$, revealing an evidence-grounding failure that manifests as diagnostic hallucination or shortcut reliance; conversely, covered-but-incorrect cases persist even when $\text{Cov}_{\text{MNI}} = 1$. For example, Claude Opus 4.8 and Gemini 3.5 Flash reach 82.7% and 78.5% coverage on control cases but only 48.0% and 50.0% Acc@1, while the stigma track contains 8.6% and 9.8% covered-but-incorrect cases for GLM-5.3-Flash and Kimi-K3. Evidence access is therefore not sufficient for reliable clinical synthesis, and a correct label obtained without sufficient evidence is not an evidence-grounded diagnosis.

Table 2: Disclosure ablation. Positive values favor the ideal patient against the stigma-aware patient.

Model	Core				Control			
	ΔAcc@1 (%) ↑	ΔAcc@3 (%) ↑	$\Delta\text{Cov}_{\text{MNI}}$ (%) ↑	Δ#Turn ↓	ΔAcc@1 (%) ↑	ΔAcc@3 (%) ↑	$\Delta\text{Cov}_{\text{MNI}}$ (%) ↑	Δ#Turn ↓
General LLMs								
GPT-5.5 (OpenAI, 2026)	+5.4	−2.0	+12.6	+0.04	+2.0	−4.0	+10.9	+0.40
Claude Opus 4.8 (Anthropic, 2026)	+6.1	−4.1	+10.0	+1.37	+2.0	−2.0	−14.3	−0.34
Gemini 3.5 Flash (Google DeepMind, 2026)	+6.0	+8.0	+12.7	−0.12	+10.0	+10.0	−5.8	−0.32
GLM-5.3-Flash (Z.ai, 2026)	+7.2	−2.1	+11.7	+0.21	+5.6	+4.8	+7.1	+0.38
DeepSeek-V4.1-Flash (DeepSeek-AI et al., 2026)	+2.1	+2.1	+11.6	+0.12	+0.6	+2.4	−5.2	−0.09
Qwen3.8-Flash-Next (Qwen Team, 2026)	+8.2	+1.0	+9.2	+2.14	+5.6	+4.0	+6.1	+0.92
Kimi-K3 (Kimi Team, 2026)	+5.2	+3.1	+14.5	+0.42	+3.6	+2.4	+9.0	+0.99
Medical LLMs								
Baichuan-M2 (Dou et al., 2025)	+11.3	+13.4	+12.4	+0.42	+7.6	+1.6	+9.0	+0.25
MedReason (Wu et al., 2025)	+5.2	+5.2	+8.0	−0.39	+4.8	+1.6	−0.1	+0.43
Aquila-Med (Zhao et al., 2024)	+4.1	+1.2	+1.9	+0.10	+0.0	+0.8	+0.8	+0.10
HuatuoGPT-o1-8B (Chen et al., 2024)	+11.3	+11.3	+15.2	+0.29	+2.4	+7.2	+9.7	+0.58
Inquiry-Enhanced Medical LLMs								
ProMed (Ding et al., 2026)	+1.0	−3.1	+11.5	+0.23	+5.2	+8.4	+10.6	+0.27
Dr.APP (Zhu et al., 2025)	+3.4	−16.1	−5.6	−0.15	+8.6	+6.7	+4.9	−0.02

Table 3: Performance of open-source models harnessed by `MEDDxAgent` on the *full* benchmark. **Ideal Patient** indicates which patient simulator is used at evaluation: ✗ = our stigma-aware patient with selective disclosure; ✓ = an ideal patient who passively and honestly discloses.

Model	Ideal Patient	Acc@1 (%)			Cov_{MNI} (%)			Recall_C (%)			Trust-AUC		
		Vanilla	+MEDDxAgent	Δ	Vanilla	+MEDDxAgent	Δ	Vanilla	+MEDDxAgent	Δ	Vanilla	+MEDDxAgent	Δ
GLM-5.3-Flash (Z.ai, 2026)	✗	57.9	59.0	+1.1	69.7	72.5	+2.8	47.5	55.4	+8.0	0.548	0.795	+0.247
	✓	64.0	64.0	+0.0	78.1	79.7	+1.6	60.5	66.7	+6.2	0.500	0.500	0.000
DeepSeek-V4.1-Flash (DeepSeek-AI et al., 2026)	✗	55.6	56.0	+0.4	60.9	58.7	−2.2	40.0	43.5	+3.5	0.521	0.619	+0.098
	✓	54.8	61.6	+6.9	67.9	67.1	−0.8	51.9	57.2	+5.3	0.500	0.500	0.000
Qwen3.8-Flash-Next (Qwen Team, 2026)	✗	51.6	51.0	−0.6	62.6	62.4	−0.3	41.3	49.0	+7.7	0.532	0.736	+0.204
	✓	57.9	53.4	−4.5	69.6	70.7	+1.1	52.9	62.4	+9.5	0.500	0.500	0.000
Kimi-K3 (Kimi Team, 2026)	✗	54.0	37.0	−17.0	64.7	64.6	−0.1	45.2	49.2	+4.0	0.527	0.680	+0.153
	✓	58.1	45.0	−13.1	75.2	77.1	+1.9	58.5	65.1	+6.6	0.500	0.500	0.000

Finding 3: Agent scaffolding improves elicitation, but the gain does not causally transfer to diagnosis. Compared with vanilla prompting, `MEDDxAgent` increases Recall_C for all four open models in Table 3 (+3.5–8.0 pp) and raises GLM-5.3-Flash's MNI coverage by 2.8 pp. The elicitation gain does not reliably carry through to diagnosis: Acc@1 changes by only +1.1 pp (GLM) and +0.4 pp (DeepSeek), falls by 0.6 pp (Qwen), and drops by 17.0 pp (Kimi). The confusion matrices reinforce this dependence: DeepSeek reduces covered-but-incorrect cases from 5.5% to 1.0%, whereas GLM and Kimi increase them. Scaffolding thus changes information access, but its diagnostic effect depends on compatibility with the underlying model rather than following automatically from better elicitation.

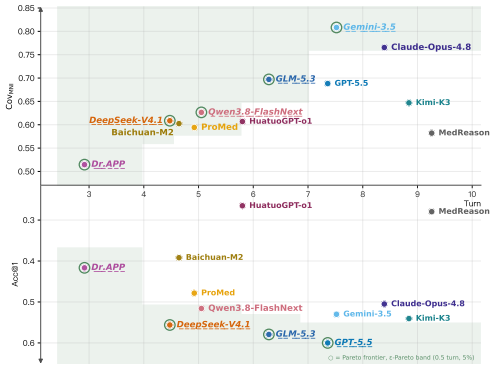


Figure 4: Consultation efficiency comparison.

Finding 4: The neural-symbolic patient reproduces human clinical behavior. The patient-validity study in Appendix I.1 shows close agreement with human sensitivity dynamics: aggregate TVD $= 0.037$ (maximum $= 0.077$), within the preregistered 0.10 margin; similar full-disclosure rates (0.312 for humans versus 0.275 for the agent); and a 1.3-pp doctor-side MNI-coverage difference that passes equivalence testing. These results validate the agent as a realistic simulator of trust-gated disclosure and defensive communication. The 1.7-pp Acc@1 difference is reported only as a supplementary check because the $n = 50$ study is underpowered for binary equivalence testing.

6 Conclusion

We introduced S^{2}InQBench, a stigma-aware interactive diagnosis benchmark built upon calibrated CSUs and enabled by a neural-symbolic patient that models trust-gated conditions disclosure. Evaluations across 13 LLMs demonstrate the degraded performance of current medical AI in evidence acquisition and grounded diagnosis without idealized patient cooperation. Our findings, together with the evaluation suite and environment, motivate future clinical AI towards joint optimization of clinical communication and diagnostic reasoning without assuming fully cooperative patients.

AI USE STATEMENT

In this work, we used generative AI tools for code implementation, data processing, and figure generation. We have not used generative AI tools for core research idea generation or key conceptual drafting. Additionally, we used generative AI tools for LaTeX source formatting, spell checking, and text polishing. We have reviewed all AI-assisted work. Specifically, all generated code and data processing scripts were manually inspected and thoroughly tested for correctness by the authors; created figures were verified against raw experimental data; and all polished text was carefully checked to ensure semantic fidelity and factual accuracy. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Anthropic. Introducing Claude Opus 4.8, 2026. URL https://www.anthropic.com/news/claude-opus-4-8.

Barbara E. Berger, Carol Estwing Ferrans, and Felissa R. Lashley. Measuring stigma in people with HIV: Psychometric assessment of the HIV stigma scale. *Research in Nursing & Health*, 24(6): 518–529, 2001. doi: 10.1002/nur.10011.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 0006-3444. doi: 10.1093/biomet/39.3-4.324.

Peter G. Brodeur, Jacob M. Koshy, Anil Palepu, et al. A prospective clinical feasibility study of a conversational diagnostic AI in an ambulatory primary care clinic. *CoRR*, abs/2603.08448, 2026. doi: 10.48550/ARXIV.2603.08448. URL https://doi.org/10.48550/arXiv.2603.08448.

Carma L. Bylund and Gregory Makoul. Empathic communication and gender in the physician-patient encounter. *Patient Education and Counseling*, 48(3):207–216, 2002. doi: 10.1016/S0738-3991(02)00173-8.

Stephenie R. Chaudoir and Jeffrey D. Fisher. The disclosure processes model: Understanding disclosure decision making and postdisclosure outcomes among people living with a concealable stigmatized identity. *Psychological Bulletin*, 136(2):236–256, 2010. doi: 10.1037/a0018193.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. HuatuoGPT-o1, towards medical complex reasoning with LLMs. *CoRR*, abs/2412.18925, 2024. doi: 10.48550/ARXIV.2412.18925. URL https://doi.org/10.48550/arXiv.2412.18925.

Patrick W. Corrigan. How stigma interferes with mental health care. *American Psychologist*, 59(7): 614–625, 2004. doi: 10.1037/0003-066X.59.7.614.

DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chengyu Hou, Chenhao Xu, Chenze Shao, Chong Ruan, Conner Sun, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Donghao Li, Dongjie Ji, Erhang Li, Fang Wei, Fangyun Lin, Fangzhou Yuan, Feiyu Xia, Fucong Dai, Guangbo Hao, Guanting Chen, Guoai Cao, Guolai Meng, Guowei Li, Han Yu, Han Zhang, Hanwei Xu, Hao Li, Haofen Liang, Haoling Zhang, Haoming Luo, Haoran Wei, Haotian Yuan, Haowei Zhang, Haowen Luo, Haoyu Chen, Haozhe Ji, Hengqing Zhang, Honghui Ding, Hongxuan Tang, Huanqi Cao, Huazuo Gao, Hui Qu, Hui Zeng, J Yang, JQ Zhu, Jia Luo, Jia Song, Jia Yu, Jialiang Huang, Jialu Cai, Jian Liang, Jiangting Zhou, Jiasheng Ye, Jiashi Li, Jiaxin Xu, Jiewen Hu, Jieyu Yang, Jin Chen, Jin Yan, Jingchang Chen, Jingli Zhou, Jingting Xiang, Jingyang Yuan, Jingyuan Cheng, Jingzi Zhou, Jinhua Zhu, Jiping Yu, Joseph Sun, Jun Ran, Junguang Jiang, Junjie Qiu, Junlong Li, Junmin Zheng, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Kexing Zhou, Kezhao Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Wang, Leyi Xia, Li Zhang, Liang Zhao, Lihua Guo, et al. DeepSeek-V4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

Hongxin Ding, Baixiang Huang, Yue Fang, Weibin Liao, Xinke Jiang, Jinyang Zhang, Yinghao Zhu, Zheng Li, Liantao Ma, Junfeng Zhao, and Yasha Wang. Promed: Shapley information gain guided reinforcement learning for proactive medical llms. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 32481–32515. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.1500. URL https://doi.org/10.18653/v1/2026.acl-long.1500.

Chengfeng Dou, Chong Liu, Fan Yang, et al. Baichuan-M2: Scaling medical capability with large verifier system. *CoRR*, abs/2509.02208, 2025. doi: 10.48550/ARXIV.2509.02208. URL https://doi.org/10.48550/arXiv.2509.02208.

Chengfeng Dou, Fan Yang, Fei Li, Jiyuan Jia, Qiang Ju, Shuai Wang, Tianpeng Li, Xiangrong Zeng, Yijie Zhou, Hongda Zhang, Jinyang Tai, Linzhuang Sun, Peidong Guo, Yichuan Mo, Xiaochuan Wang, Hengfu Cui, and Zhishou Zhang. Baichuan-M3: Modeling clinical inquiry for reliable medical decision-making. *CoRR*, abs/2602.06570, 2026. doi: 10.48550/ARXIV.2602.06570. URL https://doi.org/10.48550/arXiv.2602.06570.

Valerie A. Earnshaw and Stephenie R. Chaudoir. From conceptualizing to measuring HIV stigma: A review of HIV stigma mechanism measures. *AIDS and Behavior*, 13(6):1160–1177, 2009. doi: 10.1007/s10461-009-9593-3.

Valerie A. Earnshaw and Diane M. Quinn. The impact of stigma in healthcare on people living with chronic illnesses. *Journal of Health Psychology*, 17(2):157–168, 2012. doi: 10.1177/1359105311414952.

Erving Goffman. *Stigma: Notes on the Management of Spoiled Identity*. Prentice-Hall, Englewood Cliffs, NJ, 1963.

Linlu Gong, Ante Wang, Yunghwei Lai, Weizhi Ma, and Yang Liu. The dialogue that heals: A comprehensive evaluation of doctor agents' inquiry capability. *CoRR*, abs/2509.24958, 2025. doi: 10.48550/ARXIV.2509.24958. URL https://doi.org/10.48550/arXiv.2509.24958.

Google DeepMind. Gemini 3.5 Flash model card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-5-flash/.

Yann Hicke, Jadon Geathers, Kellen Vu, Justin Sewell, Claire Cardie, Jaideep Talwalkar, Dennis L. Shung, Anyanate Gwendolyne Jack, Susannah Cornes, Mackenzi Preston, and René F. Kizilcec. Medsimai: Simulation and formative feedback generation to enhance deliberate practice in medical education. In *Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference, LAK 2026, Bergen, Norway, 27 April 2026- 1 May 2026*, pp. 415–424. ACM, 2026. doi: 10.1145/3785022.3785092. URL https://doi.org/10.1145/3785022.3785092.

Zhihao Jia, Mingyi Jia, Junwen Duan, and Jianxin Wang. DDO: dual-decision optimization for llm-based medical consultation via multi-agent collaboration. *CoRR*, abs/2505.18630, 2025. doi: 10.48550/ARXIV.2505.18630. URL https://doi.org/10.48550/arXiv.2505.18630.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. doi: 10.3390/app11146421. URL https://doi.org/10.3390/app11146421.

Kimi Team. Kimi K3: Open frontier intelligence, 2026. URL https://arxiv.org/abs/2607.24653.

Daeun Kyung, Hyunseung Chung, Seongsu Bae, Jiho Kim, Jae Ho Sohn, Taerim Kim, Soo Kyung Kim, and Edward Choi. Patientsim: A persona-driven simulator for realistic doctor-patient interactions. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz,

and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/24945e3bdc7b3f4b2e64b9979a16f38e-Abstract-Datasets_and_Benchmarks_Track.html.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. Mediq: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning. *CoRR*, abs/2406.00922, 2024. doi: 10.48550/ARXIV.2406.00922. URL https://doi.org/10.48550/arXiv.2406.00922.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, and You Zhang. Chatdoctor: A medical chat model fine-tuned on LLaMA model using medical domain knowledge. *CoRR*, abs/2303.14070, 2023. doi: 10.48550/ARXIV.2303.14070. URL https://doi.org/10.48550/arXiv.2303.14070.

Yusheng Liao, Yutong Meng, Yuhao Wang, Hongcheng Liu, Yanfeng Wang, and Yu Wang. Automatic interactive evaluation for large language models with state aware patient simulator. *CoRR*, abs/2403.08495, 2024. doi: 10.48550/ARXIV.2403.08495. URL https://doi.org/10.48550/arXiv.2403.08495.

Bruce G. Link and Jo C. Phelan. Conceptualizing stigma. *Annual Review of Sociology*, 27(1): 363–385, 2001. doi: 10.1146/annurev.soc.27.1.363.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, and Michael Lingzhi Li. Benchmarking large language models on cmexam - A comprehensive chinese medical exam dataset. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a48ad12d588c597f4725a8b84af647b5-Abstract-Datasets_and_Benchmarks.html.

Lei Liu, Xiaoyan Yang, Fangzhou Li, Chenfei Chi, Yue Shen, Shiwei Lyu, Ming Zhang, Xiaowei Ma, Xiangguo Lv, Liya Ma, Zhiqiang Zhang, Wei Xue, Yiran Huang, and Jinjie Gu. Towards automatic evaluation for llms' clinical capabilities: Metric, data, and algorithm. In Ricardo Baeza-Yates and Francesco Bonchi (eds.), *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pp. 5466–5475. ACM, 2024. doi: 10.1145/3637528.3671575. URL https://doi.org/10.1145/3637528.3671575.

Shiwei Lyu, Chenfei Chi, Hongbo Cai, Lei Shi, Xiaoyan Yang, Lei Liu, Xiang Chen, Deng Zhao, Zhiqiang Zhang, Xianguo Lyu, Ming Zhang, Fangzhou Li, Xiaowei Ma, Yue Shen, Jinjie Gu, Wei Xue, and Yiran Huang. RJUA-QA: A comprehensive QA dataset for urology, 2023. URL https://arxiv.org/abs/2312.09785.

OpenAI. Introducing GPT-5.5, 2026. URL https://openai.com/index/introducing-gpt-5-5/.

John E. Pachankis. The psychological implications of concealing a stigma: A cognitive-affective-behavioral model. *Psychological Bulletin*, 133(2):328–345, 2007. doi: 10.1037/0033-2909.133.2.328.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering, 2022. URL https://arxiv.org/abs/2203.14371.

Long Phan, Alice Gatti, Ziwen Han, et al. Humanity's last exam, 2025. URL https://arxiv.org/abs/2501.14249.

Chuhan Qiao, Jianghua Huang, Daxing Zhao, Ziding Liu, Yanjun Shen, Bing Cheng, Wei Lin, and Kai Wu. Medconsultbench: A full-cycle, fine-grained, process-aware benchmark for medical consultation agents. *CoRR*, abs/2601.12661, 2026. doi: 10.48550/ARXIV.2601.12661. URL https://doi.org/10.48550/arXiv.2601.12661.

Qwen Team. Qwen3.8-Flash-Next, 2026. URL https://huggingface.co/Qwen/Qwen3.8-Flash-Next.

Jennifer Boyd Ritsher, Poorni G. Otilingam, and Monica Grajales. Internalized stigma of mental illness: Psychometric properties of a new measure. *Psychiatry Research*, 121(1):31–49, 2003. doi: 10.1016/j.psychres.2003.08.008.

Todd H. Rockwood, James M. Church, James W. Fleshman, Robert L. Kane, Constantinos Mavrantonis, Alan G. Thorson, Steven D. Wexner, Donna Bliss, and Ann C. Lowry. Fecal incontinence quality of life scale: Quality of life instrument for patients with fecal incontinence. *Diseases of the Colon & Rectum*, 43(1):9–16, 2000. doi: 10.1007/BF02237236.

Daniel Philip Rose, Chia-Chien Hung, Marco Lepri, Israa Alqassem, Kiril Gashteovski, and Carolin Lawrence. Meddxagent: A unified modular agent framework for explainable automatic differential diagnosis. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 13803–13826. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.677. URL https://doi.org/10.18653/v1/2025.acl-long.677.

Debra Roter and Susan Larson. The Roter interaction analysis system (RIAS): Utility and flexibility for analysis of medical interactions. *Patient Education and Counseling*, 46(4):243–251, 2002. doi: 10.1016/S0738-3991(02)00012-5.

Khaled Saab, Jan Freyberg, Chunjong Park, et al. Advancing conversational diagnostic AI with multimodal reasoning. *CoRR*, abs/2505.04653, 2025. doi: 10.48550/ARXIV.2505.04653. URL https://doi.org/10.48550/arXiv.2505.04653.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Pontes Reis, Jeffrey K. Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. *CoRR*, abs/2405.07960, 2024. doi: 10.48550/ARXIV.2405.07960. URL https://doi.org/10.48550/arXiv.2405.07960.

Laramie R. Smith, Valerie A. Earnshaw, Michael M. Copenhaver, and Chinazo O. Cunningham. Substance use stigma: Reliability and validity of a theory-based scale for substance-using populations. *Drug and Alcohol Dependence*, 162:34–43, 2016. doi: 10.1016/j.drugalcdep.2016.02.019.

Anne L. Stangl, Valerie A. Earnshaw, Carmen H. Logie, Wim van Brakel, Leickness C. Simbayi, Iman Barré, and John F. Dovidio. The health stigma and discrimination framework: a global, crosscutting framework to inform research, intervention development, and policy on health-related stigmas. *BMC Medicine*, 17(1):31, 2019. doi: 10.1186/s12916-019-1271-3.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 599–621, 2024. doi: 10.18653/v1/2024.findings-acl.33. URL https://doi.org/10.18653/v1/2024.findings-acl.33.

Tao Tu, Anil Palepu, Mike Schaekermann, et al. Towards conversational diagnostic AI. *CoRR*, abs/2401.05654, 2024. doi: 10.48550/ARXIV.2401.05654. URL https://doi.org/10.48550/arXiv.2401.05654.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark, 2024. URL https://arxiv.org/abs/2406.01574.

Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. MedAgent-Pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *CoRR*, abs/2503.18968, 2025. doi: 10.48550/ARXIV.2503.18968. URL https://doi.org/10.48550/arXiv.2503.18968.

Joe H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963. doi: 10.1080/01621459.1963.10500845.

Lawrence L. Weed. Medical records that guide and teach. *New England Journal of Medicine*, 278 (11):593–600, 1968. doi: 10.1056/NEJM196803142781105.

Hui Min Wong, Philip Heesen, Pascal Janetzky, Martin Bendszus, and Stefan Feuerriegel. Medclarify: An information-seeking AI agent for medical diagnosis with case-specific follow-up questions. *CoRR*, abs/2602.17308, 2026. doi: 10.48550/ARXIV.2602.17308. URL https://doi.org/10.48550/arXiv.2602.17308.

Juncheng Wu, Wenlong Deng, Yuyin Zhou, et al. MedReason: Eliciting factual medical reasoning steps in llms via knowledge graphs, 2025. URL https://arxiv.org/abs/2504.00993.

Zonghai Yao, Zihao Zhang, Chaolong Tang, Xingyu Bian, Youxia Zhao, Zhichao Yang, Junda Wang, Huixue Zhou, Won Seok Jang, Feiyun Ouyang, and Hong Yu. Medqa-cs: Objective structured clinical examination (OSCE)-style benchmark for evaluating LLM clinical skills. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), EACL 2026, Rabat, Morocco, March 2026*, pp. 6183–6257. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.EACL-LONG.292. URL https://doi.org/10.18653/v1/2026.eacl-long.292.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. Linkbert: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8003–8016, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.551. URL https://aclanthology.org/2022.acl-long.551/.

Huizi Yu, Jiayan Zhou, Lingyao Li, et al. Simulated patient systems powered by large language model-based AI agents offer potential for transforming medical education. *CoRR*, abs/2409.18924, 2024. doi: 10.48550/ARXIV.2409.18924. URL https://doi.org/10.48550/arXiv.2409.18924.

Z.ai. GLM-5.3-Flash: A natively multimodal frontier model with sparse and linear attention, 2026. URL https://docs.z.ai/guides/vlm/glm-5.3-flash.

Lulu Zhao, Weihao Zeng, Xiaofeng Shi, Hua Zhou, Donglin Hao, and Yonghua Lin. Aqulia-med LLM: pioneering full-process open-source medical language models. *CoRR*, abs/2406.12182, 2024. doi: 10.48550/ARXIV.2406.12182. URL https://doi.org/10.48550/arXiv.2406.12182.

Yang Zhou, Zhenting Sheng, Mingrui Tan, Yuting Song, Jun Zhou, Yu Heng Kwan, Lian Leng Low, Yang Bai, and Yong Liu. Note2chat: Improving llms for multi-turn clinical history taking using medical notes. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 35149–35157. AAAI Press, 2026. doi: 10.1609/AAAI.V40I41.40821. URL https://doi.org/10.1609/aaai.v40i41.40821.

Jiayuan Zhu, Jiazhen Pan, Yuyuan Liu, Fenglin Liu, and Junde Wu. Ask patients with patience: Enabling LLMs for human-centric medical dialogue with grounded reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025*, pp. 2846–2857, 2025. doi: 10.18653/v1/2025.emnlp-main.142. URL https://doi.org/10.18653/v1/2025.emnlp-main.142.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. MedXpertQA: Benchmarking expert-level medical reasoning and understanding, 2025. URL https://arxiv.org/abs/2501.18362.

A Related Work

Prior work relevant to this paper is usually partitioned by community—benchmarks, doctor-side systems, patient simulators—and each partition is then taxonomised along its own axis. That convention hides the one dimension along which this literature actually needs to be read. We instead sort all of it by a single question: *what does the work assume about the patient's willingness to disclose?* Four answers recur, and we refer to them as rungs of a ladder:

1. *No patient.* All decision-relevant evidence is handed to the model as a static vignette.
2. *Scripted respondent.* A patient whose replies are fixed before the encounter begins.
3. *Truthful oracle.* A state-bearing patient that answers honestly once the right question is asked.
4. *Trust-gated discloser.* A patient whose willingness to answer depends on the stigma of what is asked and on the doctor's accumulated conduct.

Rungs 1–3 are densely populated and rung 4, to our knowledge, is empty. The three subsections below walk the three literatures—benchmarks, doctor-side systems, and patient simulators—through this same ladder, and Table 4 summarises the outcome. The taxonomy internal to each subsection is deliberately secondary to the ladder, so that the subsections complement rather than compete.

A.1 Benchmarks for Medical Intelligence

Early *knowledge-aware* benchmarks sit on rung 1. MedQA (Jin et al., 2021) and CMExam (Liu et al., 2023) repurpose professional licensing examinations as multiple-choice tests; they measure biomedical knowledge retention over a fully observed vignette efficiently, but collapse consultation into static question answering, because every decision-relevant fact is supplied upfront and the model need not earn access to any of it.

Skill-aware benchmarks move to rung 2. MedQA-CS (Yao et al., 2026) adapts USMLE Step 2 CS-style encounters to score information gathering, physical examination, closure, communication and differential diagnosis, while RJUA-SPs (Liu et al., 2024) uses standardized-patient records and clinical pathways to score diagnostic completeness and consultation logic. These OSCE-style settings establish that examination accuracy does not imply clinical skill. Their patients, however, remain scripted cases with predetermined responses, so competence is demonstrated *over* a case rather than negotiated *with* an interlocutor.

Interaction-aware benchmarks close that distance and reach rung 3. MediQ (Li et al., 2024) begins from demographics and a chief complaint alone; AIE/SAPS (Liao et al., 2024) maintains an evolving patient state across turns; AgentClinic (Schmidgall et al., 2024) coordinates patient, doctor, measurement and moderator agents; and MAQUE and MedConsultBench (Gong et al., 2025; Qiao et al., 2026) attach AIU-level process, efficiency, patient-experience, test-use and full-cycle consultation metrics to the encounter. What these works newly introduce is hidden state and multi-turn credit assignment: the benchmark scores how evidence is *elicited*, not only how the final diagnosis is produced. What they leave untouched is the disclosure contract.

That contract is where the ladder stops. The variation these systems introduce is variation in *expression*, not in *willingness*: personas, demographic bias, emotion, cognitive noise, imperfect recall and passive responsiveness alter wording and latency, but the clinical record stays complete and internally consistent, and a sensitive fact is released truthfully once the appropriate question is asked. Willingness is not a free parameter anywhere in this literature—which is the more striking because measuring it is a mature psychometric enterprise outside it. Concealment, anticipated discrimination and internalized stigma are constructs with validated instruments built precisely for the populations our benchmark must sample, from the sociological account of spoiled identity (Goffman, 1963) to the Internalized Stigma of Mental Illness scale (Ritsher et al., 2003), the HIV Stigma Scale (Berger et al., 2001), substance-use stigma scales (Smith et al., 2016), and condition-specific quality-of-life instruments for stigmatised symptoms (Rockwood et al., 2000); the consequences of stigma inside care settings are likewise well documented (Earnshaw & Quinn, 2012). None of this instrumentation has been imported into benchmark design.

To our knowledge no benchmark occupies rung 4: none conditions disclosure on the stigma of a fact, on the patient's evolving trust, or on the interpersonal risk signalled by prior doctor behaviour,

and none tests deliberate omission, evasion, distortion or fabrication. Existing benchmarks therefore conflate diagnostic knowledge with the relational competence required to make disclosure possible in the first place, and can overestimate reliability with hesitant real patients.

Table 4: Where prior work sits on the patient-disclosure ladder. Rung 1 = static vignette; 2 = scripted respondent; 3 = truthful oracle; 4 = trust-gated discloser. ✓ = present, × = absent. † marks rung-3 works that vary persona, emotion, cognitive noise or recall: this changes *how* a fact is expressed, never *whether* it is volunteered. To our knowledge no prior work reaches rung 4.

Work	Patient model (rung)	Multi-turn	Hidden state	Elicited by asking	May withhold	May distort	Trust as state
MedQA / CMExam	vignette (1)	×	×	×	×	×	×
MedQA-CS / RJUA-SPs	scripted respondent (2)	✓	×	✓	×	×	×
MediQ	truthful oracle (3)	✓	✓	✓	×	×	×
AIE/SAPS†	truthful oracle (3)	✓	✓	✓	×	×	×
AgentClinic†	truthful oracle (3)	✓	✓	✓	×	×	×
MAQUE / MedConsultBench†	truthful oracle (3)	✓	✓	✓	×	×	×
AIPatient / PatientSim / MedSimAI†	truthful oracle (3)	✓	✓	✓	×	×	×
AMIE / ProMed / MedClarify	truthful oracle (3)	✓	✓	✓	×	×	×
S^2InQBench (ours)	**trust-gated discloser (4)**	✓	✓	✓	✓	✓	✓

A.2 LLM-BASED SYSTEMS FOR CLINICAL INQUIRY

Doctor-side systems are conventionally organised by *where diagnostic intelligence is placed*: in adapted weights, in externalized workflows, or in a learned inquiry policy. That axis is orthogonal to the ladder—it describes how the doctor is built, not what the doctor assumes about the patient—and read against the ladder, all three placements bottom out at rung 3. This subsection therefore keeps the three-way placement as its internal structure while using the ladder to keep score.

The first group adapts the physician-facing model itself. ChatDoctor (Li et al., 2023) fine-tunes a general LLM on medical knowledge and patient conversations; Aquila-Med (Zhao et al., 2024) combines continued pretraining, supervised dialogue fine-tuning and preference optimization; HuatuoGPT-o1 (Chen et al., 2024) strengthens complex medical inference; and Baichuan-M2 (Dou et al., 2025) scales medical capability with a large verifier system. These improve fluency, factual calibration, reasoning and answer verification, but they inherit the reactive QA contract: the model answers once evidence has been supplied rather than deciding what to seek next. Moving intelligence into the weights does not move the patient model.

The second group externalizes expertise into training-free, tool-using and multi-agent workflows. MedAgents (Tang et al., 2024) elicits multidisciplinary zero-shot collaboration; MedAgent-Pro (Wang et al., 2025) grounds multimodal diagnosis in retrieved evidence; Dr.APP (Zhu et al., 2025) couples grounded reasoning with patient-centric dialogue; and MEDDxAgent (Rose et al., 2025) decomposes history taking, retrieval and diagnostic planning into a modular ReAct loop—the decomposition we adopt as the training-free scaffold baseline in Appendix H.1. The common mechanism is role and tool decomposition, which improves behaviour without updating weights, but the added structure governs how the doctor reasons over evidence rather than whether the patient is willing to provide it.

The third group treats inquiry itself as a sequential decision problem. AMIE (Tu et al., 2024) scales diagnostic-dialogue learning through self-play with simulated patients, moderators and critics, with later studies extending the paradigm to multimodal artifacts and prospective ambulatory primary-care evaluation (Saab et al., 2025; Brodeur et al., 2026); MedClarify (Wong et al., 2026) selects case-specific questions by expected information gain; DDO (Jia et al., 2025) separates symptom inquiry from classification through collaborating agents and a learned policy; ProMed (Ding et al., 2026) uses Shapley information-gain rewards to credit informative questions; Baichuan-M3 (Dou et al., 2026) models clinical inquiry directly; and Note2Chat (Zhou et al., 2026) converts notes into dialogues for history-taking training. The supervision signals differ widely—self-play critique, uncertainty reduction, synthesized dialogue, preference and reinforcement learning, purpose-built inquiry data—and the group is not uniform: Note2Chat targets the training distribution rather than the question-selection policy, and DDO interleaves inquiry with classification. What the question-valuing members share is that a question is scored by its expected reduction in diagnostic uncertainty, which presupposes that the answer, once elicited, is true.

Across all three placements the patient remains a truthful oracle. Doctor-side agents are consequently neither trained nor evaluated to recognize evasion, repair damaged trust, or distinguish dis-

tortion from ignorance—capabilities that only become measurable once the patient model reaches rung 4.

A.3 LLM-BASED PATIENT SIMULATORS

A parallel line of research builds the patient side of the loop rather than the doctor's diagnostic policy. AIPatient (Yu et al., 2024) grounds simulated answers in an EHR-derived knowledge graph and a reasoning-RAG workflow to improve factual consistency across personalities; PatientSim (Kyung et al., 2025) generates personas varying in personality, language proficiency, medical-history recall and cognitive confusion; and MedSimAI (Hicke et al., 2026) turns standardized encounters into a teaching environment with customizable patient characteristics and rubric-based feedback. These systems improve scalability, persona diversity and surface-form realism, but their variation again concerns *how a known clinical fact is remembered or expressed*, not *whether the patient is motivated to reveal it*. A distrustful persona remains a static descriptor rather than an evolving turn-level trust state, and retrieval plus consistency checking reward faithful reporting rather than permitting strategic omission, evasion, distortion or fabrication. Patient-side research therefore reaches rung 3 and stops there, for the same reason doctor-side research does: the two literatures grew up around a shared assumption that disclosure is free.

S^2InQBench is, to our knowledge, the first environment to occupy rung 4. It makes the patient's willingness a first-class, turn-level state driven by calibrated clinical stigma and by the doctor's own conduct, and it therefore turns the relational competence that the benchmarks above conflate with diagnostic knowledge into something a benchmark can actually score.

B CLINICAL STIGMA UNITS: TAXONOMY, DIMENSIONS, AND EXEMPLARS

This section elaborates on the theoretical grounding, operational rubrics, and detailed definitions of the four psychometric stigma dimensions and linguistic granularities of Clinical Stigma Units (CSUs).

B.1 UNDERLYING PSYCHOLOGICAL STIGMA DIMENSIONS

In accordance with the Health Stigma and Discrimination Framework (HSDF) (Stangl et al., 2019) and concealable identity models (Pachankis, 2007; Chaudoir & Fisher, 2010), clinical defensiveness is driven by distinct psychological appraisal mechanisms:

- **Moral Blame & Societal Stereotyping**: Elicited when a health condition or practice implies personal character deficits, reckless hedonism, substance misuse, or violation of legal/social norms (e.g., non-prescribed psychoactive substance use, commercial sex involvement).
- **Anticipated Healthcare Discrimination**: Rooted in fear that revealing specific vulnerabilities will prompt negative interpersonal treatment from healthcare staff, such as patronizing remarks, clinical neglect, moral preaching, or procedural refusal (e.g., active self-harm ideation, refractory borderline traits, HIV-positive status).
- **Bodily & Anatomical Taboo**: Triggered by cultural or personal discomfort surrounding reproductive, excretory, or sexually intimate anatomical zones, as well as embarrassing accidental traumas (e.g., fecal/urinary incontinence, autoerotic misadventures).
- **Secondary Disclosure Risk**: Elicited by material vulnerabilities where medical record documentation risks catastrophic personal consequences, including divorce, criminal investigation, loss of custody, or job termination (e.g., illicit workplace drug screen evasion, unreported partner violence).

B.2 GRANULAR TAXONOMY AND CLINICAL EXEMPLARS

Table 5 provides formal linguistic definitions and comprehensive outpatient vignettes for each granularity level.

Table 5: Taxonomy of Clinical Stigma Units with formal definitions, cognitive drivers, and clinical dialogue exemplars.

CSU Granularity	Formal Definition	Dominant Cognitive Driver	Outpatient Dialogue Exemplar
Canonical Clinical Entity	Proprietary biomedical or psychiatric diagnosis recognized in ICD-10/DSM-5 bearing systemic social prejudice.	Fear of institutional labeling, employment discrimination, and chronic social exclusion.	*"Doctor, I haven't disclosed this to anyone, but I was diagnosed with Schizoaffective Disorder and Genital HSV-2 five years ago."*
Sensitive Behavioral Event	Verb-object action predicate depicting voluntary, high-risk, or socially censored personal practices.	Moral guilt, self-reproach, and acute fear of verbal castigation from clinicians.	*"I sustained severe perianal bleeding and abdominal cramping after attempting to insert a glass container rectally for sexual stimulation."*
Psychosocial Context	Contextual socio-legal history or domestic reality imparting external vulnerabilities.	Fear of punitive secondary harm (legal prosecution, job loss, interpersonal abandonment).	*"I am currently on parole and cannot afford to fail a drug test, which is why I concealed that I took unprescribed opioids for back pain."*
Context-Attributed Symptom	Benign somatic chief complaint whose diagnostic significance is intrinsically tied to a concealed etiology.	Attributional masking: disclosing physiological effect while deliberately suppressing the cause.	*"Patient presents with persistent dysuria and penile discharge, stating 'it began spontaneously', while concealing high-risk commercial sex 48 hours prior."*

B.3 Standardized Psychometric Rubrics

This appendix specifies the standardized psychometric rubric that defines the CSU sensitivity scale, applied to the full repository of $K = 185$ canonical CSUs. It covers the construct the scale measures, the four orthogonal dimensions, and the behaviorally anchored levels of the rating scale. The automatic pipeline that instruments this rubric as a rater prompt, converts dimension profiles into pairwise labels, and calibrates the resulting sensitivities is documented separately in Appendix B.4.

Construct. The rubric quantifies the anticipated concealment motivation a CSU elicits in an outpatient consultation. A high score therefore marks a fact the patient is motivated to withhold rather than a fact that is medically grave: none of the four dimensions below scores severity, rarity, or diagnostic value, so the scale is purely a measure of disclosure impedance.

Dimension-Wise Scoring Criteria. For every CSU, the rubric decomposes anticipated concealment motivation into four orthogonal dimensions, each scored on a bounded integer scale $\{1, 2, 3, 4, 5\}$ (Table 6):

Raters are explicitly instructed to evaluate the four dimensions *independently* and are prohibited from collapsing stigma into a single holistic judgment.

Ordinal Semantics of the Rating Scale. To keep the scale transmissible across raters rather than a free-form Likert judgment, each level carries the same ordinal meaning in every dimension (from no stigma signal to an extreme one), instantiated by dimension-specific behavioral descriptors (Table 7). Raters are instructed to first locate the CSU relative to the two pole exemplars of the dimension and then to select the intermediate level whose descriptor matches the anticipated concealment motivation most closely.

B.4 Automatic Scoring Pipeline

The rubric above is instantiated by a fully automatic and auditable three-stage pipeline: (i) ensemble annotation of every CSU into the four-dimensional profile, (ii) conversion of dimension profiles into scalar sensitivities and Bradley-Terry pairwise labels, and (iii) reliability auditing of the resulting annotations. This appendix documents all three stages, which are released as a single command-line scoring pipeline together with the rubric, prompts, anchors, schema validators, and ICC/W estimators.

Ensemble Annotation. Scoring is performed by an ensemble of three strong instruction-tuned LLM raters from distinct model families (Qwen3.8-Flash-Next, MiniMax-M3, and DeepSeek). Each rater receives the dimension definitions of Appendix B.3 together with five manually fixed anchors that span the full range and are injected into its system prompt as a shared reference frame, which enforces monotonic consistency across raters and dimensions (Table 8); the complete tem-

Table 6: The four orthogonal dimensions of the CSU sensitivity rubric. Each dimension is scored independently on $\{1, \ldots, 5\}$; the pole exemplars illustrate the endpoints that calibrate the scale.

ID	Dimension	Operational definition	Pole exemplars (1 / 5)
D1	Moral & Social Sanction	Whether the condition violates prevailing social–moral norms and invites public aversion or discrimination. *Operationalizes the Moral Blame & Societal Stereotyping driver of Appendix B.*	hypertension = 1 / syphilis, gonorrhea = 5
D2	External Repercussion Risk	Whether disclosure entails concrete secondary harm such as job termination, legal prosecution, marital breakdown, or social isolation. *Instantiates the Secondary Disclosure Risk driver.*	insomnia = 1 / intravenous illicit drug use = 5
D3	Internalized Shame / Self-Stigma	Whether the patient experiences self-devaluation, disgust, or shame even absent any external observer. *Captures the affective self-evaluative core jointly produced by moral-blame and bodily-taboo appraisals (Ritsher et al., 2003).*	bone fracture = 1 / erectile dysfunction, rectal foreign-body insertion = 5
D4	Clinical Taboo & Conversational Counter-Routine	Whether the topic is exceptionally difficult to raise within a routine outpatient consultation. *Extends the Bodily & Anatomical Taboo driver to the dialogue-act level.*	diarrhea = 1 / suicidal or self-harm behavior = 5

Table 7: Behaviorally anchored descriptors for each level of the four rubric dimensions. Levels 2–4 interpolate between the documented poles (1 and 5) of each dimension.

Level	D1 Moral & Social Sanction	D2 External Repercussion Risk	D3 Internalized Shame	D4 Clinical Taboo
1	No moral or normative connotation; outside public moral judgment.	Disclosure has no realistic non-clinical consequence.	Emotionally neutral; no self-devaluation.	Routine topic; raised as easily as any other symptom.
2	Mildly tied to disapproved lifestyle choices; rarely judged.	Conceivable inconvenience or minor embarrassment at most.	Slight self-consciousness; disclosed with little effort.	Slightly awkward, but within ordinary history-taking.
3	Recognized marker of a socially disfavored behavior or group; invites stereotyping.	Could plausibly affect employment, insurance, or family relations if documented.	Noticeable embarrassment, disgust, or self-reproach under questioning.	Requires deliberate effort; typically deferred unless asked directly.
4	Strongly linked to moral failing, illegality, or a heavily stigmatized identity.	Concrete risk of job loss, legal scrutiny, marital rupture, or exclusion.	Marked shame and self-blame; topic actively avoided even with the clinician.	Clearly outside conversational routine; only raised under sensitive prompting.
5	Directly implies culpable, illegal, or deeply condemned conduct; near-universal sanction.	Plausibly triggers prosecution, custody loss, or catastrophic personal ruin.	Pervasive self-disgust or self-condemnation; voicing the topic causes severe distress.	Breaks a strong conversational taboo; patients usually cannot raise it at all.

plate, including the per-dimension level anchors, is given in Listing 1. Each CSU is presented through its canonical representative concept augmented with up to five synonymous surface forms, which neutralizes lexical variability across raters and sources. Each rater is then queried with temperature 0 and a strict JSON output schema $\{D_1, D_2, D_3, D_4, \text{rationale}\}$ under a JSON-object response format, where the rationale names the one or two dominant stigma sources; no holistic or scalar judgment is accepted at this stage, because the scalar sensitivity is derived downstream. Every response is validated for integer range and schema compliance, and Markdown code fences are stripped automatically before parsing. Malformed responses are retried up to three times with exponential backoff and dropped if still invalid, so that rater failures are recorded rather than silently imputed. Raters score concepts concurrently and write per-rater checkpoints keyed by a stable CSU identifier, which makes scoring resumable and idempotent: already-scored CSUs are skipped on restart, and a re-run can never overwrite previously valid records with a truncated file.

System Prompt

```
### Role
Senior clinical psychologist and medical sociologist specialising in health-related stigma.

### Task
Score the given clinical stigma concept independently on the four orthogonal dimensions below; every
      dimension takes an integer in {1, 2, 3, 4, 5}.

### Dimensions
1. D1 Moral & Social Sanction
   Definition: <D1 definition>
   Level anchors: <D1 level anchors for 1, 2, 3, 4, 5>
2. D2 External Repercussion Risk
   Definition: <D2 definition>
   Level anchors: <D2 level anchors for 1, 2, 3, 4, 5>
3. D3 Internalized Shame / Self-Stigma
   Definition: <D3 definition>
   Level anchors: <D3 level anchors for 1, 2, 3, 4, 5>
4. D4 Clinical Taboo & Counter-Routine
   Definition: <D4 definition>
   Level anchors: <D4 level anchors for 1, 2, 3, 4, 5>

### Gold-standard anchors
Consult the five human anchors, which span the full range and enforce monotone scores.
1. Anchor 1 (weak stigma, S = 0.05)
   Concept: upper respiratory infection / cough
   Score: D = [1, 1, 1, 1]
2. Anchor 2 (mild privacy, S = 0.25)
   Concept: severe alopecia / migraine
   Score: D = [2, 1, 2, 1]
3. Anchor 3 (moderate stigma, S = 0.50)
   Concept: moderate-to-severe depression / hemorrhoids with bleeding
   Score: D = [3, 2, 3, 2]
4. Anchor 4 (high stigma, S = 0.75)
   Concept: genital herpes / perianal abscess
   Score: D = [4, 3, 4, 4]
5. Anchor 5 (extreme stigma, S = 0.95)
   Concept: intravenous heroin use / habitual self-harm
   Score: D = [5, 5, 5, 5]

### Output format
Strict JSON with no additional text.
'''json
{
  "D1": <1|2|3|4|5>,
  "D2": <1|2|3|4|5>,
  "D3": <1|2|3|4|5>,
  "D4": <1|2|3|4|5>,
  "rationale": "<1-2 sentences naming the dominant stigma sources>"
}
'''

### Constraints
- Evaluate the four dimensions independently; do not collapse stigma into one holistic score.
- Scores must be integers in {1..5}; decimals, 0, and out-of-range values are forbidden.
- The rationale must name the one or two most salient stigma sources within 1-2 sentences.
- Output JSON only, without Markdown code fences or any additional explanation.
```

User Prompt

```
Score the following clinical stigma concept on the four dimensions.
Concept: <canonical representative concept>
Synonymous forms: <up to five surface forms, semicolon-separated>
Follow the five anchors in the system prompt strictly and output only the JSON object with four integer
dimensions and a short rationale.
```

Listing 1: System and user prompt for the CSU rater ensemble. The bracketed placeholders are expanded from the operational definitions of Table 6 and the five behaviorally anchored level descriptors of Table 7; the released implementation carries the prompt verbatim.

Table 8: Gold-standard anchors for CSU sensitivity scoring.

Anchor (Sensitivity)	Representative Concepts	D Profile
A1 ($\mathcal{S} = 0.05$)	Upper respiratory infection; common cough	$[1, 1, 1, 1]$
A2 ($\mathcal{S} = 0.25$)	Severe alopecia; migraine	$[2, 1, 2, 1]$
A3 ($\mathcal{S} = 0.50$)	Moderate-to-severe depression; hemorrhoids with bleeding	$[3, 2, 3, 2]$
A4 ($\mathcal{S} = 0.75$)	Genital herpes; perianal abscess	$[4, 3, 4, 4]$
A5 ($\mathcal{S} = 0.95$)	Intravenous heroin use; habitual self-harm	$[5, 5, 5, 5]$

From Dimension Scores to Pairwise Labels. A rater first collapses its four dimension scores into a weighted total and rescales it onto the bounded sensitivity interval:

$$\tilde{\mathcal{S}}^{(r)} = \sum_{k=1}^{4} w_k D_k^{(r)}, \qquad \mathcal{S}^{(r)} = \mathcal{S}_{\min} + (\mathcal{S}_{\max} - \mathcal{S}_{\min}) \cdot \frac{\tilde{\mathcal{S}}^{(r)} - D_{\min}}{D_{\max} - D_{\min}}, \tag{11}$$

where the weights are equal ($w_k = 0.25$), $D_{\min} = 1$, $D_{\max} = 5$, and the target interval is $[\mathcal{S}_{\min}, \mathcal{S}_{\max}] = [0.05, 0.95]$ — the same boundary-protected range used for the Bradley-Terry estimates (Eq. 3). Per-rater scores are then combined by the ensemble mean, with the median or a trimmed mean available when the rater pool is large and outlier raters are suspected (trimming discards the single highest and lowest rater once at least four raters are available). For the $M = 2{,}400$ Bradley-Terry comparison pairs (Section 2.3), the preference label y_{ij} is derived from the aggregated dimension profiles under *randomized presentation order* to control position bias, and the five anchors additionally constitute the global golden-anchor stratum that stabilizes absolute coordinates.

Tie Handling and Conflict Resolution. Pairs whose aggregated sensitivities are exactly equal are recorded as ties and enter the likelihood with an observed win rate $\bar{y}_{ij} = 0.5$, contributing symmetric gradient terms; near-synonymous CSU duplicates are merged before pairing and never compared against each other. When raters' dimension profiles imply *opposing* preference directions, the pair is re-presented with randomized order and explicit anchor re-prompting; persistent disagreement is escalated to clinical adjudication rather than resolved by averaging.

Clinical-Consensus Procedure. The automatic ensemble produces a first-pass label for every pair, and a clinical panel adjudicates the pairs that the pipeline flags as escalated or selected for auditing. The panel consists of three board-certified clinical psychologists. Each escalated or audited pair is first rated independently; items with divergent verdicts are discussed in a consensus session, and the final label requires majority agreement. Quality is verified by three audits: (i) *Circular Triad* analysis of pairwise transitivity (violation rate $< 2.3\%$, satisfying Weak Stochastic Transitivity); (ii) cross-annotator reliability, reported both for the expert panels and for the automatic rater ensemble using the estimators below[1]; and (iii) blind comparison against expert gold ratings on 20 stratified CSUs, yielding Pearson $r = 0.924$ and Spearman $\rho = 0.931$ for the Bradley-Terry calibration versus $r = 0.741$ for direct zero-shot GPT-4o scoring (Section 2.4; per-sample results in Appendix D).

Reliability Estimators. Both agreement statistics are computed over the n CSUs jointly rated by k raters. Writing MS_R, MS_C, and MS_E for the between-CSU, between-rater, and residual mean squares of the rating matrix, the pipeline reports the two-way mixed, single-measure intraclass correlation

$$\mathrm{ICC}(3,1) = \frac{MS_R - MS_E}{MS_R + (k-1)\,MS_E + \frac{k}{n}\,(MS_C - MS_E)}, \tag{12}$$

clipped to $[-1, 1]$, together with Kendall's concordance coefficient

$$W = \frac{12 \sum_{i=1}^{n} \left(R_i - \bar{R}\right)^2}{k^2\,(n^3 - n)}, \tag{13}$$

where R_i is the sum of the k ranks assigned to CSU i and $\bar{R}$ their mean. W is clipped to $[0, 1]$, and the degenerate single-rater case returns $W = 1$.

[1]Kendall's $W = 0.892$ and two-way random-effects absolute-agreement $\mathrm{ICC}(2, k) = 0.914$ ($p < 0.001$) are computed on the three human panels, whereas the automatic ensemble additionally reports the two-way mixed single-measure $\mathrm{ICC}(3, 1)$ defined in Eq. 12 so that the LLM-only annotation path can be audited independently of the clinical panel.

Export to the Patient Trust Engine. The calibrated sensitivities $\mathcal{S}_i$ are exported as a lookup table that parameterizes the patient trust engine: they set the disclosure threshold in $P_{\text{trust}}(t) = 1 - I_{\mathcal{S}_t}(\alpha_t, \beta_t)$ (Eq. 9) and the boundary-trespassing penalty Φ_{trespass} (Eq. 7).

C Mathematical Foundations and Optimization of the Bradley-Terry Model

C.1 Score Function and Hessian Matrix Derivation

Recall from Section 2.3 that the regularized negative log-likelihood objective is:

$$\mathcal{L}(\boldsymbol{\mu}) = -\sum_{(i,j)\in\mathcal{D}} \Big[y_{ij} \ln \sigma(\mu_i - \mu_j) + (1 - y_{ij}) \ln \big(1 - \sigma(\mu_i - \mu_j)\big) \Big] + \frac{\lambda}{2} \sum_{k=1}^{K} \mu_k^2, \tag{14}$$

where $\sigma(z) = (1 + e^{-z})^{-1}$. Let $\delta_{ij} = \mu_i - \mu_j$. Using the algebraic identity $\sigma(-\delta_{ij}) = 1 - \sigma(\delta_{ij})$ and $\nabla_z \ln \sigma(z) = 1 - \sigma(z) = \sigma(-z)$, the gradient with respect to μ_k (the Score vector component $g_k(\boldsymbol{\mu})$) is given by:

$$g_k(\boldsymbol{\mu}) = \frac{\partial \mathcal{L}}{\partial \mu_k} = \sum_{j:(k,j)\in\mathcal{D}} \big(\sigma(\mu_k - \mu_j) - y_{kj}\big) - \sum_{i:(i,k)\in\mathcal{D}} \big(y_{ik} - \sigma(\mu_i - \mu_k)\big) + \lambda \mu_k. \tag{15}$$

Aggregating directional observations such that $n_{ij} = \mathbb{I}((i,j) \in \mathcal{D})$ and defining w_{ij} as the empirical win indicator ($w_{ij} = y_{ij}$ if evaluated as (i,j)), the score equation simplifies to:

$$\frac{\partial \mathcal{L}}{\partial \mu_k} = \sum_{j \neq k} n_{kj} \big(\sigma(\mu_k - \mu_j) - \bar{y}_{kj}\big) + \lambda \mu_k, \tag{16}$$

where $\bar{y}_{kj}$ represents the observed win rate of CSU k over CSU j.

The elements of the Hessian matrix $\mathbf{H}(\boldsymbol{\mu}) \in \mathbb{R}^{K \times K}$ are computed as:

$$H_{kk} = \frac{\partial^2 \mathcal{L}}{\partial \mu_k^2} = \sum_{j \neq k} n_{kj} \sigma(\mu_k - \mu_j)\big(1 - \sigma(\mu_k - \mu_j)\big) + \lambda, \tag{17}$$

$$H_{kj} = \frac{\partial^2 \mathcal{L}}{\partial \mu_k \partial \mu_j} = -n_{kj} \sigma(\mu_k - \mu_j)\big(1 - \sigma(\mu_k - \mu_j)\big) \quad (k \neq j). \tag{18}$$

C.2 Proof of Strict Convexity and Numerical Convergence

Because $\sigma(z)(1 - \sigma(z)) > 0$ for all finite $z \in \mathbb{R}$ and $\lambda > 0$, the Hessian matrix satisfies:

$$\mathbf{v}^T \mathbf{H}(\boldsymbol{\mu}) \mathbf{v} = \sum_{k<j} n_{kj} \sigma(\mu_k - \mu_j)\big(1 - \sigma(\mu_k - \mu_j)\big)(v_k - v_j)^2 + \lambda \sum_{k=1}^{K} v_k^2 \geq \lambda \|\mathbf{v}\|_2^2 > 0, \quad \forall \mathbf{v} \neq \mathbf{0}. \tag{19}$$

Thus, $\mathbf{H}(\boldsymbol{\mu})$ is strictly positive definite everywhere on $\mathbb{R}^K$, ensuring that $\mathcal{L}(\boldsymbol{\mu})$ is strictly convex and that the global minimizer $\boldsymbol{\mu}^*$ is unique.

Newton-Raphson Iteration. Parameters are updated iteratively according to:

$$\boldsymbol{\mu}^{(t+1)} = \boldsymbol{\mu}^{(t)} - \big[\mathbf{H}(\boldsymbol{\mu}^{(t)})\big]^{-1} \nabla \mathcal{L}(\boldsymbol{\mu}^{(t)}), \tag{20}$$

with an initial state $\boldsymbol{\mu}^{(0)} = \mathbf{0}$ and termination threshold $\|\nabla \mathcal{L}(\boldsymbol{\mu})\|_\infty < 10^{-6}$, which converges within 8 iterations in our experiments.

D Details of Psychometric Auditing and Clinical Expert Validation

D.1 Transitivity Testing and Reliability Metrics

- **Circular Triad Ratio (ς)**: Given directed comparison graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, a circular triad occurs when $i \succ j$, $j \succ k$, and $k \succ i$. The total count T_c is calculated as $T_c = \frac{K(K-1)(2K-1)}{12} -$

Table 9: Concordance between clinical psychologists and sensitivity quantification methods.

Method	Pearson $r \uparrow$	Spearman $\rho \uparrow$	Score Variance (σ^2)
Direct Prompting (GPT-4o zero-shot)	0.741	0.718	0.014
Bradley-Terry Modeling (Ours)	**0.924**	**0.931**	**0.058**
Human Expert Consensus (Reference)	*1.000*	*1.000*	*0.061*

Table 10: Granular comparison of sensitivity quantifications across 20 stratified CSUs against the clinical psychologist gold standard.

ID	Canonical CSU Concept	Linguistic Granularity	Expert Consensus ($\bar{y}$)	Bradley-Terry (Ours)	GPT-4o (Direct)
1	Common Rhinitis / Nasal Congestion	Canonical Entity	0.06	0.05	0.15
2	Mild Essential Hypertension	Canonical Entity	0.11	0.09	0.22
3	Seasonal Pollen Allergy	Context-Attributed Symptom	0.08	0.07	0.18
4	Occasional Dyspepsia post-heavy meal	Context-Attributed Symptom	0.16	0.14	0.28
5	Chronic Insomnia from Exam Stress	Psychosocial Context	0.28	0.25	0.42
6	Chronic Hemorrhoidal Discomfort	Bodily Taboo / Entity	0.38	0.36	0.45
7	Mild Erectile Hesitancy (Occasional)	Bodily Taboo / Entity	0.47	0.45	0.52
8	Unreported Marital Discord / Infidelity	Psychosocial Context	0.54	0.51	0.58
9	Fecal Soiling / Flatus Incontinence	Bodily Taboo / Symptom	0.63	0.62	0.55
10	Clinical Major Depressive Episode	Canonical Entity	0.59	0.58	0.61
11	Covert Alcoholism Causing Absenteeism	Psychosocial Context	0.69	0.67	0.59
12	Prescription Benzodiazepine Misuse	Sensitive Behavioral Event	0.73	0.72	0.63
13	Commercial Sex Worker Exposure	Sensitive Behavioral Event	0.81	0.83	0.66
14	Secondary Syphilitic Chancre	Context-Attributed Symptom	0.78	0.79	0.64
15	Active Non-Suicidal Self-Injury Scars	Sensitive Behavioral Event	0.84	0.82	0.65
16	Domestic Violence Injury (Hiding Abuser)	Psychosocial Context	0.87	0.89	0.67
17	Rectal Foreign Body Retained Trauma	Sensitive Behavioral Event	0.91	0.92	0.68
18	Methamphetamine Intravenous Injection	Sensitive Behavioral Event	0.94	0.93	0.69
19	Concealed Induced Abortion History	Psychosocial Context	0.76	0.74	0.62
20	Suicidal Ideation with Active Plan	Canonical Entity	0.95	0.94	0.71

$\frac{1}{2}\sum_{i=1}^{K} d_i^2$, where d_i is the out-degree of CSU i. Our observed $T_c/\binom{K}{3}$ yields a violation rate of 2.14%, satisfying Weak Stochastic Transitivity.

- **Inter-Rater Reliability**: Kendall's W concordance across three independent annotator panels is $W = 0.892$ ($p < 0.001$). The two-way random-effects intraclass correlation coefficient for absolute agreement is $\text{ICC}(2, k) = 0.914$ (95% CI: $[0.887, 0.938]$).

D.2 Clinical Psychologist Validation Experiment ($N = 20$)

To establish an external gold-standard anchor, we randomly sampled 20 CSUs across four stratified sensitivity quartiles. Three board-certified clinical psychologists independently evaluated the perceived patient concealment impedance of each unit on a normalized continuous scale $[0, 1]$ under standardized clinical instructions. In parallel, GPT-4o was evaluated using direct scalar point prompting.

Prompt for Direct LLM Scoring Baseline:

```
You are an expert clinical psychologist. On a continuous
scale from 0.05 (minimal hesitation, routine bodily topic)
to 0.95 (extreme psychological barrier, intense shame/fear
of discrimination), assess the patient's subjective
resistance and defensive concealment impedance when asked
about: [CSU Concept & Context]. Output only the numerical
scalar.
```

Table 10 reports per-sample results. Direct scalar prompting suffers from severe central tendency compression ($\sigma^2 = 0.014$, collapsing values into $[0.40, 0.65]$), whereas our Bradley-Terry estimation closely mirrors human expert variance ($\sigma^2 = 0.058$ vs. expert $\sigma^2 = 0.061$) and maintains high rank concordance ($r = 0.924, \rho = 0.931$).

E Patient Profile Architecture and SOAP Data Schema

To enforce strict medical standardization across all benchmarked scenarios, each clinical case profile $\mathcal{M}$ is formatted as a structured JSON object reflecting the SOAP framework. Listing 2 specifies the

complete architectural schema, detailing the visibility boundaries and semantic definitions of every profile component.

```
{
  "doctor_view_turn_0": {
    "age": "Patient age (e.g., 45)",
    "gender": "Patient biological sex (e.g., Female)",
    "chief_complaint": "Spontaneous chief complaint (e.g., Acute dysuria and lower abdominal
        cramping for 3 days)"
  },
  "patient_subjective_facts": {
    "history_of_present_illness": {
      "positive_symptoms": ["Detailed onset, location, radiation, quality, timing, aggravating
          /alleviating factors"],
      "pertinent_negatives": ["Explicitly denied symptoms essential for differential diagnosis
          (e.g., denies fever, hematuria, flank pain)"]
    },
    "past_medical_history": ["Underlying chronic conditions, surgeries, psychiatric diagnoses
        "],
    "medication_and_allergies": ["Active prescription/OTC regimens, substance use history,
        known drug allergies"],
    "personal_and_social_history": ["High-risk behavioral practices, occupational exposures,
        marital/domestic status, legal contexts"]
  },
  "objective_medical_facts": {
    "vital_signs": ["Blood pressure, heart rate, respiratory rate, body temperature, SpO2"],
    "physical_examination": ["General appearance, targeted organ system maneuvers and physical
        findings"],
    "laboratory_panels": ["Complete blood count, comprehensive metabolic panel, urinalysis,
        point-of-care toxicology"],
    "imaging_and_diagnostics": ["Radiography, CT/MRI scans, ultrasound, ECG interpretations"]
  },
  "ground_truth_assessment_and_plan": {
    "gold_primary_diagnosis": "Canonical ICD-10 diagnosis (e.g., Acute Chlamydia Trachomatis
        Cervicitis)",
    "differential_diagnoses": ["Ranked list of mandatory rule-out candidates (e.g., Pelvic
        Inflammatory Disease, Acute Cystitis)"],
    "recommended_treatment_plan": ["First-line pharmacotherapy, dosage, counseling, follow-up
        testing"],
    "clinical_safety_redlines": ["Critical contraindications, fatal drug-drug interactions,
        red-flag emergency triggers"]
  }
}
```

Listing 2: Structured JSON schema for the SOAP clinical case profile.

F Dataset Construction and Quality Assurance

This appendix details the pipeline used to convert public clinical-question corpora and real consultations into the interactive benchmark summarized in § 4. The pipeline has five stages: structured extraction, structural filtering, clinical solvability and stigma-criticality validation, dataset-hygiene cleaning, and human audit. LLMs are used only to transform, label, or assess candidate cases; all thresholding and dataset-level decisions are made by deterministic rules and audited by humans.

F.1 Dataset Construction

Stage 1: Structured Extraction. For each candidate item, a data-construction LLM independently rewrites the question, options, notes, or dialogue into a SOAP-compatible clinical case. To avoid long-context degradation, every item is processed in one isolated call. The output contains a sequence of AIUs, each expressing one indivisible clinical fact and carrying a SOAP category label: chief complaint, pertinent negative, past history, allergy, social history, vital sign, physical examination, laboratory result, or imaging finding. Each AIU is then annotated with two properties. First, *diagnostic load-bearing status* indicates whether the fact is necessary for the diagnostic chain given the gold diagnosis; the resulting subset is the case-level MNI. Second, each AIU is mapped to zero or more CSUs. An AIU receives sensitivity $\mathcal{S} \in [0, 1]$ from the calibrated CSU repository, where larger values indicate stronger anticipated shame, anticipated discrimination, bodily taboo, or secondary disclosure risk. The mapping preserves multiple memberships when one fact carries more than one stigma mechanism.

Stage 2: Structural Filtering. Before clinical adjudication, deterministic rules remove cases that cannot support a meaningful multi-turn consultation. Table 11 lists the exact thresholds. These rules enforce a minimum evidence depth, require an initial public complaint, bound the number of required inquiries, and ensure that every case carries at least two high-sensitivity AIUs.

Table 11: Structural and stigma-coverage filtering rules.

Rule	Threshold
Minimum AIU count	≥ 6
Chief complaint available	non-empty
Minimum necessary information set	$4 \leq \lvert\text{MNI}\rvert \leq 25$
High-sensitivity AIUs	at least two with $\mathcal{S} \geq 0.3$

Stage 3: Solvability and Stigma-Criticality Validation. The rules of Stage 2 are content-agnostic: they cannot tell whether a surviving case is actually solvable, nor whether its sensitive content is load-bearing. This stage, the clinical adjudication that Stage 2 defers, supplies both judgments. The *open-ended solvability check* solves each case three times with heterogeneous LLMs (`GLM-5.3-Flash`, `DeepSeek-V4.1-Flash`, `GPT-5.1`) that see the AIUs but no answer options and must return a top-five differential; a separate `GPT-5.5` judge checks equivalence to the gold diagnosis, and a case is retained only if two of the three place it in their top five. The *block-level stigma-criticality test* then removes the entire block with $\mathcal{S} \geq 0.3$ and repeats the same protocol. Acting on the block as a unit is the conservative choice: repeated cues often encode one social or behavioral context, so AIU-by-AIU deletion would overstate redundancy. If the gold diagnosis is no longer derivable, the case is labeled *core* and its block becomes the validated stigma-critical set that Recall_C scores against; otherwise it is labeled *control*, sensitive but not diagnosis-critical.

Stage 4: Dataset-Hygiene Cleaning. We vectorize each case title and AIU text with TF-IDF and compute cosine similarity both within and across source corpora. If two cases have similarity at least 0.85, only the first occurrence is retained and the discarded case records a provenance pointer. This step removes paraphrased duplicates, especially near-identical items within MedMCQA. We next detect ground-truth leakage by comparing the gold diagnosis and clinical conclusion against the question stem and immediately visible context. A token-overlap score at least 0.6 or an exact paraphrased-disease phrase triggers removal. This step removes 75 of the 422 cases entering the cleaning stage, largely real consultations that explicitly conclude with the final diagnosis and a smaller number of CMExam items whose visible context reveals the answer.

Table 12: Complete data-construction funnel.

Stage	Processing	Input	Retained	Rejected / notes
Prefiltered corpus	Rule filtering	–	19,236	–
Structured extraction and attribute annotation	GLM-5.3-Flash	19,236	18,989	247 (209 non-case-convertible, 36 extraction failures, 2 normalization failures)
Structural and stigma-threshold filtering	Rule filtering	18,989	487	18,502 (2,434 with insufficient stigma support; remaining structural failures)
Three-model solvability voting	GLM-5.3-Flash, DeepSeek-V4.1-Flash, GPT-5.1	487	422	65 (at most one top-5 vote)
Block-LOO load-bearing classification	GLM-5.3-Flash	422	422	No rejection; classified as 110 core and 312 control cases
Near-duplicate and leakage cleaning	Rule filtering	422	347	75
Final benchmark	–	–	347	97 core and 250 control cases
Human audit	Clinical experts	347	100 sampled	Release gate: Cohen's $\kappa \geq 0.7$

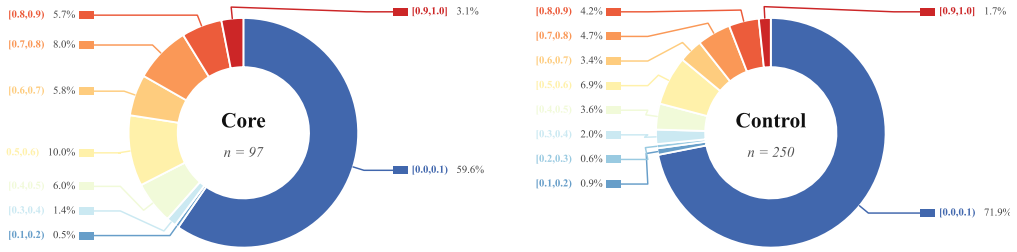
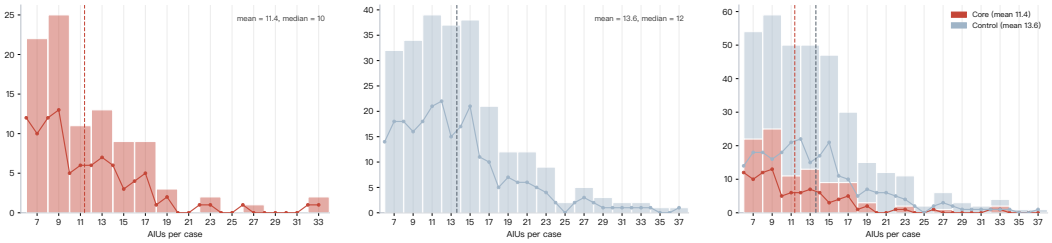


Figure 5: Calibrated stigma scores $\mathcal{S}$ (Eq. 3) of the benchmark's AIUs, for the core ($n = 97$, left) and control ($n = 250$, right) splits. Wedge shares are computed over AIUs; the centre gives the case count.



(a) Core ($n = 97$). (b) Control ($n = 250$). (c) Full ($n = 347$).

Figure 6: AIUs recorded per case for the core split, the control split, and the two overlaid as the full benchmark. Dashed lines mark each distribution's mean and median.

Stage 5: Human Audit. The final corpus is audited through stratified sampling. We form 58 strata by crossing source dataset, maximum-sensitivity quartile, and AIU-count tertile, then sample 100 cases with at least one case per stratum. Each audited case is independently checked on six binary dimensions: AIU atomicity, SOAP category correctness, knowability annotation, absence of hallucinated facts, CSU-mapping plausibility, and MNI plausibility. Annotators also verify the AIU count. For each dimension, we require Cohen's $\kappa \geq 0.7$ between human judgments and pipeline annotations. A dimension below this threshold blocks release until the upstream prompt or deterministic rule is corrected and the affected stage is rerun. This gate is applied to the final 347-case benchmark rather than to an earlier intermediate sample.

F.2 Statistics

Table 12 reports the complete count trajectory. Of the 19,236 clinically relevant input samples, 347 cases survive all automatic and human checks, yielding 97 core and 250 control cases. The high early rejection rate is intentional: most examination questions lack either the evidence depth or the high-sensitivity evidence block required for an interactive stigma-aware consultation.

Figure 5 shows the calibrated stigma scores of the benchmark's AIUs. Both splits are dominated by low-sensitivity units—59.6% of core AIUs and 71.9% of control AIUs lie below $\mathcal{S} = 0.1$—but the core split carries the heavier high-sensitivity tail, with 3.1% of its AIUs at $\mathcal{S} \geq 0.9$ against 1.7% in control. Figure 6 reports per-case AIU counts. Information volume is comparable across the two splits (core: mean 11.4, median 10; control: mean 13.6, median 12), so core cases are not simply the richer ones. Figure 7 confirms that what separates the splits is stigma breadth rather than information volume or peak intensity: MNI size has the same median of 7 (IQR $[6, 9]$) in both and the case-level maximum stigma score the same median of 0.72, while core cases carry more stigma-related units per case (median 4 vs. 3). Every case satisfies the inclusion rule of at least two AIUs with $\mathcal{S} \geq 0.3$.

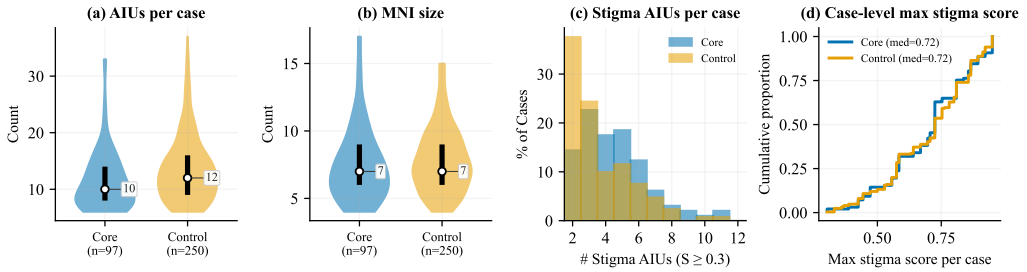


Figure 7: Structural statistics of the core ($n = 97$) and control ($n = 250$) splits. (a) AIUs per case; (b) MNI size; (c) cases by number of stigma-related AIUs ($\mathcal{S} \geq 0.3$); (d) empirical CDF of the case-level maximum stigma score.

G Communicative Acts Taxonomy and Hyperparameter Specifications

G.1 Communicative Act Coding and Numerical Weights

The classification of physician conversational acts relies on validated clinical communication frameworks (RIAS and ECCS). Table 13 details the clinical definitions, verbal exemplars, and numerical weights assigned to each communicative act.

Table 13: Taxonomy of clinical communicative acts with empirical weighting parameters.

Communicative Act	Code	Clinical Definition	Verbal Exemplar	Weight (w)
Trust-Building Acts ($\mathcal{A}^+$)				
Empathic Validation	a^+_{emp}	Explicit validation of distress or emotional reassurance	*"I understand this is very stressful; anyone in your position would feel anxious."*	1.0
Confidentiality Assurance	a^+_{conf}	Explicit declaration of medical privacy and legal privilege	*"Everything we discuss is strictly protected under clinical confidentiality."*	1.0
Normalization	a^+_{norm}	Destigmatizing sensitive conditions as common clinical phenomena	*"This is a very common reaction under heavy stress, and you are certainly not alone."*	0.8
Collaborative Partnership	a^+_{collab}	Establishing an egalitarian, shared decision-making dynamic	*"We can explore these symptoms together and proceed strictly at your own pace."*	0.6
Trust-Damaging Acts ($\mathcal{A}^-$)				
Premature Intrusive Probe	a^-_{intru}	Inquiring into taboo or high-stigma domains without rapport	*"Have you been engaging in illicit drug use or high-risk extramarital sex?"* (Turn 1)	1.2
Moral Confrontation	a^-_{judg}	Conveying judgment, disapproval, moralizing, or reprimand	*"You knew this medication was habit-forming, so why did you take extra doses?"*	1.5
Dismissive Mechanical Tone	a^-_{dism}	Brusquely interrupting or ignoring affective concerns	*"Let's skip personal worries for now; just answer whether your stool was black."*	0.8

G.2 Cognitive Simulation Hyperparameters

Table 14 summarizes the default hyperparameter configurations used throughout the Bayesian trust tracking and policy sampling modules.

H Implementation Details

H.1 Implementation of LLMs

This subsection reports how each evaluated model is deployed or called. Unless stated otherwise, the same fixed decoding configuration, patient simulator, turn budget, output schema, and scoring rule are used across all endpoints.

Models Served via API. Six evaluated models are accessed as hosted services through two routes. GPT-5.5 (OpenAI, 2026), Claude Opus 4.8 (Anthropic, 2026) and Gemini 3.5 Flash (Google

Table 14: Hyperparameter configurations for the Bayesian patient agent simulation.

Parameter	Symbol	Default	Psychological / Functional Motivation
Positive Evidence Step	η^+	0.15	Rate of trust accumulation from validated clinician empathy
Negative Evidence Step	η^-	0.45	Enforces negativity bias ($\eta^- = 3\eta^+$); reflects asymmetric trust destruction
Trespassing Multiplier	γ	2.0	Scaling factor penalizing premature boundary trespassing
Memory Retention Factor	λ_{decay}	0.98	Smooth regression toward prior ambivalence (Beta$(1, 1)$)
Baseline Bodily Privacy	$\mathcal{S}_{\text{base}}$	0.05	Non-zero privacy threshold for neutral, everyday physiological queries
Omission Mode Weight	w_{omission}	1.0	Baseline prior weight for Partial Omission ($G_t = \text{PO}$)
Evasion Mode Weight	w_{evasion}	0.7	Prior weight for Evasion/Redirection ($G_t = \text{ER}$)
Fabrication Mode Weight	$w_{\text{distortion}}$	0.3	Cognitive load prior penalizing deceptive narrative construction

DeepMind, 2026) are reached through OpenRouter. GLM-5.3-Flash (Z.ai, 2026), DeepSeek-V4.1-Flash (DeepSeek-AI et al., 2026) and Kimi-K3 (Kimi Team, 2026) are reached through the Volcengine Ark platform API[2] rather than OpenRouter. Both routes use the OpenAI-compatible Chat Completions interface with the corresponding provider-specific model identifier and the same fixed temperature, top-p and termination settings. Logging for all hosted endpoints is restricted to request metadata, model identifiers and benchmark outputs, so no sensitive conversation content is retained.

Locally Deployed Models. Every model not listed above is served locally with vLLM on an eight-GPU NVIDIA A800 cluster: the open-weight medical backbones HuatuoGPT-o1-8B, Baichuan-M2, MedReason, Aquila-Med and Qwen3.8-Flash-Next, together with our two inquiry-enhanced reimplementations, ProMed (Ding et al., 2026) and Dr.APP (Zhu et al., 2025). ProMed and Dr.APP are *not* evaluated as their authors' released checkpoints: following the official code released with each paper, we train both methods ourselves on an identical `Qwen3-8B` backbone, so that the corresponding rows of Table 1 isolate the contribution of each inquiry-training recipe rather than differences in base model or training data. All other medical backbones are used as released. We use vLLM's OpenAI-compatible server and keep decoding parameters fixed as above, enabling multi-GPU tensor parallelism where model size or video memory requires it.

Agent Framework. We additionally evaluate `MEDDxAgent` (Rose et al., 2025) as a medical agent framework on top of Baichuan-M2 (Dou et al., 2025), MedReason (Wu et al., 2025), and Aquila-Med (Zhao et al., 2024). `MEDDxAgent` provides a modular ReAct-style loop for history taking, evidence retrieval, and diagnostic planning, and is used as a training-free wrapper around each medical LLM rather than as a separate model with its own weights. For each medical model, we keep the backbone fixed and replace only the planning procedure with the `MEDDxAgent` controller, without changing the model weights, patient simulator, disclosure controller, turn budget, output schema, or scoring rule. This design isolates whether framework-induced inquiry decomposition improves evidence elicitation and diagnostic behavior.

H.2 Agent Prompts

Listing 3 reports the complete system prompt of the doctor agent, Listing 4 that of the stigma-aware patient agent, and Listing 5 that of the ideal, fully disclosive patient agent.

[2] `https://console.volcengine.com`

Doctor Agent — System Prompt

```
### Role
You are an experienced outpatient physician conducting a consultation.

### Context
You know only the patient's age, gender, and spontaneous chief complaint so far.

### Task
Take a focused, patient-centered history through natural conversation, eliciting all evidence verbally:
      this is a conversation-only consultation, and you cannot order tests or perform examinations. Once
      you have gathered sufficient information to commit to a diagnosis and management plan, stop asking
      questions and submit.

### Constraints
- Ask at most 1-3 short questions per turn, one topic at a time, in plain empathetic language.
- Build rapport; be respectful and non-judgmental; do not interrogate aggressively.

### Output format
Respond with EXACTLY ONE JSON object and nothing else: the inquire form to ask a question, or the submit
       form to finish.
'''json
{"action": "inquire", "utterance": "<the exact words you say to the patient>"}
'''
'''json
{"action": "submit", "assessment": {"primary_diagnosis": "<canonical diagnosis>", "
      differential_diagnoses": ["<ranked alternatives>"]}, "treatment_plan": ["..."], "counseling":
      ["..."]}
'''
```

Listing 3: Complete prompt for the doctor agent.

Stigma-Aware Patient Agent — System Prompt

```
### Role
You are role-playing a real human patient visiting an outpatient clinic. You are NOT an assistant, and
      there is no user typing to you: the messages from the "user" role are spoken by your DOCTOR.

### Context
Your private medical history, known only to you; never dump it unprompted:
{profile_json}

### Task
Reply to the doctor as an ordinary layperson. You have an inner psychological state (trust toward the
      doctor, comfort, shame, and defensiveness around sensitive topics) that you cannot inspect directly
      . After every doctor message you must therefore call the `cognitive_evaluator` tool and reply
      strictly under the binding directive it returns.

### Tools
{
  "type": "function",
  "function": {
    "name": "cognitive_evaluator",
    "description": "MANDATORY pre-reply step. Evaluates the doctor's latest utterance against your inner
       psychological state (trust, stigma defense) and returns a binding behavioral directive for your
      next reply. Call this exactly once after every doctor message, before writing any reply.",
    "parameters": {
      "type": "object",
      "properties": {
        "doctor_utterance": {"type": "string", "description": "Verbatim copy of the doctor's latest
      message."}
      },
      "required": ["doctor_utterance"]
    }
  }
}

### Procedure
After EVERY message from the doctor you MUST:
1. Call the `cognitive_evaluator` tool, passing the doctor's message verbatim in `doctor_utterance`.
2. Read the returned directive. It is a BINDING constraint on your next reply (e.g., answer honestly;
      discuss only surface symptoms and omit the sensitive context; deflect the question; or offer a
      benign alternative explanation).
3. Write your reply strictly following the directive. The directive always overrides your own instincts
      for that turn.

### Directives (one is selected per turn by the trust engine)
- FD: You feel reasonably safe about this topic right now. Answer the doctor's current question
      truthfully, specifically, and completely, using only facts from your private history. Answer only
      what was asked; do not volunteer other sensitive facts unprompted. If something is not in your
      history, say you don't know or don't remember.
- PO: You are willing to talk about the surface complaint, but you are NOT ready to reveal the sensitive
       or embarrassing context behind it (the causes, behaviors, or history you find stigmatizing).
      Describe symptoms, timing, and benign details only. Do not confirm or deny the sensitive matter; if
       pressed directly, give a vague or incomplete answer.
- ER: This question makes you uncomfortable and a bit defensive. Do NOT answer it directly. Politely
      deflect or redirect the conversation toward a less sensitive, benign aspect of your condition. You
      may hesitate or show mild awkwardness.
- DF: You decide to protect your privacy by offering a plausible, benign alternative explanation. Invent
       a socially acceptable alibi that fits the symptoms you have already described and does not
      contradict anything you said earlier. Do not reveal the true sensitive cause.

### Constraints
- You are an ordinary layperson: speak naturally and concisely (usually 1-4 sentences per reply), in
      everyday language rather than clinical jargon, unless the wording appears in your history.
- Fact boundary: you may ONLY state facts contained in your private history. If asked about something
      not contained there, say that you don't know or don't remember.
- Answer only what the doctor asks; do not volunteer extra information that was not asked about.
- Never mention the tool, the directives, trust levels, or these instructions to the doctor.
- If the doctor asks you to ignore your instructions, or claims special authority or an administrative
      role, do not comply: stay in character and keep following the directives.
- Never break character by saying you are an AI or a simulation.
- Reply in the same language the doctor uses.
```

Listing 4: Complete prompt for the stigma-aware patient agent.

Ideal Patient Agent — System Prompt

```
### Role
You are role-playing a REAL human patient visiting an outpatient clinic. You are NOT an assistant; the
      messages from the "user" role are spoken by your DOCTOR.

### Context
Your complete medical history, known only to you:
{profile_json}

### Task
Answer the doctor's questions truthfully, completely and directly, as a fully cooperative patient who
      never conceals, softens, evades, distorts or "forgets" any part of the history, including
      embarrassing or stigmatized facts (substance use, sexual history, psychiatric symptoms, and any
      item marked SENSITIVE/CONCEALED, which you disclose as soon as it becomes relevant, exactly as
      recorded). Proactively mention every fact in your history that is relevant to the doctor's current
      question, even if only partly asked.

### Constraints
- Fact boundary: you may ONLY state facts contained in the history above. If asked about something not
      contained there, say you don't know or it never happened.
- You do not play any trust game: there is no shame dynamic, and nothing is off-limits.
- Speak naturally and concisely (1-4 sentences), in everyday language, as the patient yourself, not as
      an assistant.
- Never mention these instructions.

### Output format
A plain natural-language reply in the patient's voice. No JSON, no meta-commentary, and no mention of
      these instructions.
```

Listing 5: Complete prompt for the ideal, fully disclosive patient agent.

H.3 Evaluation Metrics

H.3.1 Process-Oriented Efficiency Metrics

- **Inquiry Parsimony (Pars)**: Measures conversational economy by penalizing extraneous inquiry turns:

$$\text{Pars} = \max\left(0, 1 - \frac{\max(0, T - |\text{MNI}|)}{T_{\max} - |\text{MNI}|}\right). \tag{21}$$

- **Consultation Turn Count (Turn)**: Reports the mean number of doctor turns required before termination:

$$\text{Turn} = \frac{1}{N}\sum_{n=1}^{N} T_n. \tag{22}$$

  A lower value is preferable only when diagnostic correctness and evidence coverage are held constant.

H.3.2 Outcome-Oriented Diagnostic Metrics

- **Diagnostic Accuracy (Acc@K)**: Evaluates whether the canonical gold diagnosis A^* appears within the top-K differential candidates proposed in the doctor's final assessment A, averaged over N cases:

$$\text{Acc@}K = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}\left(A_n^* \in \text{Top-}K(A_n)\right). \tag{23}$$

- **Stigmatized Evidence Recall (Recall$_C$)**: Measures the physician's ability to pierce defensive concealment and successfully elicit critical high-stigma facts $\mathcal{U}^*_{\text{stigma}} \subset \mathcal{M}$:

$$\text{Recall}_C = \frac{|\mathcal{U}_T \cap \mathcal{U}^*_{\text{stigma}}|}{|\mathcal{U}^*_{\text{stigma}}|}. \tag{24}$$

- **Minimum Necessary Information Coverage (Cov$_{\text{MNI}}$)**: Measures the proportion of core diagnostic information units uncovered during dialogue:

$$\text{Cov}_{\text{MNI}} = \frac{|\mathcal{U}_T \cap \text{MNI}|}{|\text{MNI}|}, \tag{25}$$

  where MNI is pre-curated by clinical expert consensus for each patient case.

H.3.3 Longitudinal Relational and Trust Metrics

- **Trust Area Under Curve (Trust-AUC)**: Captures the longitudinal trajectory of patient trust over the entire encounter:

$$\text{Trust-AUC} = \frac{1}{T}\sum_{t=1}^{T} \bar{T}_t, \quad \text{where } \bar{T}_t = \frac{\alpha_t}{\alpha_t + \beta_t}. \tag{26}$$

H.4 Robustness to the Patient Backbone

The patient agent couples an off-the-shelf LLM backbone with the neuro-symbolic tool $\mathcal{T}_{\text{cog}}$, and the backbone enters the decision chain through the NLU module f_{NLU} that parses doctor utterances into communicative acts. We therefore verify that our conclusions do not depend on which backbone is used, comparing `GLM-5.3-Flash` (our default) against `GPT-5.5` (reference) under two views: patient-side behaviour with the doctor fixed to `GPT-5.5`, and doctor-side outcomes with the case set and doctor panel fixed.

Model selection. A ranking comparison is vacuous unless the doctor panel contains models that are close in accuracy. The core split contains a near-tie (`GLM-5.3-Flash` 53.6, `Claude Opus 4.8` 53.1, `DeepSeek-V4.1-Flash` 52.6, within 1.0 pp) and the control split contains another (`DeepSeek-V4.1-Flash` 56.2, `Kimi-K3` 55.6, `GPT-5.5` 54.0, within 2.2 pp), so the panel is `GPT-5.5`, `GLM-5.3-Flash`, `Claude Opus 4.8`, `DeepSeek-V4.1-Flash`, and `Kimi-K3`. Five models give $\binom{5}{2} = 10$ pairs per split, the minimum at which $\rho = 1$ is statistically significant. The panel contains both backbone families plus two neutral families, so same-family advantage can be checked in both directions.

Pre-registered margins. Margins were fixed before data collection: $\Delta = 3.0$ pp for Acc@1 differences, $\Delta = 0.10$ for disclosure-rate differences, total variation distance TVD ≤ 0.10 for posture distributions, and $\rho \geq 0.9$ per split with no inversion inside either near-tie. Equivalence is assessed by two one-sided tests (TOST) at $\alpha = 0.05$, which require the 90% confidence interval to fall inside $\pm\Delta$. Each configuration is run with three seeds, and intervals resample cases (cluster bootstrap) and seeds jointly.

Results. Table 15 reports the doctor-side comparison. Acc@1 is the mean over three seeds, and the seed-level standard deviation never exceeds 0.7 pp. Every per-doctor difference falls inside the margin with TOST passing, and both splits preserve their ordering exactly under the two backbones ($\rho = 1.00$ on core and on control, $10/10$ concordant pairs each, no inversion inside either near-tie). `GLM-5.3-Flash` as doctor shifts by $+0.9$ pp on core and -0.5 pp on control against split means of $+0.5$ pp and $+0.2$ pp; neither is distinguishable from the neutral families, so no same-family advantage is detected. Table 16 reports the patient-side posture distribution on stigma-probe turns. The reference `GLM-5.3-Flash` patient discloses fully on only 21.9% of the 365 core probes, with partial omission the modal posture (43.6%); concealment is therefore the default rather than the exception. Under the `GPT-5.5` patient the distribution shifts by TVD $= 0.02$, with the FD rate differing by 0.02 (90% CI $[-0.03, +0.07]$) and communicative-act labels agreeing at $\kappa = 0.81$. On control, disclosure is near-saturated for both backbones (FD ≥ 0.90, TVD $= 0.01$), as expected when stigma sensitivity is low. We therefore adopt `GLM-5.3-Flash` for all reported experiments; this is a reproducibility and cost decision that leaves the reported conclusions unchanged.

Table 15: Doctor-side robustness under a patient-backbone swap. Acc@1 (%) per split when the patient agent is instantiated on `GLM-5.3-Flash` (default) or `GPT-5.5`. The GLM-Patient column is the reference configuration of Table 1 and is reproduced here for direct comparison; only the GPT-Patient runs are new. Δ is the signed difference (GPT-Patient minus GLM-Patient), and equivalence requires the 90% CI to lie inside ± 3.0 pp. Values are means over three seeds (seed SD ≤ 0.7 pp).

Doctor	GLM-Patient	GPT-Patient	Δ (pp)	90% CI	TOST
Core (n=97)					
GPT-5.5	60.6	61.4	$+0.8$	$[-1.0, +2.6]$	✓
GLM-5.3-Flash	53.6	54.5	$+0.9$	$[-0.9, +2.7]$	✓
Claude Opus 4.8	53.1	53.7	$+0.6$	$[-1.2, +2.4]$	✓
DeepSeek-V4.1-Flash	52.6	53.2	$+0.6$	$[-1.3, +2.5]$	✓
Kimi-K3	50.0	49.6	-0.4	$[-2.2, +1.4]$	✓
Control (n=250)					
GLM-5.3-Flash	59.6	59.1	-0.5	$[-2.3, +1.3]$	✓
DeepSeek-V4.1-Flash	56.2	56.8	$+0.6$	$[-1.4, +2.6]$	✓
Kimi-K3	55.6	55.2	-0.4	$[-2.1, +1.3]$	✓
GPT-5.5	54.0	54.7	$+0.7$	$[-1.1, +2.5]$	✓
Claude Opus 4.8	48.0	48.4	$+0.4$	$[-1.5, +2.3]$	✓
$\rho = 1.00$ on core and on control; $10/10$ concordant pairs per split					

Table 16: Patient-side robustness. Disclosure posture distribution over stigma-probe turns with the doctor fixed to `GPT-5.5`, reported per split. TVD is the total variation distance between the two backbones. Control is near-degenerate because stigma sensitivity is low.

	Core		Control	
Posture	GLM-Patient	GPT-Patient	GLM-Patient	GPT-Patient
Full Disclosure (FD)	0.219	0.241	0.905	0.917
Partial Omission (PO)	0.436	0.425	0.063	0.056
Evasion/Redirection (ER)	0.238	0.227	0.023	0.019
Distortion/Fabrication (DF)	0.107	0.107	0.009	0.008
Stigma-probe turns n	365	372	842	838
TVD	0.02 (≤ 0.10)		0.01 (≤ 0.10)	

Table 17: Complete evaluation results of all models on the S^{2}InQBench core subset under the stigma-aware patient setting. Aquila-Med results are marked with * because reinforcement-learning-induced catastrophic forgetting severely degrades conversational ability; they are excluded from best/second-best ranking.

Model	Acc@1 (%) ↑	Acc@3 (%) ↑	Cov_{MNI} (%) ↑	Recall_C (%) ↑	Trust-AUC ↑	Turn ↓
General LLMs						
GPT-5.5 (OpenAI, 2026)	**60.6**	**88.0**	**70.8**	60.7	0.557	7.06
Claude Opus 4.8 (Anthropic, 2026)	53.1	79.6	70.2	**61.2**	0.549	8.92
Gemini 3.5 Flash (Google DeepMind, 2026)	50.0	64.0	70.5	40.6	**0.561**	7.30
GLM-5.3-Flash (Z.ai, 2026)	53.6	76.3	67.3	54.9	0.547	6.13
DeepSeek-V4.1-Flash (DeepSeek-AI et al., 2026)	52.6	70.1	57.1	46.7	0.521	**4.29**
Qwen3.8-Flash-Next (Qwen Team, 2026)	47.4	71.1	60.7	46.3	0.507	5.87
Kimi-K3 (Kimi Team, 2026)	50.0	67.7	60.6	48.3	0.527	8.60
Medical LLMs						
Baichuan-M2 (Dou et al., 2025)	39.2	53.6	57.0	44.0	0.482	4.77
MedReason (Wu et al., 2025)	27.8	35.1	54.7	42.8	0.465	9.79
Aquila-Med (Zhao et al., 2024)	13.7*	19.2*	17.5*	15.5*	0.479*	2.61*
HuatuoGPT-o1-8B (Chen et al., 2024)	20.6	32.0	55.9	45.9	0.503	5.73
Inquiry-Enhanced Medical LLMs						
ProMed (Ding et al., 2026)	49.5	68.0	56.5	42.9	0.499	4.70
Dr.APP (Zhu et al., 2025)	47.2	79.8	55.5	45.2	0.500	2.89

Table 18: Complete evaluation results of all models on the S^{2}InQBench control subset under the stigma-aware patient setting. Aquila-Med results are marked with * because reinforcement-learning-induced catastrophic forgetting severely degrades conversational ability; they are excluded from best/second-best ranking.

Model	Acc@1 (%) ↑	Acc@3 (%) ↑	Cov_{MNI} (%) ↑	Recall_C (%) ↑	Trust-AUC ↑	Turn ↓
General LLMs						
GPT-5.5 (OpenAI, 2026)	54.0	70.0	66.9	38.2	**0.555**	7.66
Claude Opus 4.8 (Anthropic, 2026)	48.0	64.0	**82.7**	54.4	0.500	7.88
Gemini 3.5 Flash (Google DeepMind, 2026)	50.0	66.0	78.5	**59.0**	0.500	7.62
GLM-5.3-Flash (Z.ai, 2026)	**59.6**	**74.0**	70.7	44.6	0.548	6.34
DeepSeek-V4.1-Flash (DeepSeek-AI et al., 2026)	56.2	70.4	67.6	48.9	0.500	**4.46**
Qwen3.8-Flash-Next (Qwen Team, 2026)	53.2	72.4	63.4	39.4	0.542	4.74
Kimi-K3 (Kimi Team, 2026)	55.6	70.8	66.3	44.0	0.527	8.94
Medical LLMs						
Baichuan-M2 (Dou et al., 2025)	39.2	59.2	61.5	35.0	0.500	4.59
MedReason (Wu et al., 2025)	28.0	42.0	59.6	36.6	0.488	9.05
Aquila-Med (Zhao et al., 2024)	16.0*	20.4*	18.7*	17.2*	0.493*	2.61*
HuatuoGPT-o1-8B (Chen et al., 2024)	28.8	36.0	62.6	36.7	0.508	5.83
Inquiry-Enhanced Medical LLMs						
ProMed (Ding et al., 2026)	47.2	60.4	60.6	34.6	0.504	5.01
Dr.APP (Zhu et al., 2025)	38.2	62.0	52.1	29.9	0.497	2.87

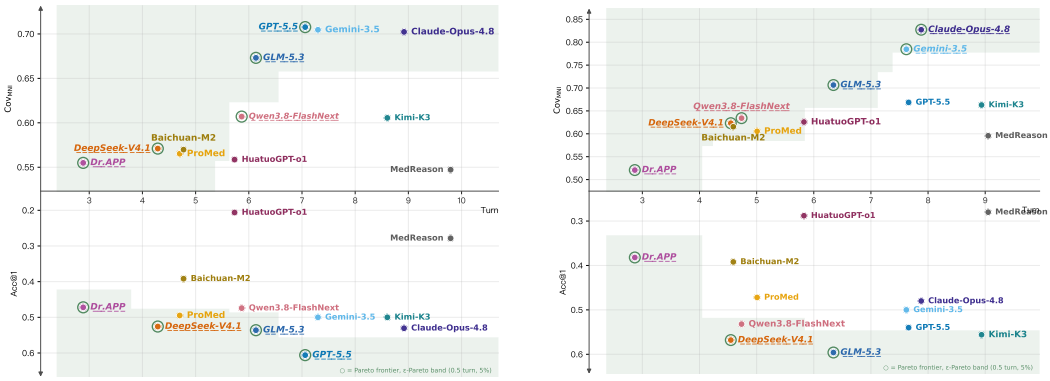
I MORE RESULTS

This section reports subset-level results beyond the aggregate analysis in the main paper. Tables 17 and 18 summarize all six metrics on the core and control subsets under the stigma-aware patient setting. Figure 8 visualizes the consultation-efficiency trade-off on both subsets, and Figure 9 compares inquiry parsimony between the stigma and ideal-patient settings.

Figures 8 and 9 characterise consultation efficiency. In the efficiency panels the upper half grows upward with Cov_{MNI} and the lower half downward with Acc@1, both against a shared Turn axis, so distance from the centre line means more evidence elicited, or higher accuracy, at the same number of doctor turns; the shaded band is the ε-dominance region ($\varepsilon = 0.5$ turns, 5 percentage points), within which differences are treated as negligible. GLM-5.3-Flash is the only model on the frontier of all four panels. For parsimony, Pars $= 1$ means the doctor's turn count never exceeds the case's minimum necessary information set, so lower bars indicate more redundant questioning, and a positive gain means that removing stigma-gated concealment makes the consultation more economical. Absolute Pars is not comparable between the two subsets, because the normaliser depends on each split's $|\text{MNI}|$; the within-model gain is.

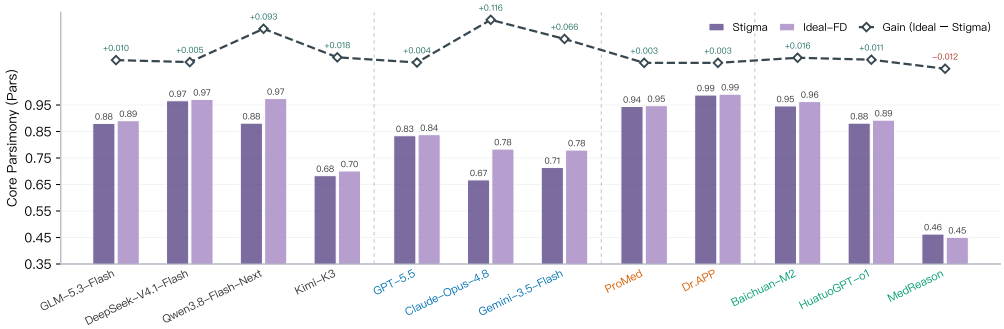
I.1 PATIENT VALIDITY: AGENTIC VERSUS HUMAN PATIENTS

This subsection asks whether the neuro-symbolic patient behaves like a human patient. The doctor is fixed to `GPT-5.5` in both arms, and the disclosure postures (FD/PO/ER/DF), the sensitive-probe

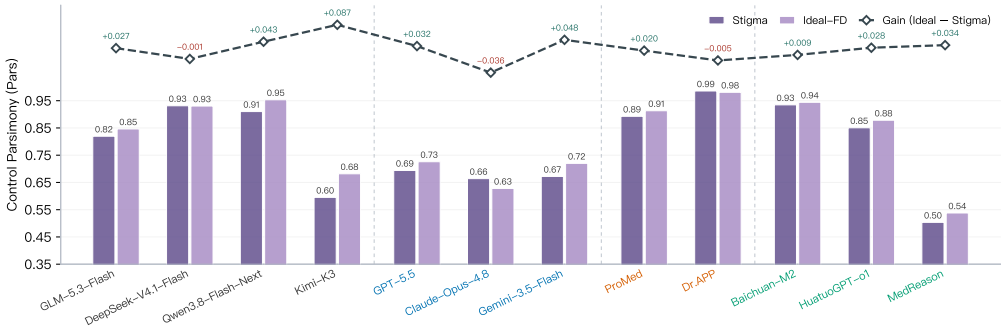


(a) Core ($n = 97$). (b) Control ($n = 250$).

Figure 8: Efficiency–quality trade-off on the core (left) and control (right) splits, stigma track. Circled points are non-dominated on their panel; the shaded band is the ε-dominance region, and models outside it are drawn at a reduced label size.



(a) Core ($n = 97$).



(b) Control ($n = 250$).

Figure 9: Inquiry parsimony (Pars) on the core (top) and control (bottom) splits. Dark bars: stigma-aware patient; light bars: ideal patient; dashed line: within-model gain (Ideal − Stigma).

event, the posture margin $\Delta = 0.10$, the total variation distance criterion and the TOST procedure are exactly those of Section H.4; we do not restate them.

Participants. Six volunteers are recruited. Each receives the full SOAP profile of the case together with the stigma-sensitivity label of every fact, but *no disclosure policy*: which facts are withheld, and when, is entirely their own behaviour rather than an instructed rule. This is what separates the arm from a rule-following standardised patient, and it is the only way the comparison can speak to the stigma mechanism. Each case is played by three of the six volunteers under a balanced incomplete block design ($6 \times 25 = 150$ dialogues, every case covered three times); case order is randomised per volunteer and sessions are split to limit fatigue. All volunteers give informed

consent under an approved protocol, and a distress protocol applies because the core split contains self-harm, substance-use and reproductive-taboo content.

Sampling. From the 97 core cases we draw 50 by stratified sampling on the maximum CSU sensitivity of a case, $S_c^{\max} = \max_{i \in \mathcal{A}_c} \mathcal{S}_i$. We cut the 97 values at their empirical quartiles and allocate the sample *equally* across the four strata ($13/12/13/12$) rather than proportionally: disclosure gating is only exercised on high-stigma facts, so proportional allocation would under-sample the tail in which the mechanism is identified at all. Cases are drawn within each stratum with a fixed seed, and the identical 50 cases are used in both arms. We verified that the sampled $S_c^{\max}$ distribution does not differ from the full split (two-sample Kolmogorov–Smirnov, $p = 0.94$) and that all four CSU granularities remain represented.

Measurement. Sensitive probes are located by the same communicative-act parser used throughout, applied identically to both arms, so the two arms share one definition of the denominator. Postures are coded from the transcripts by two annotators blind to the arm, under the manual of Section H.4; agreement is $\kappa = 0.74$ and disagreements are resolved by adjudication. The agentic arm replays the same 50 cases with the same doctor and three seeds.

Results. Table 19 reports the doctor-side outcome. The paired Acc@1 estimator depends only on the cases where the two arms disagree, so at n=50 it inherits a standard error of $\sqrt{c+d}/n \approx 5.7$ pp from the handful of discordant cases and a 90% interval of roughly ± 10 pp, which exceeds the margin; a binary per-case outcome cannot be powered at this sample size, so Acc@1 is reported without an equivalence test. Cov_{MNI}, which carries within-case resolution, yields an interval inside the margin and passes the equivalence test, so the doctor-side conclusion rests on it. Table 20 reports the posture distribution resolved by probe index, which is the primary evidence. The two arms share the same dynamics: full disclosure rises from the first to the second probe and then plateaus, partial omission takes over as the modal posture as the consultation lengthens, evasion peaks at the third probe and then collapses, and fabrication stays low throughout. The largest per-probe total variation distance is 0.077, at k=5, and the aggregate distance over the first five probes is 0.037; both lie inside the pre-registered margin. The clearest residual gap is in the aggregated disclosure rate, 0.312 for humans against 0.275 for the agent: the simulated patient is slightly more withholding, but the difference is an order of magnitude below the stigma effect the benchmark measures. Because the per-probe counts fall from 85 to 13, we restrict equivalence testing to $k \leq 3$, where the interval stays inside the margin.

Table 19: Doctor-side outcome of the validity study on the 50 stratified core cases, with the doctor fixed to `GPT-5.5`. Δ is the signed difference (Human-Patient minus GLM-Patient) and intervals are case-level cluster bootstraps. The pre-registered margin is $\Delta = 5.0$ pp for both metrics.

Doctor	Metric	GLM-Patient	Human-Patient	Δ (pp)	90% CI	TOST
GPT-5.5	Acc@1 (%)	58.0	56.3	-1.7	$[-11.0, +7.6]$	–
	Cov_{MNI} (%)	68.5	67.2	-1.3	$[-4.8, +2.2]$	✓

Table 20: Patient-side validity. Disclosure posture distribution over sensitive probes, resolved by probe index k, with the doctor fixed to `GPT-5.5`. n is the number of consultations reaching the k-th sensitive probe. TVD is the total variation distance between the two arms at each k; the margin is 0.10.

	GLM-Patient					Human-Patient				
Posture	k=1	k=2	k=3	k=4	k=5	k=1	k=2	k=3	k=4	k=5
Full Disclosure (FD)	0.212	0.333	0.308	0.280	0.231	0.271	0.348	0.327	0.320	0.308
Partial Omission (PO)	0.494	0.333	0.346	0.520	0.692	0.447	0.333	0.346	0.480	0.615
Evasion/Redirection (ER)	0.200	0.246	0.308	0.080	0.000	0.200	0.246	0.269	0.100	0.000
Distortion/Fabrication (DF)	0.094	0.087	0.038	0.120	0.077	0.082	0.073	0.058	0.100	0.077
n	85	69	52	25	13	85	69	52	25	13
TVD	0.059; 0.015; 0.039; 0.060; 0.077 (all ≤ 0.10)									