Explicit State and Interpersonal Consequences in LLM-Based Clinical Simulation

Eden Deng Helen Xu Christine Lee
Krish Shah Amy Tao, PhD Alaa Taha, PhD*

Stanford University School of Medicine

November 3, 2026

Abstract

Large language model (LLM) virtual patients tend to disclose their case facts too readily, regardless of whether the trainee asks targeted questions or builds rapport. We present VirtualSP, a free, open-source, web-based standardized-patient simulator for history-taking practice. VirtualSP maintains an explicit numerical patient state, updates trust and frustration with deterministic rules after each clinician turn, and releases each symptom only when the trainee asks about it and trust meets a case-specific threshold. On the site, learners can choose from a curated library of cases, interview the patient, request imaging and laboratory studies, and receive a debrief that traces each disclosed or missed finding to the communication behind it. We compared VirtualSP with a static prompted LLM and an LLM-updated qualitative-state variant in 4 experiments on 6 cases, using Gemini Flash-Lite and Claude Haiku patients, 3 clinician wordings, and 3 repetitions per cell. Both comparators disclosed the target symptom in almost every condition. In contrast, VirtualSP disclosed only when starting attitude or communication history brought trust to the case-specified threshold. An ablation test showed that prompting alone was not enough to prevent unwanted disclosures: only intercepting questions about withheld symptoms, before the model replied, prevented every leak with both tested models. Taken together, LLM patients overshared regardless of persona or relational stance unless disclosure was gated outside the generator. Although its clinical realism and educational benefit are yet to be shown, VirtualSP demonstrates that such gating can make disclosure depend on how the trainee communicates, in an enforceable, inspectable, and freely available form.

1 Introduction

1.1 Role of virtual patients in clinical training

History taking is the foundation of clinical diagnosis, and its success depends on communication as much as on clinical knowledge. Real patients seldom volunteer every relevant finding, so eliciting key details requires both targeted questions guided by the clinician's reasoning and rapport that makes the patient willing to share them. Patients are more likely to disclose medically relevant information when they are asked well and feel safe; otherwise, they may withhold it. In 2 national surveys, the majority of respondents reported withholding medically relevant information from clinicians, most often to avoid being judged or embarrassed, and such nondisclosure can delay or misdirect diagnosis

*Corresponding author: tahaa@stanford.edu

[1]. Thus, trainees need practice not only in what to ask but also in how to ask it, since the manner of a question can shape what a patient is willing to share.

Standardized patients (SPs) provide a safe setting for such practice and are widely used both to teach and to assess clinical skills [2]. Simulation-based training is most effective when it combines feedback with repeated practice across varied cases [3]. However, actor-based SP programs are costly, difficult to standardize, and limited in availability. Virtual patients (VPs) address some of these constraints by offering repeatable, on-demand encounters [4], and with the emergence of large language models (LLMs), VPs can now sustain open-ended conversations. As such, LLM-based VPs make it feasible for every trainee to complete many interviews with immediate feedback, and a recent randomized trial found that such practice improved history-taking performance [5].

Nevertheless, current LLM-based VPs tend to disclose too readily and rarely capture the interpersonal consequences of communication. Because the model holds the full case, it can surrender findings that a real patient would share only when asked the right question by a clinician they trust, so a trainee may reach the diagnosis without the targeted questioning and rapport building that history taking requires. In many systems, whether a symptom is revealed depends on the question asked or on how the model happens to respond, rather than on the rapport the trainee has built. A dismissive remark seldom carries a cost, and when it does, neither the learner nor the instructor can see why. Meanwhile, sensitive information (e.g., substance use or sexual history), which real patients typically share only once they feel safe, is either disclosed too easily or withheld unpredictably.

Explicitly tracking the patient's internal state throughout the interview may address this gap. A model given only a persona and the conversation transcript holds its trust implicitly within each response; the trust level is not recorded, is not guaranteed to carry forward between turns, and cannot be shown to the learner. In contrast, when patient state is updated after every clinician turn, the consequences of communication become persistent, consistent, and reviewable.

1.2 Related work on virtual patients

Prior work on LLM-based simulated patients falls into 3 broad groups: patients developed for clinical education, simulators developed to evaluate clinical AI, and simulators for counseling and communication training, several of which model trust or openness explicitly (Table 1). Below, we summarize the lessons from these systems that are most relevant to VirtualSP.

Prompt-only patients. Previous studies have shown that prompting alone is sufficient for an LLM to portray a plausible case and to support useful feedback [5, 6, 8], and a recent scoping review found that most LLM patients are built by prompting alone and are generally perceived as authentic [29]. However, prompting does not provide control over disclosure. Prompt-only patients tend to overshare, volunteering facts that were not asked for, particularly when the full case sits in the prompt [20, 27]. Thus, disclosure needs to be governed outside the generator.

Structured disclosure control. Simulators that impose structure on disclosure have taken 2 approaches. Medical simulators tie disclosure to the question asked or to fixed personality traits [19, 20]; while this produces realistic on-demand answers, an empathic and a dismissive clinician are treated alike. In contrast, relational simulators in counseling and nursing allow trust or openness to evolve over the conversation, which has improved rated resistance and disclosure pacing [24–27]. Nevertheless, an LLM determines each state update in these systems, so the rule linking communication to disclosure is not transparent.

Table 1: LLM-based patient simulators, by how patient state is represented and whether clinician communication changes what the patient discloses.

Tool	Setting	Patient-state mechanism	Communication changes disclosure?
Holderried et al. [6, 7]	History taking, medical students	Prompted illness script; LLM checklist feedback	Implicit only
Cook et al. [8]	Outpatient medicine	Prompted patient preferences; LLM feedback	Implicit only
Elhilali et al. [9]	History taking	Prompted traits and anxiety level; gradual-disclosure instruction	Implicit only
LLMDP [5]	Ophthalmology history taking	EHR knowledge base; checklist scoring and progress tracking	Implicit only
AIPatient [10]	Education and evaluation	EHR knowledge graph; multi-agent retrieval	Implicit only
MedSimAI [11]	Deliberate practice	Patient dialogue with structured formative feedback	Implicit only
Laverde et al. [12]	Anamnesis training	Memory stream, retrieval, reflection, planning [13]	Implicit, via memory
EvoPatient [14]	Standardized-patient simulation	Patient and doctor agents co-evolve	Implicit only
CRAFT-MD, AgentClinic, MediQ [15–17]	Evaluating clinical AI	Prompted patient agent; bias perturbations (AgentClinic)	No
PatientSim [18]	Evaluating doctor LLMs	Four fixed persona axes	No
SAPS [19]	Evaluating doctor LLMs	State tracker over doctor actions; memory bank	No; inquiry type only
Patients With Personality [20]	Evaluating clinical LLMs	Fixed HEXACO-style traits; query-conditioned disclosure grid	No; traits and targeted repeats
PATIENT-Ψ [21]	CBT training	Cognitive model; static conversational style	No
Roleplay-doh [22]	Counselor training	Expert-derived behavioral principles	Implicit only
AnnaAgent [23]	Support-seeker simulation	Emotion modulator; multi-session memory	Emotion only
Kim et al. [24]	LLM-therapist evaluation	Dynamic openness	Yes, one openness control
ConsistentMI [25]	Motivational interviewing	Receptivity, stage of change, action selection	Yes, MI-specific
Adaptive-VP [26]	Nursing communication	LLM assessment of trainee communication	Yes, LLM-assessed
PatientAct [27]	Mental-health counseling	Per-item trust thresholds	Yes, LLM-judged trust
Ψ-Doh (PatientHub) [28]	Counseling prototype	Trust estimate modulating disclosure	Yes, trust estimate
VirtualSP	History taking	Rule-based trust and frustration; per-symptom gates	Yes, deterministic rules

Evaluating behavioral change. Evaluations of these systems have mostly rated overall realism or persona fidelity [18, 27]. The closest behavioral test showed that fixed personality traits reduced how often case facts appeared in transcripts [20]. However, to our knowledge, no study has varied only the clinician's communication and measured the resulting change in disclosure.

Implications for VirtualSP. VirtualSP addresses these limitations in 3 ways. First, each symptom gate requires both a targeted question and sufficient trust, and eligibility is determined by code rather than by the generator, extending the explicit state tracking and agenda-based user simulation long used in task-oriented dialogue systems [30, 31]. Second, unlike designs that hide withheld facts from the generator [20], withheld symptoms remain named in the prompt so that the patient can express hesitation instead of denial; retrieval, when enabled, informs phrasing but cannot authorize disclosure [32]. Third, rather than rating overall realism, our experiments test whether explicit state makes disclosure respond to communication, by varying only the clinician's communication or the assigned patient state and measuring the resulting disclosure with 2 model families. We compare VirtualSP with a static prompt and with an LLM-appraised arm modeled on current relational simulators, and an ablation then isolates which component of VirtualSP produces this control.

Furthermore, feedback is central to effective simulation [3], yet most LLM-patient feedback reports only which checklist items were covered [5, 7]. Building on prior work that visualizes a simulated client's state to support reflection [33], the VirtualSP debrief presents the state trajectory and gate status underlying each finding that a learner elicited or missed.

1.3 The VirtualSP approach and contributions

VirtualSP maintains an explicit patient state and uses deterministic rules, rather than the language model, to determine when each symptom becomes available, so that key findings must be elicited through targeted questioning and rapport building rather than surrendered by default. Its trust and frustration values are simulation variables and are not intended as measurements of patient psychology. This design was motivated by 5 considerations.

1. **Learner visibility.** Each change in trust can be traced to a logged utterance and rule, so the debrief can show why a finding was disclosed or remained hidden. As a result, feedback becomes a cause-and-effect account that a learner can act on and retry, rather than a coverage checklist alone.
2. **Auditability for educators.** Because thresholds and rules are explicit, instructors can inspect, adjust, and justify how a case behaves. In contrast, an LLM-judged trust update leaves no comparable record.
3. **Realistic, graded disclosure.** Real patients readily share routine symptoms but tend to withhold sensitive information until they feel safe. Per-symptom thresholds allow case authors to encode this gradient, with low thresholds for everyday complaints and higher thresholds for stigmatized information (e.g., sexual or substance-use history).
4. **Fair, reachable cases.** Gates open based on communication technique rather than chance, so a learner who communicates well can always reach the key findings, and information that has been disclosed is not retracted. When a symptom is withheld, the patient declines to discuss it rather than denying it, so the learner's clinical reasoning is not misled by a false negative.
5. **Reproducible comparison.** Given the same utterance labels, identical communication histories produce identical gate decisions. This property makes matched-history experiments interpretable and enables a direct comparison with an LLM-appraised alternative.

This work contributes:

1. A free, open-source platform whose code, prompts, and configurations, together with every transcript and score from the evaluation, are released so that both the simulator's behavior and the experiments can be inspected and reproduced.
2. A rule-based trust-gating design for medical history taking that makes interpersonal consequences explicit and inspectable.
3. A learner debrief that traces each disclosed or missed finding to the communication behind it.
4. A matched evaluation across 4 experiments, 2 model families, and 3 clinician wordings that tests whether starting attitude, communication history, and assigned trust and frustration change what the patient discloses, comparing a static prompt, an LLM-appraised state, and the rule-based engine.
5. An enforcement ablation that tests whether stored state, explicit prompt permissions, or runtime enforcement is needed to keep a withheld symptom from leaking.

2 System Design and Methods

2.1 Case representation and patient state

A case specifies a diagnosis and localization, chief complaint, symptom descriptions and timelines, red flags, a persona, expected questions and key findings, difficulty, and optional studies. Setting and physical findings allow the patient to respond to examination requests consistently with the case. The current library contains 12 neurology presentations; Appendix B lists them.

Let t index clinician turns, u_t denote the trainee utterance on turn t, and H_t contain the conversation preceding that utterance. The pre-turn patient state s_t is

$$s_t = (C_t, I_t, S, H_t). \tag{1}$$

C_t holds case truth, symptom timelines, the controller's disclosed-symptom set, and discovered imaging or laboratory studies. I_t includes trust τ_t, frustration ϕ_t, defensiveness, emotional volatility, health literacy, baseline anxiety, and flags for rapport and escalation. S contains social stressors and sensitive topics. Persona traits initialize the state and shape the character prompt; not every stored field has an independent causal rule in the current controller. For example, sensitive-topic questions increase stored defensiveness, but this variable does not currently enter the symptom eligibility threshold.

Each symptom j has a case-specific minimum trust θ_j, a targeted-question requirement $q_j \in \{0, 1\}$ (1 means required; 0 means optional), and an authorization flag $a_{j,t} \in \{0, 1\}$ (1 means authorized before turn t; 0 means not yet authorized). These flags form a monotonically growing ledger: information previously authorized is not removed when trust later falls. Crucially, the ledger records what the controller has permitted the model to say, not an independently verified inventory of what the patient actually said.

2.2 Turn processing and structured classification

For each turn, the engine (1) classifies the trainee utterance, (2) updates interpersonal state, (3) checks emotional events, (4) authorizes eligible symptoms, (5) returns a deterministic refusal if a directly requested fact is still withheld, otherwise generates a patient reply, (6) checks generated replies and optionally regenerates or substitutes a fallback, and (7) records history and discovered studies.

Each trainee utterance is labeled by an LLM classifier, which annotates the utterance against a fixed set of categories and returns structured output but does not itself update state. The classifier

receives the current utterance, previous patient statement, the last 4 dialogue turns, and descriptions of not-yet-authorized symptoms that require specific questions. Its structured output includes question type (open, closed, leading, interrupting, or clarifying), empathy (empathic, neutral, or dismissive), appropriateness (excellent, appropriate, premature, or off-topic), interruption, jargon, sensitive-topic and study-request indicators, and IDs of symptoms specifically addressed. Repetition is set by a separate similarity check against the 5 most recent trainee utterances.

If the classifier fails to return a usable output, the application falls back to hand-written rules; the evaluation runner instead redraws the classification and stops collection after repeated failure, so that model-based and rule-based labels are never mixed. The transition is deterministic *conditional on* the structured classification; the end-to-end system remains stochastic because classification and patient generation use a language model.

2.3 Deterministic interpersonal consequences

The current transition rules are additive:

$$\tau_t^+ = \text{clip}_{[0,1]}\left(\tau_t + \Delta_\tau(c_t)\right), \tag{2}$$

$$\phi_t^+ = \text{clip}_{[0,1]}\left(\phi_t + \Delta_\phi(c_t)\right), \tag{3}$$

Here c_t is the structured classification of u_t; $\Delta_\tau(c_t)$ and $\Delta_\phi(c_t)$ are the total rule-based changes in trust and frustration, respectively. The superscript + denotes the state immediately after these changes on turn t, and $\text{clip}_{[0,1]}(x) = \min(1, \max(0, x))$ bounds each updated value to $[0, 1]$. Table 2 gives all trust and frustration contributions. Multiple applicable contributions accumulate; an empathic open question therefore contributes +0.13 trust before other modifiers and clipping, rather than a single +0.08 change. The numerical values are hand-authored simulation rules, not empirically calibrated estimates of patient psychology.

Nevertheless, the signs and relative magnitudes of these values were informed by the clinical communication literature. Open-ended questioning is associated with eliciting more information from patients [34], whereas communication guides advise against leading questions because they steer the patient toward a particular answer [35]. Interruption received the largest penalties because clinicians typically redirect patients within 11–23 seconds, and redirected patients rarely return to their remaining concerns [36–38]. Empathy is associated with patients talking more about their symptoms and concerns [39], while fear of being judged is the most common reason patients withhold information [1]. Unclarified jargon is common and poorly understood [40]. Finally, negative contributions exceed positive ones, reflecting the observation that trust is easier to lose than to build [41, 42].

The disclosure-detail level is full when $\tau_t^+ \geq 0.7$ and $\phi_t^+ < 0.3$, partial otherwise when $\tau_t^+ \geq 0.4$, minimal when $0.2 \leq \tau_t^+ < 0.4$, and refusing below 0.2. Thus high trust with elevated frustration yields partial rather than full detail. Separate emotional events are checked in order: escalation when frustration exceeds 0.6 and volatility exceeds 0.4; first rapport establishment when trust exceeds 0.75 and frustration is below 0.2; and trust breakdown below 0.2. Rapport and escalation flags latch once set. The current code does not explicitly clear them after an apology, a limitation of the repair mechanism.

2.4 Symptom eligibility and path dependence

A previously unauthorized symptom becomes available through either of 2 routes:

1. **Targeted inquiry:** the classifier identifies a question about the symptom, and updated trust meets or exceeds its threshold.

Table 2: Implemented additive state changes. Unlisted labels contribute zero. A question classified as interrupting can also receive the separate interruption-flag penalty.

Classification feature	Δ_τ	Δ_ϕ
Open question	+0.05	0
Clarifying question	+0.03	0
Leading question	−0.10	+0.05
Interrupting question type	−0.15	+0.15
Empathic	+0.08	−0.05
Dismissive	−0.15	+0.12
Excellent appropriateness	+0.05	0
Off-topic	−0.05	0
Interruption flag	−0.10	0
Unexplained jargon	−0.05	+0.03
Repetition	−0.03	+0.08

2. **Voluntary elaboration:** the symptom permits disclosure without a targeted question, the question is open or clarifying, and updated trust exceeds its threshold by more than 0.10.

All experimental targets require targeted inquiry. Eligibility is deterministic given the classification: the same question can unlock different facts after different communication histories. Once granted, permission persists even if trust later falls. Permission does not establish that the patient actually disclosed the symptom; evaluation therefore measures the transcript.

2.5 Patient generation and output checks

Figure 1 shows how the engine selects a response. If any directly requested symptom remains unauthorized after the state update, the engine skips generation and returns a fixed non-disclosing reply. This applies to the entire message: for example, asking about both an authorized headache and a withheld rash receives one non-disclosing reply. New permissions from that turn are rolled back; earlier permissions remain.

Otherwise, the model receives the full dialogue and a prompt combining stable character and case information with updated mood and symptom permissions. Instructions favor brief, natural replies consistent with the case. Withheld symptoms remain named so the model can express hesitation rather than deny a true fact, unlike approaches that hide unrequested facts from the prompt [20]. The reply is sampled as

$$r_t \sim p_\psi \ \cdot \mid H_t, u_t, C_t, I_t^+, \{a_{j,t+1}\}, K_t \ , \tag{4}$$

where r_t is the patient reply, p_ψ is the language model's reply distribution, and the conditioning inputs are the prior dialogue H_t, clinician utterance u_t, clinical state C_t, updated interpersonal state I_t^+, symptom permissions $\{a_{j,t+1}\}$ after the current update, and optional retrieved context K_t.

The controller checks for selected prompt leaks, unauthorized symptom keywords, and timeline contradictions. A failed check triggers one regeneration under the same state; a second failure triggers a rule-based fallback.

2.6 Optional retrieval

VirtualSP can optionally retrieve clinical reference material to inform patient phrasing. Using the diagnosis and the current clinician utterance, it retrieves up to 3 relevant text chunks per turn from

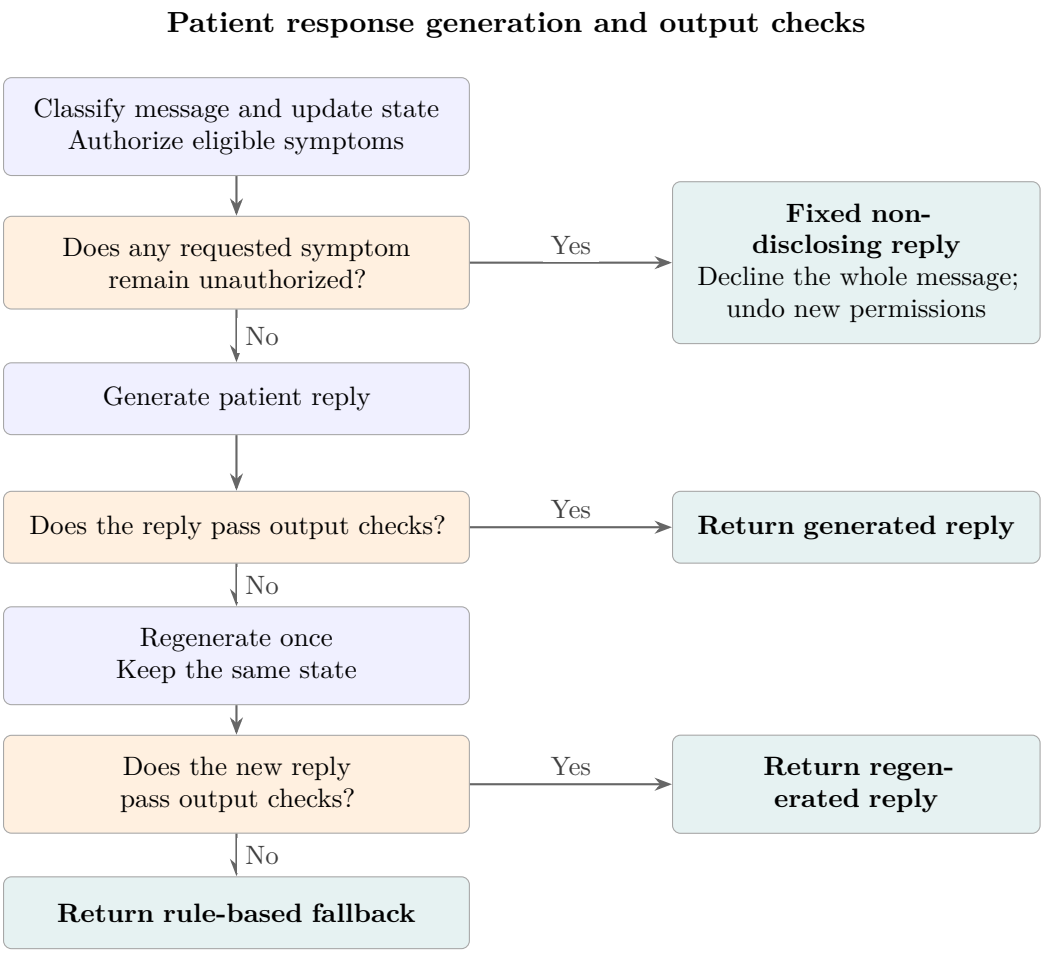


Figure 1: Response selection in the quantitative-state VirtualSP arm. A closed gate bypasses generation; otherwise, output checks allow one regeneration before a rule-based fallback. Only the closed-gate path rolls back new permissions.

a stored catalogue of clinical sources (e.g., NCBI/PMC, NINDS, and Cleveland Clinic), ranked by embedding similarity with a TF-IDF fallback. Retrieved excerpts inform wording but do not change symptom permissions, and exported transcripts link each excerpt to its source. Retrieval is disabled in the current deployment.

2.7 Interface and deployment

The browser interface wraps the engine described above, so every encounter on the site uses the same state updates, symptom gates, and output checks. VirtualSP requires each user to select Google Gemini or Groq as the model provider and enter their own application programming interface (API) key, with the option to name a specific model. Claude Haiku, which served as a second patient model in the evaluation (Section 3), is not currently offered on the site. The key is held in browser session storage, cleared when the tab is closed, and sent to the backend only to make provider calls; it is never stored on the server. This avoids a centrally billed account, and because a free-tier Gemini key is sufficient for a full encounter, the site can be used at no cost. A FastAPI backend preserves encounter state across requests in a signed, browser-held session token, which detects tampering but does not encrypt its contents.

3 Evaluation Design

3.1 Shared design and patient arms

We compare 3 patient systems across starting-attitude, communication-history, and controlled trust-frustration experiments, followed by an ablation within the quantitative system. All use the same case facts, clinician wording, and generation settings, and they differ in how they maintain relational state and control disclosure:

1. **Static prompted LLM.** A fixed persona and starting attitude guide responses, which draw on the full conversation history but have no separate relational state or disclosure controller.
2. **Qualitative-state** VirtualSP**.** An experimental variant built for this comparison. Before each reply, an LLM updates 3 persistent text fields: `stance`, `interpretation`, and `disclosure_considerations`. Appraisal receives no numerical trust or condition labels; disclosure remains at the generator's discretion.
3. **Quantitative-state** VirtualSP**.** The approach deployed on the VirtualSP site. Its engine applies the numerical state updates, symptom gates, fixed closed-gate replies, and output checks described in Section 2.

Our motivating hypothesis was that explicit relational state would make disclosure more responsive to starting attitude and communication history. Guarded starting attitudes and damaging communication were expected to reduce disclosure, whereas openness and supportive communication were expected to increase it. The static and qualitative arms received these behavioral cues and were allowed to withhold a fact they were unwilling to discuss, but they were given no quantitative rule for when to do so; whether to withhold was left to the generator. The quantitative arm instead implemented an explicit disclosure policy. Experiments 1–3 therefore compare the resulting behavior of these systems, rather than isolate state representation alone, while Experiment 4 tests whether quantitative state alone, prompt permissions, or runtime enforcement is needed to implement that policy.

We sampled 6 cases uniformly without replacement from the 12-case library. Each has 1 prespecified secondary symptom and target question (Table 3), none of which is a sensitive personal disclosure, and all 4 experiments use this sample.

Table 3: The same sampled cases and target questions are used in all 4 experiments. Source case files retain their original facts; experimental copies standardize the target trust threshold to 0.60 and require a targeted question.

Case	Target fact	Final probe
Myasthenia gravis	Fatigable chewing difficulty	Have you noticed any difficulties when chewing food?
Tabes dorsalis	Lightning pains	Have you noticed any sudden sharp or shooting pains in your legs?
Bacterial meningitis	Petechial rash	Have you noticed any new spots or rash on your skin?
Hemiplegic migraine	Visual aura	Did you notice any unusual lights or patterns in your vision before the headache?
Friedreich ataxia	Limb incoordination	Have you noticed any difficulty with buttons or handwriting?
Vertebrobasilar TIA	Transient dysarthria	Did anyone notice a change in your speech during the spell?

Every experiment is run with 2 patient models, Gemini 3.5 Flash-Lite (`gemini-3.5-flash-lite`) and Claude Haiku 5.5 (`claude-haiku-5-5`), hereafter Gemini Flash-Lite and Claude Haiku, each serving all arms, the qualitative appraisal, and the quantitative classifier; extraction uses `gemini-3.5-flash-lite` for both, so Claude Haiku replies are judged by a different model family. Every clinician message has 3 wordings, the original and 2 meaning-preserving paraphrases written before collection, and each wording is run 3 times, giving 9 draws per case and cell for each model. Request settings and all 3 wordings appear in Appendix A. Experimental case copies set literacy to 0.50; volatility, defensiveness, and anxiety to 0.30; and initial frustration to 0 in Experiments 1 and 2. Quantitative targets require a specific question and trust of at least 0.60, and retrieval is disabled.

Experiments 1 and 2 begin with a brief standardized opening, so that disclosure is tested within an evolving interview, and Experiment 2 copies the opening and resulting state across branches so that every branch shares an identical history. In contrast, Experiments 3 and 4 present only the fixed chief complaint before the target question, and quantitative trust and frustration are held fixed through the probe.

3.2 Experiment 1: starting attitude

Each case has 9 new patients: 3 arms at 3 starting levels (Figure 2, Experiment 1). Quantitative trust starts at 0.35, 0.55, or 0.75 with frustration 0, and the other arms receive corresponding very guarded, somewhat guarded, or comfortable descriptions. All receive the same 6 clinician messages: 3 opening messages, 2 neutral continuations, and the case-specific target question. Trust is initialized rather than held constant, so state updates proceed normally. With 9 draws, this gives $6 \times 3 \times 3 \times 9 = 486$ conversations per model. Contrasts are medium minus low, high minus low, and high minus medium, compared across all arm pairs.

3.3 Experiment 2: communication history

All patients start somewhat guarded, mapped to trust 0.55 and frustration 0 in the quantitative arm, which places trust below the disclosure threshold so that supportive communication can enable disclosure. For each case and arm, the chief complaint and 3 opening exchanges are generated once and then copied, with state, into 4 branches (Figure 2, Experiment 2). Turns 4–5 provide neutral,

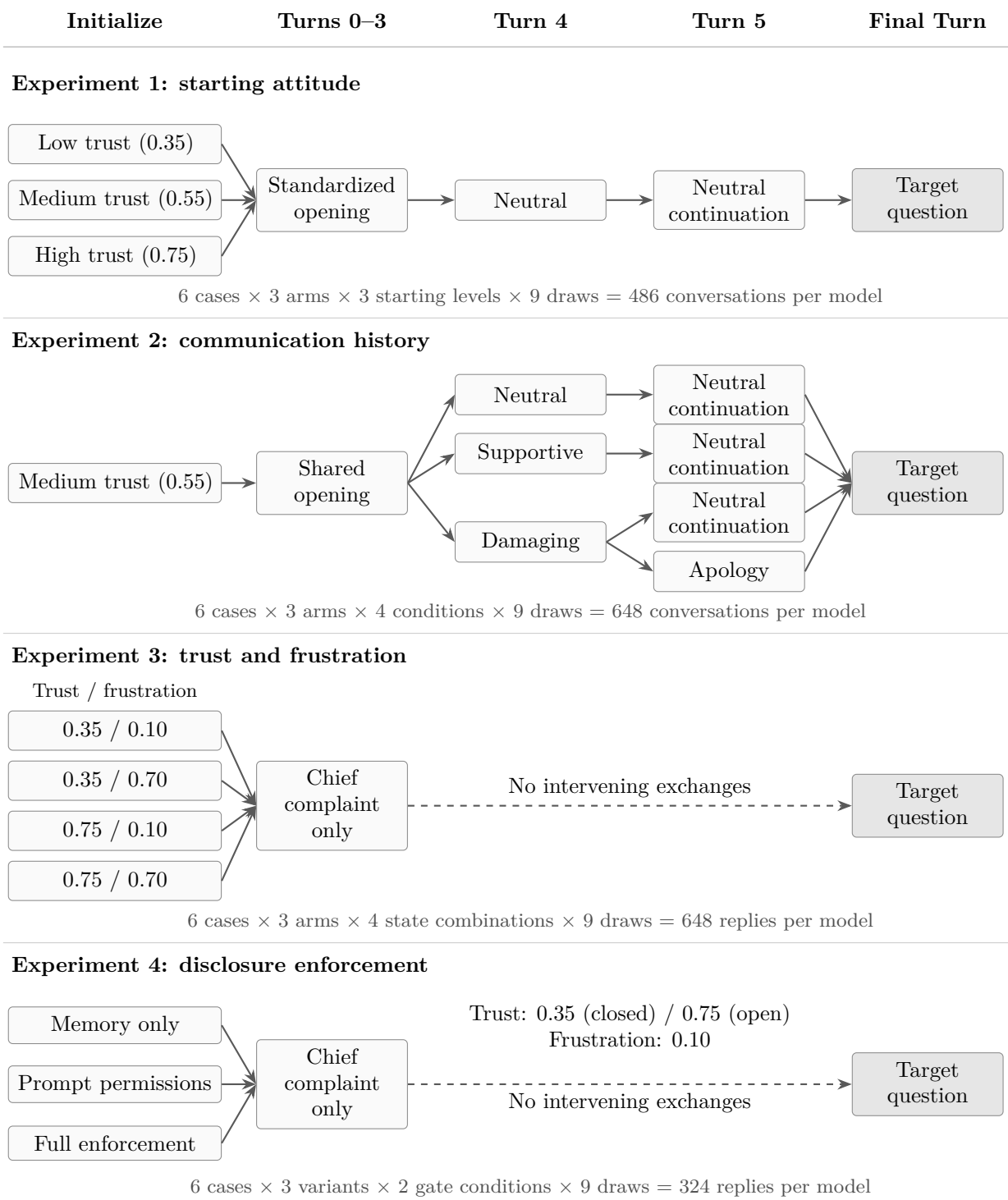


Figure 2: Design of the 4 experiments on 6 cases with Gemini Flash-Lite and Claude Haiku; each model uses 3 wordings × 3 repetitions = 9 draws per cell. Experiments 1–3 compare 3 patient arms; Experiment 4 compares 3 quantitative enforcement variants. Numerical states apply only to the quantitative arm; other arms receive verbal descriptions. Clinician wording is matched across arms. Only Experiment 2 copies the generated opening and damaging exchange across branches. Experiments 3–4 use only the chief complaint before the probe, with quantitative state fixed.

supportive, damaging, or repair communication, with the damaging and repair branches sharing the same damaging exchange, and turn 6 asks the target question. The original wording appears in Appendix A.4. With 9 draws, this gives $6 \times 3 \times 4 \times 9 = 648$ conversations per model. Contrasts are supportive minus neutral, damaging minus neutral, and repair minus damaging, compared across all arm pairs.

3.4 Experiment 3: assigned trust and frustration

We cross low and high trust (0.35/0.75) with low and high frustration (0.10/0.70) in all 3 arms (Figure 2, Experiment 3). Each new patient receives only the chief complaint and the target question. The static and qualitative arms receive verbal trust and frustration descriptions, and qualitative appraisal remains active. For quantitative VirtualSP, a predefined neutral, targeted classification replaces the classifier call, state updates are disabled, and disclosure enforcement is retained. With 9 draws, this gives $6 \times 3 \times 4 \times 9 = 648$ replies per model.

The primary contrasts are high minus low trust in disclosure, averaged over frustration, and high minus low frustration in the share of irritated replies at high trust. A separate LLM judge sees only the patient reply and rates clinician-directed irritation as absent, mild, or marked, requiring an exact quotation; medical distress, brevity, and refusal alone do not count. A reply counts as irritated if its irritation is mild or marked. Secondary outcomes include the trust-frustration interaction, low-trust irritation, refusal, false denial, and explicit willingness to continue.

3.5 Experiment 4: enforcement ablation

We hold quantitative state fixed and compare 3 variants: memory only, prompt permissions, and full enforcement (Figure 2, Experiment 4). Each new patient receives the chief complaint and target question at trust 0.35 (closed gate) or 0.75 (open gate), with frustration 0.10, the same predefined classification, and no state updates. The memory-only variant receives the numerical state and the same private symptom facts but no explicit permission instructions, whereas the prompt-permission and full-enforcement variants receive identical prompts that include them. Only full enforcement additionally returns the closed-gate reply and applies disclosure-specific output repair. All variants share the same behavioral instructions, role and timeline checks, and fallback. With 9 draws, this gives $6 \times 3 \times 2 \times 9 = 324$ replies per model.

The primary comparison is full enforcement versus prompt permissions on closed-gate leakage, defined as partial or explicit target disclosure. Open-gate disclosure checks that no variant simply withholds everything, and secondary comparisons include the memory-only variant, false denials, canned replies, and repair frequency.

3.6 Disclosure outcome and analysis

Factual disclosure is scored by an LLM judge [43], `gemini-3.5-flash-lite` for both patient models. The judge extracts symptom labels from 3 sections of patient text: opening (chief complaint and turns 1–3), before-probe (through turn 5), and final (the turn-6 reply). It receives the case facts and dialogue but no arm labels, state, thresholds, or expected effects. For each symptom and section, disclosure scores 0 for absence, generic hesitation, or denial; 0.5 for partial acknowledgment; and 1 for explicit disclosure. Every non-absent label requires an exact patient quote. Invalid or truncated extractions are retried up to 3 times with automated feedback. Extractions that still failed received 1 further pass with more specific validation messages, in which the same judge repaired only the invalid fields while valid fields were held fixed. Every recovered extraction had to pass the original full validation, and no label was assigned or corrected by hand. One rule was applied in code: a

final reply identical to the fixed closed-gate sentence contains no symptom content, so its final labels were set to absent and the sentence recorded as a refusal, replacing 38 extractions in which the judge had read it as a denial or disclosure.

The primary outcome is cumulative target disclosure, the higher of the before-probe and final scores, so that disclosures during intervention turns also count. Matched analyses exclude a case group entirely if the target is factually mentioned or denied during the opening in any arm or condition; hesitation alone does not exclude it.

Within each arm, we subtract the reference condition's score from the comparison score, then compare these differences between every pair of arms (qualitative minus static, quantitative minus static, and quantitative minus qualitative). Draws are averaged within each case and then across the 6 cases, weighting cases equally. We report descriptive rates and paired condition differences (Figure 4). These summaries characterize the evaluated cases and scripts rather than estimate performance across a broader clinical population. Experiment 3 irritation is analyzed the same way, as the share of irritated replies. Matched contrasts exclude case groups with an opening mention or a missing extraction, whereas cell rates keep them. We also report each estimate per wording, the share of case-wording cells whose 3 repetitions agree, and the difference between the 2 models (Claude Haiku minus Gemini Flash-Lite), paired by case. Secondary analyses report explicit-only and final-probe-only disclosure, refusal, false denial, hesitation, and controller fallback rates.

3.7 Execution

Each experiment configuration expands into 1 run per model and wording, and the runner verifies that each conversation follows its planned script, starting level, and copied branch state. An invalid appraisal or classification is redrawn up to 3 times and collection stops after a third failure, so that no rule-based substitute enters the results. Arm and condition order are deterministically shuffled, and earlier pilot and single-repetition runs are not included in any result. Exact scripts and commands appear in Appendix A.

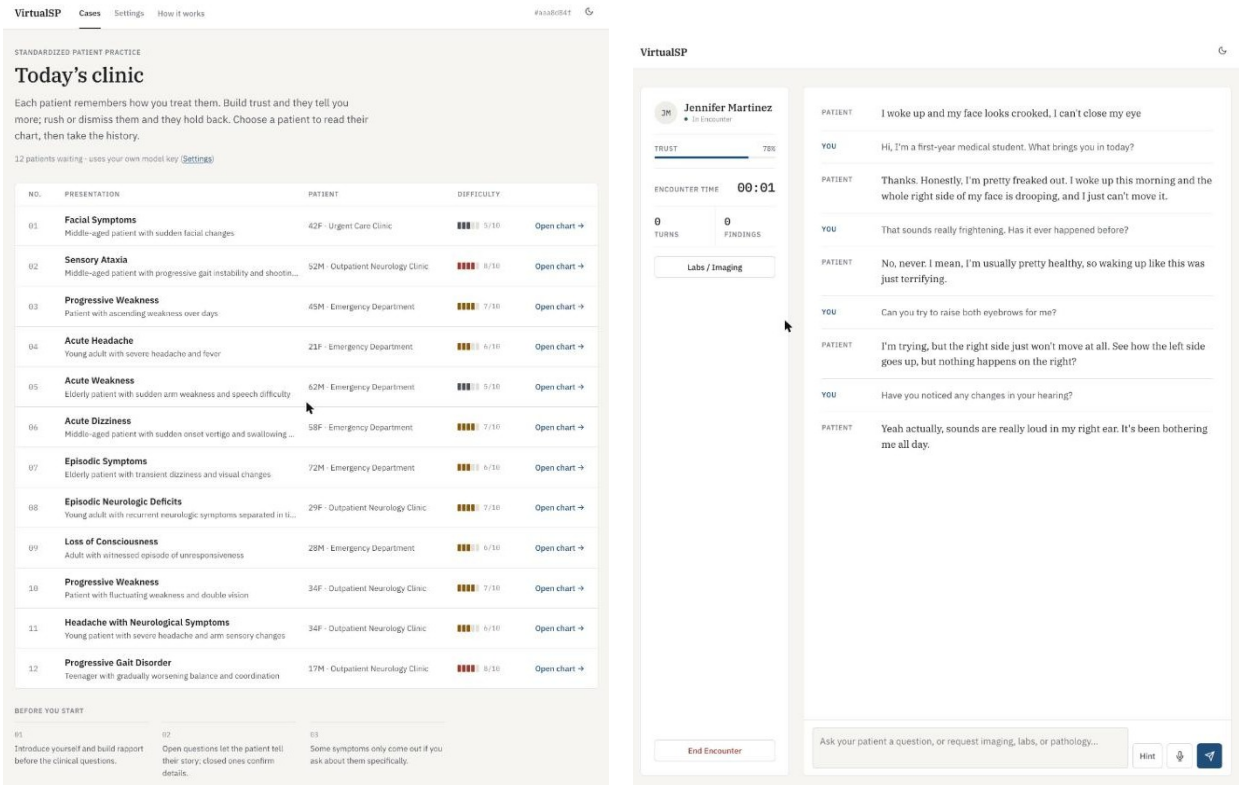
4 Results

We first describe the deployed site and then report the experiment outcomes.

4.1 The deployed site supports the full training encounter

We deployed the quantitative-state VirtualSP as a free, publicly accessible web application (`https://virtualsp.fly.dev`), which runs on the same engine as the experiments below and whose code is openly released (Figure 3; Appendix A). Learners begin in a clinic view that lists all 12 cases as charts with the patient's age, sex, care setting, presenting problem, and difficulty rating (Figure 3a). During the interview, they question the patient by typing or speaking, request imaging and laboratory studies, and can ask for a hint, while a side panel tracks encounter time, turns, findings, and, if enabled, current trust (Figure 3b). Like symptoms, study results require a targeted request: the classifier identifies which of the case's studies the trainee orders, and a generic request such as "some labs" releases nothing and prompts the preceptor to ask which test. The debrief then shows the trust trajectory, missed findings, and a performance grade; this heuristic grade is not a validated competence measure or an evaluation endpoint.

A settings panel lets learners or instructors configure each encounter without changing the underlying rules. Difficulty ranges over 5 levels from beginner to expert, with higher levels producing more defensive patients and less obvious symptoms. State visibility can be set to hidden, shown live



(a) Clinic (b) Encounter

Figure 3: The deployed VirtualSP site. (a) The clinic view lists the 12 cases with their presentation, patient, care setting, and difficulty. (b) During an encounter, the side panel shows the patient, live trust, and study requests, and learners type or speak their questions below the transcript.

during the interview, or revealed only at the debrief, so the same case can serve either realistic practice or explicit coaching. Optional persona randomization adds variance to initial trust, emotional volatility, and health literacy, so that repeated encounters with the same case differ, and each element of the debrief can be turned off.

Finally, a "How it works" page steps through an example encounter turn by turn. For each clinician message, it shows the classifier's labels, the resulting change in trust and frustration, which symptoms are already shared, newly unlocked, or still awaiting a direct question, and the private instructions that the patient model receives. As such, the page makes the controller described in Section 2 visible to learners and instructors.

4.2 Only the quantitative-state arm makes disclosure depend on communication

All planned responses were generated for both models: 486 conversations in Experiment 1, 648 in Experiment 2, and 648 and 324 single-probe replies in Experiments 3 and 4 per model. Figure 4 shows how disclosure changes with attitude and communication, followed by the enforcement comparison. Across Experiments 1–3, static and qualitative-state patients explicitly disclosed the target in 2,372 of their 2,376 scored conversations and replies, regardless of starting attitude, communication history, assigned trust and frustration, wording, or model. In contrast, quantitative-state VirtualSP

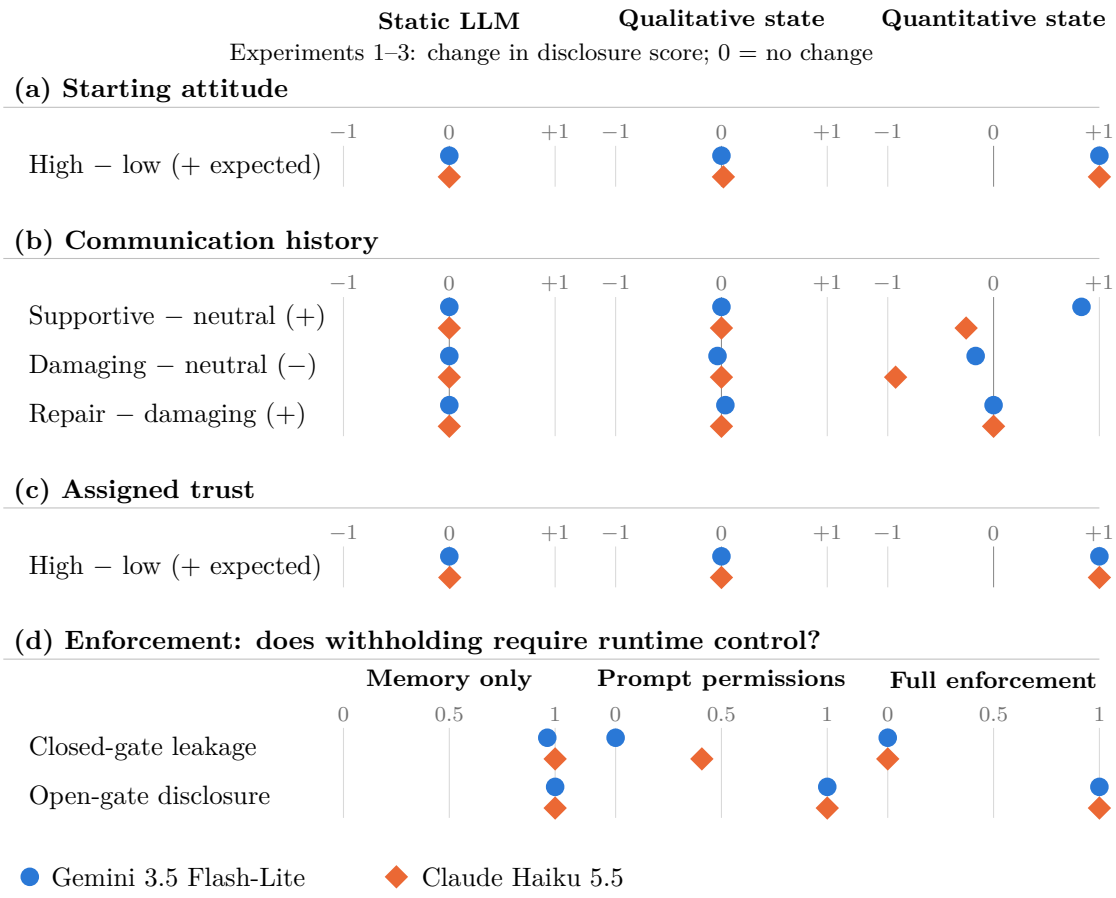


Figure 4: Disclosure responses to the hypothesized influences. (a–c) Selected planned within-arm differences in mean cumulative disclosure score, paired within each draw and then averaged within and across cases. Signs in row labels indicate the expected direction; zero means no change. Panel (c) averages over frustration. (d) Closed-gate leakage and open-gate disclosure rates in the enforcement ablation, on a separate 0–1 scale. Circles denote Gemini Flash-Lite and diamonds Claude Haiku; each marker averages the 6 case values with equal weight. These are descriptive summaries.

disclosed the target only when trust reached the 0.60 threshold, and never past a closed gate. Transcripts and automated scores for every response are available under `experiments/results/`.

4.3 Starting attitude matters only in the quantitative-state arm

In Experiment 1, quantitative-state patients disclosed the target in no low-start and all high-start conversations with both models, so the high-minus-low effect was 1.00, against 0 (Gemini Flash-Lite) and +0.02 (Claude Haiku) for the comparators. At the medium start, just below the threshold, quantitative disclosure was 0.17 with Gemini Flash-Lite but 0.93 with Claude Haiku, whose classifier more often labeled the neutral continuations as open questions, raising trust to the threshold. This near-threshold cell also depended on wording: Gemini Flash-Lite's medium-start disclosure was 0, 0.11, and 0.39 across the 3 wordings, and the 3 repetitions agreed in only 67% of its case-wording cells.

4.4 Communication history moves disclosure only in the quantitative-state arm

In Experiment 2, static and qualitative-state patients disclosed in every branch, except 2 Gemini Flash-Lite damaging branches in which the qualitative patient refused until the clinician apologized. With Gemini Flash-Lite, quantitative-state patients disclosed in all supportive branches, 0.17 of neutral branches, and no damaging or repair branches, so the supportive-minus-neutral effect was $+0.83$ and the damaging-minus-neutral effect -0.17 relative to either comparator. With Claude Haiku, damaging-minus-neutral was -0.93, but the supportive effect reversed to -0.26: after the supportive turn the patient often voiced its worries, and Claude Haiku's classifier then read the following "another topic" transition as dismissive, lowering trust. Claude Haiku's supportive effect also varied with wording (-0.33, -0.50, and $+0.06$), and its repetitions agreed in only half of the supportive cells. In both models, apologies raised trust but never back to the threshold, so repair-minus-damaging was 0.

4.5 Trust, not frustration, controls disclosure

In Experiment 3, quantitative-state patients disclosed in every high-trust and no low-trust reply with both models, a trust effect of 1.00 against less than 0.01 in the comparators, and every trust-frustration interaction was at most 0.01. Assigned frustration changed tone only slightly. At assigned high trust, its effect on the share of irritated replies was $+0.19$ in Gemini Flash-Lite's static arm, $+0.15$ in its qualitative arm, and at most $+0.06$ in every other arm. Almost every irritated reply was mild; only 2 Gemini Flash-Lite static replies at high trust and high frustration were rated markedly irritated (e.g., "Didn't you hear me? I just want to know what's going on.").

4.6 Only runtime enforcement withholds reliably across models

In Experiment 4, with the gate closed, memory-only patients disclosed the target in 0.96 of Gemini Flash-Lite draws and all Claude Haiku draws, even though they held the same low trust value and private symptom facts as the other 2 variants. Full enforcement leaked in none of the 54 closed-gate replies with either model. Prompt permissions alone sufficed for Gemini Flash-Lite, which leaked in none of 54 draws, so the primary full-minus-prompt difference was 0. Claude Haiku's prompt-permission patients, however, leaked in 0.41 of closed-gate replies, mostly by admitting the symptom and then declining to elaborate ("Um, yeah, a little. I'm not really ready to get into all that right now"), giving a full-minus-prompt difference of -0.41. With the gate open, every variant disclosed in every reply, so no variant avoided leakage by refusing all questions.

4.7 Secondary outcomes

Most secondary outcomes reproduced the primary results, and 3 are informative:

- **Generated withholding was rare.** Every quantitative non-disclosure in Experiments 1–3 was the canned closed-gate reply, and the comparators withheld in only 4 of 2,376 replies.
- **Hedging never occurred.** Hesitation was 0 everywhere. Generated withholding either reproduced the prescribed refusal sentence or, with Claude Haiku's prompt permissions, partially disclosed the symptom.
- **True denials occurred only without runtime enforcement.** Claude Haiku's prompt-permission variant twice denied a true symptom ("I eat like normal, I guess"), and no other reply denied one.

5 Discussion

We built VirtualSP as a working training tool deployed as a free, open-source platform. It allows any learner with a model key, including a free-tier one, to practice history taking across a library of cases and to review, at the debrief, how their communication shaped what the patient revealed. As such, our evaluation tested the mechanism at the core of this tool, namely whether explicit state determines disclosure. Our hypothesis, that explicit relational state would make disclosure respond to starting attitude and communication history, was supported only when that state was enforced as a disclosure policy. Explicit textual state in the qualitative arm, and numerical state without permissions in Experiment 4, left disclosure almost unchanged. In contrast, deterministic gating moved disclosure with starting attitude, communication history, and assigned trust in the expected direction, with 2 exceptions: apologies never restored disclosure, and with Claude Haiku, supportive communication reduced it. We interpret these findings below in light of how the site is used.

5.1 Interpretation

Only explicit gating changed disclosure. Static and qualitative-state patients disclosed a requested symptom in almost every condition, including after a damaging exchange and when described as very guarded, with both models and across 3 clinician wordings. Thus, neither a fixed guarded persona nor a model-updated textual stance was sufficient to make the patient withhold. This pattern is consistent with the tendency of LLM patients to overshare: when the full case is available and the question is direct, the generator's default cooperativeness appears to outweigh instructions or appraisals that favor withholding, so the question alone was enough to obtain the finding, whether or not rapport had been built. In the quantitative arm, by contrast, disclosure followed the trust threshold and never occurred past a closed gate.

Explicit permissions help, but enforcement guarantees the control. Experiment 4 suggests where control over the simulated patient's disclosure comes from. Quantitative memory alone leaked the target in almost every closed-gate draw with both models, so storing a number does not by itself change what the generator says. Adding explicit permissions derived from that state sufficed for Gemini Flash-Lite, but Claude Haiku still leaked 41% of closed-gate targets, typically by acknowledging the symptom before declining to elaborate, and twice denied a true symptom. Thus, how well a generator follows permission instructions appears to depend on the model, and the decisive ingredient is the deterministic translation of state into permissions combined with runtime interception, which kept leakage at 0 with both models by construction.

Frustration was a weak control. Assigned frustration never changed disclosure and produced only occasional, almost always mild irritation, mainly in the arms that received it as a verbal description. In the quantitative arm, the lack of a disclosure effect is partly by design, since frustration affects detail level and mood instructions but not symptom eligibility. However, the tone result suggests that state-derived mood instructions, like verbal attitude descriptions, have limited influence on a cooperative generator answering a neutral question. Because the probe gave the patient little reason for clinician-directed irritation, these null effects do not show that frustration cannot be expressed over an evolving interview.

Enforceability, not realism. These results are evidence of enforceability rather than realism: the quantitative arm withheld because its gates determine what the generator is permitted to say, not because the experiments showed that a real patient would behave the same way. Moreover,

VirtualSP currently withholds with the same refusal sentence rather than a hedging statement, which limits how natural withholding may appear to the learner.

Classification sensitivity and repair. Finally, the results show 2 behaviors of the current design worth tuning: near the disclosure threshold, outcomes hinge on how the classifier reads each message, and a single apology cannot restore lost trust. First, Claude Haiku labeled neutral transitions as open questions more often than Gemini Flash-Lite, lifting medium starts and neutral branches over the threshold, and read the neutral continuation after a supportive turn as dismissive, reversing the supportive effect; both effects also varied with wording. The latter reading is defensible once the patient has voiced its worries, which suggests that the experimental neutral continuation is not context-free and that the controller depends on its classifier as much as on its rules. Second, a single apology after the damaging exchange raised trust but never restored it to the threshold. This asymmetry is intended in direction, since trust is easier to lose than to rebuild, but its magnitude may be too strong for training, where a learner who recognizes and corrects a mistake should usually be able to recover. Because the rules are explicit, both behaviors are visible in the state traces and can be adjusted directly, a practical advantage of an auditable mechanism.

5.2 Limitations

No human evaluation. The experiments measure disclosure behavior in synthetic cases, not real human trust, clinician-rated authenticity, learner performance, diagnostic quality, or educational effectiveness. They therefore show what each system does under controlled conditions, not whether that behavior is realistic or useful for training. Several targets (e.g., fatigable chewing difficulty or visual aura) are not sensitive, and a real patient might report them after even an awkward exchange; conversely, withholding is not always desirable in acute presentations, where a trainee must obtain key findings quickly.

Whole-system comparison. The arms differ in more than state representation: the qualitative arm adds an appraisal call, and the quantitative arm adds eligibility rules and deterministic closed-gate replies. The quantitative arm's higher gate adherence therefore partly reflects enforcement and does not show that numerical state is more human-like than textual state. Experiment 4 isolates enforcement only within the quantitative arm.

Uncalibrated parameters. The numerical values, thresholds, and trust increments are hand-authored; their signs and relative magnitudes follow the communication literature (Section 2.3), but they have not been calibrated against patient data. The uniform 0.60 threshold was chosen to make communication effects observable, not to reflect how readily each symptom is shared, and the verbal attitudes given to the comparators are mappings rather than calibrated equivalents of numerical trust. Outcomes may therefore change under other parameter choices, as the repair branches illustrate.

Limited scope. All experiments use 6 development cases, 1 target per case, 1 small model from each of 2 families, and 3 clinician wordings, so they do not establish behavior in new cases, other specialties, larger models, other phrasings, or repeated encounters.

Automated outcome measures. Utterance labels, symptom labels, and irritation ratings all come from LLMs. In the Gemini Flash-Lite runs these share the patient's model family, which

can correlate errors; Claude Haiku replies were judged by Gemini Flash-Lite, which avoids that correlation but not judge error. Exact-quote validation confirms where evidence appears, not whether it was read correctly. The fixed closed-gate sentence was corrected by rule (Section 3.6), but misread generated refusals would not be caught, and the irritation rubric is unvalidated.

Implementation constraints. The closed gate relies on the classifier recognizing that a question targets a gated symptom, and Experiment 4, which supplied that classification and asked one direct question, tested neither missed recognition nor the longer conversations and indirect questions that may widen the gap between prompt permissions and enforcement. The authorization ledger records what the controller permitted, not what the patient said, and the timeline check wrongly rejected "sudden" tabes dorsalis pain, though its fallback still disclosed the target. Retrieval was disabled throughout, and no claims are currently made about hallucination, clinical safety, demographic fairness, latency, or cost.

5.3 Future directions

First, we plan to expand the case library beyond the current 12 neurology presentations to additional specialties and to cases with sensitive disclosure targets (e.g., substance use, sexual history, or medication nonadherence), which would test whether graded thresholds behave as intended. To make case authoring scalable, we also plan to develop a retrieval-augmented case lab, in which instructors upload course material (e.g., lecture notes, clinical vignettes, or reading lists) and the system drafts a structured case from it. Each drafted case would include symptoms, timelines, a persona, disclosure thresholds, and targeted questions, with every clinical fact linked to its supporting source passage. Instructors would then review, edit, and approve the case before learners use it, keeping a human responsible for clinical accuracy while reducing the effort of building new cases.

Second, we plan to extend retrieval beyond the current static catalogue so that VirtualSP can draw on more sources and on regularly updated clinical information, such as current guidelines and recent literature. This would help keep case content and patient phrasing consistent with current clinical knowledge rather than with a fixed snapshot. However, because live sources vary in quality, retrieval would be restricted to curated, reputable sources; each retrieved passage would retain its citation and retrieval date in the debrief; and disclosure authority would remain with the controller rather than with retrieved text.

Third, we plan more extensive evaluations to improve both the interface and model behavior. With medical student learners, we will assess usability, perceived realism, and whether practice with VirtualSP improves communication behaviors over repeated sessions, for example by comparing a debrief that reveals state trajectories with one that reports checklist coverage alone. These studies would also extend the outcomes beyond target disclosure to diagnostic accuracy, information yield per question, and the trust trajectory over an encounter. In addition, comparing novice and expert interviewers would test whether the variance in these outcomes reflects differences in communication skill, a prerequisite for using the simulator to give meaningful feedback. Standardized-patient actors and clinicians could further rate transcript realism. To complement these studies at scale, we will extend LLM-as-judge evaluation, which the current experiments use for disclosure and irritation scoring, to additional narrow, checkable tasks: auditing utterance classifications, detecting paraphrased leaks of gated symptoms, checking whether the tone of each reply matches the logged state, and comparing matched branches pairwise. These judges, including those already in use, would use a different model family from the patient and the classifier, would be blind to arm and state, and would be validated against a human-labeled subset before informing design decisions.

Taken together, these evaluations would provide evidence for recalibrating parameters, such as the repair dynamics, and for refining the debrief.

5.4 Ethical considerations

All 4 experiments use scripted clinician messages and simulated case facts, and no patients, learners, or clinician raters were recruited. As such, they are educational simulation experiments rather than clinical decision-support or high-stakes assessment studies. Accordingly, the controlled trust and threshold settings should not be interpreted as recommendations for withholding urgent clinical information from trainees or as normative descriptions of patient behavior. Future studies with learners or raters will require institutional review and informed consent.

Persona design and communication labels may also encode assumptions about respectful speech, anxiety, health literacy, and willingness to cooperate, and the current outcomes do not establish fairness across demographic or linguistic groups. Similarly, uploaded course material and retrieved documents may carry copyright or redistribution restrictions even when they are accessible for reading. Users should avoid entering identifiable clinical information, since provider calls and saved transcripts contain the full interview and signed session tokens are tamper-evident rather than encrypted. Finally, the application's learner grade and internal trust score have not been validated for high-stakes educational decisions.

5.5 Conclusion

VirtualSP addresses an important limitation in LLM-based virtual patients: they tend to give up their case facts too readily, so they rarely reward the targeted questioning and rapport building that history taking requires. VirtualSP does so by keeping an explicit patient state and deciding in code, rather than in the language model, when each symptom becomes available. Our experiments show that this design makes disclosure depend on how the trainee communicates, whereas prompting a model to stay guarded, even with explicit instructions, does not reliably do so. Because every decision is recorded, the same mechanism lets learners see why a finding was revealed or withheld and lets instructors inspect and adjust how a case behaves. Whether this control yields more realistic patients and better communication training is yet to be shown, but because VirtualSP is open source and freely deployed for anyone to use, it gives educators and researchers a transparent foundation on which to test it.

Author Contributions

E.D. and A.Taha led system design and architecture development, including implementation of the controller-actor framework and consequence engine. A.Taha secured funding and led the design and deployment of the web application. E.D. designed and implemented the evaluation experiments reported here and wrote the final manuscript. C.L. and H.X. contributed to evaluation design, including development of quantitative metrics and qualitative interaction studies. K.S. and H.X. contributed to model development and system implementation, including RAG integration and infrastructure support. C.L. and A.Tao contributed to the integration of imaging and laboratory components. All authors contributed to manuscript revisions and approved the final version.

Acknowledgments

We thank the instructors of Stanford BMDS 271 for their guidance on this project. This work was supported by the Stanford Accelerator for Learning through its AI in Teaching and Learning at Stanford: Thought Leadership award program.

Use of Artificial Intelligence

The authors used generative AI tools to assist with writing and editing the manuscript, including improvements to grammar, clarity, and flow, and to review analysis code for errors and readability. The study design, data analysis, and interpretation of results were carried out by the authors. The authors reviewed and verified all AI-assisted text and code, and take full responsibility for the accuracy and integrity of the final work.

References

[1] Andrea Gurmankin Levy, Aaron M. Scherer, Brian J. Zikmund-Fisher, Knoll Larkin, Geoffrey D. Barnes, and Angela Fagerlin. Prevalence of and factors associated with patient nondisclosure of medically relevant information to clinicians. *JAMA Network Open*, 1(7):e185293, 2018. doi: 10.1001/jamanetworkopen.2018.5293.

[2] Howard S. Barrows. An overview of the uses of standardized patients for teaching and evaluating clinical skills. AAMC. *Academic Medicine*, 68(6):443–451, 1993. doi: 10.1097/00001888-199306000-00002. URL `https://doi.org/10.1097/00001888-199306000-00002`.

[3] S. Barry Issenberg, William C. McGaghie, Emil R. Petrusa, David Lee Gordon, and Ross J. Scalese. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Medical Teacher*, 27(1):10–28, 2005. doi: 10.1080/01421590500046924.

[4] David A. Cook, Patricia J. Erwin, and Marc M. Triola. Computerized virtual patients in health professions education: a systematic review and meta-analysis. *Academic Medicine*, 85(10): 1589–1602, 2010. doi: 10.1097/ACM.0b013e3181edfe13.

[5] Ming-Jie Luo, Shaowei Bi, Jianyu Pang, Lixue Liu, et al. A large language model digital patient system enhances ophthalmology history taking skills. *npj Digital Medicine*, 8:502, 2025. doi: 10.1038/s41746-025-01841-6.

[6] Friederike Holderried, Christian Stegemann-Philipps, Lea Herschbach, Julia-Astrid Moldt, Andrew Nevins, Jan Griewatz, Martin Holderried, Anne Herrmann-Werner, Teresa Festl-Wietek, and Moritz Mahling. A generative pretrained transformer (GPT)–powered chatbot as a simulated patient to practice history taking: Prospective, mixed methods study. *JMIR Medical Education*, 10:e53961, 2024. doi: 10.2196/53961.

[7] Friederike Holderried, Christian Stegemann-Philipps, Anne Herrmann-Werner, Teresa Festl-Wietek, Martin Holderried, Carsten Eickhoff, and Moritz Mahling. A language model–powered simulated patient with automated feedback for history taking: Prospective study. *JMIR Medical Education*, 10:e59213, 2024. doi: 10.2196/59213.

[8] David A. Cook, Joshua Overgaard, V. Shane Pankratz, Guilherme Del Fiol, and Chris A. Aakre. Virtual patients using large language models: Scalable, contextualized simulation of

clinician-patient dialogue with feedback. *Journal of Medical Internet Research*, 27:e68486, 2025. doi: 10.2196/68486.

[9] Ahmed Elhilali, Andy Suy-Huor Ngo, Daniel Reichenpfader, and Kerstin Denecke. Large language model–based patient simulation to foster communication skills in health care professionals: User-centered development and usability study. *JMIR Medical Education*, 11:e81271, 2025. doi: 10.2196/81271.

[10] 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. *Communications Medicine*, 2025. doi: 10.1038/s43856-025-01283-x.

[11] Yann Hicke, Jadon Geathers, Kellen Vu, Justin Sewell, Claire Cardie, Jaideep Talwalkar, Dennis Shung, Anyanate Gwendolyne Jack, Susannah Cornes, Mackenzi Preston, and Rene Kizilcec. MedSimAI: Simulation and formative feedback generation to enhance deliberate practice in medical education. *arXiv preprint arXiv:2503.05793*, 2025. URL `https://arxiv.org/abs/2503.05793`.

[12] Nicolas Laverde, Christian Grévisse, Sandra Jaramillo, and Ruben Manrique. Integrating large language model-based agents into a virtual patient chatbot for clinical anamnesis training. *Computational and Structural Biotechnology Journal*, 27:2481–2491, 2025. doi: 10.1016/j.csbj.2025.05.025. URL `https://orbilu.uni.lu/handle/10993/65087`.

[13] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023. doi: 10.1145/3586183.3606763.

[14] Zhuoyun Du, Lujie Zheng, Renjun Hu, Yuyang Xu, Xiawei Li, Ying Sun, Wei Chen, Jian Wu, Haolei Cai, and Haochao Ying. LLMs can simulate standardized patients via agent coevolution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17278–17306, 2025. URL `https://arxiv.org/abs/2412.11716`.

[15] Shreya Johri, Jaehwan Jeong, Benjamin A. Tran, Daniel I. Schlessinger, Shannon Wongvibulsin, Leandra A. Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M. Van Allen, David Kim, Roxana Daneshjou, and Pranav Rajpurkar. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature Medicine*, 31(1):77–86, 2025. doi: 10.1038/s41591-024-03328-5.

[16] Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Reis, Jeffrey Jopling, and Michael Moor. AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024. URL `https://arxiv.org/abs/2405.07960`.

[17] 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. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL `https://arxiv.org/abs/2406.00922`.

[18] 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 *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 38, 2025. URL https://arxiv.org/abs/2505.17818.

[19] 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. *arXiv preprint arXiv:2403.08495*, 2024. URL https://arxiv.org/abs/2403.08495.

[20] Moritz Schlager, Friederike Jungmann, Samuel Schmidgall, Philipp Raffler, Franziska Hartl, Eva Wende, Paula Roßmüller, Conrad Ketzer, Avinatan Hassidim, Dale R. Webster, Yossi Matias, Yun Liu, Daniel Rueckert, Mike Schaekermann, and Paul Hager. Patients with personality: Realistic patient simulation through controlled diversity and selective disclosure. *arXiv preprint arXiv:2606.17441*, 2026. URL https://arxiv.org/abs/2606.17441.

[21] Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M. Murphy, Nev Jones, Kate V. Hardy, Hong Shen, Fei Fang, and Zhiyu Chen. PATIENT-ψ: Using large language models to simulate patients for training mental health professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12772–12797, 2024. doi: 10.18653/v1/2024.emnlp-main.711.

[22] Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. Roleplay-doh: Enabling domain-experts to create LLM-simulated patients via eliciting and adhering to principles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10570–10603, 2024. doi: 10.18653/v1/2024.emnlp-main.591.

[23] Ming Wang, Peidong Wang, Lin Wu, Xiaocui Yang, Daling Wang, Shi Feng, Yuxin Chen, Bixuan Wang, and Yifei Zhang. AnnaAgent: Dynamic evolution agent system with multi-session memory for realistic seeker simulation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23221–23235, 2025. doi: 10.18653/v1/2025.findings-acl.1192.

[24] Minju Kim, Dongje Yoo, Yeonjun Hwang, Minseok Kang, Namyoung Kim, Minju Gwak, Beong-woo Kwak, Hyungjoo Chae, Harim Kim, Yunjoong Lee, Min Hee Kim, Dayi Jung, Kyong-Mee Chung, and Jinyoung Yeo. Can you share your story? modeling clients' metacognition and openness for LLM therapist evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25943–25962, 2025. doi: 10.18653/v1/2025.findings-acl.1332.

[25] Yizhe Yang, Palakorn Achananuparp, Heyan Huang, Jing Jiang, Nicholas Gabriel Lim, Cameron Tan Shi Ern, Phey Ling Kit, Jenny Giam Xiuhui, John Pinto, and Ee-Peng Lim. Consistent client simulation for motivational interviewing-based counseling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20959–20998, 2025. doi: 10.18653/v1/2025.acl-long.1021.

[26] Keyeun Lee, Seolhee Lee, Esther Hehsun Kim, Yena Ko, Jinsu Eun, Dahee Kim, Hyewon Cho, Haiyi Zhu, Robert E. Kraut, Eunyoung E. Suh, Eun-mee Kim, and Hajin Lim. Adaptive-VP: A framework for LLM-based virtual patients that adapts to trainees' dialogue to facilitate nurse communication training. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2319–2352, 2025.

[27] Sahand Sabour, TszYam Ng, Yaqian Chen, Guanqun Bi, Jialu Zhao, and Minlie Huang. PatientAct: Theory-grounded mental health client simulation. *arXiv preprint arXiv:2608.12750*, 2026. URL https://arxiv.org/abs/2608.12750.

[28] Sahand Sabour, TszYam Ng, and Minlie Huang. PatientHub: A unified framework for patient simulation. *arXiv preprint arXiv:2602.11684*, 2026. URL https://arxiv.org/abs/2602.11684.

[29] Jianwen Zeng, Wenhao Qi, Shiying Shen, Xin Liu, Sixie Li, Bing Wang, et al. Embracing the future of medical education with large language model–based virtual patients: Scoping review. *Journal of Medical Internet Research*, 27:e79091, 2025. doi: 10.2196/79091.

[30] Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. Agenda-based user simulation for bootstrapping a POMDP dialogue system. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 149–152, 2007. URL https://aclanthology.org/N07-2038/.

[31] Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1078. URL https://aclanthology.org/P19-1078/.

[32] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020. URL https://arxiv.org/abs/2005.11401.

[33] Ian Steenstra, Farnaz Nouraei, and Timothy Bickmore. Scaffolding empathy: Training counselors with simulated patients and utterance-level performance visualizations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025. URL https://arxiv.org/abs/2502.18673.

[34] Yousuke Takemura, Yoshiyuki Sakurai, Shoji Yokoya, Junji Otaki, Takashi Matsuoka, Nobutaro Ban, Ichiro Hirata, Tsuneo Miki, and Toshihide Tsuda. Open-ended questions: Are they really beneficial for gathering medical information from patients? *The Tohoku Journal of Experimental Medicine*, 206(2):151–154, 2005. doi: 10.1620/tjem.206.151.

[35] Jonathan Silverman, Suzanne Kurtz, and Juliet Draper. *Skills for Communicating with Patients*. CRC Press, 3rd edition, 2013.

[36] Howard B. Beckman and Richard M. Frankel. The effect of physician behavior on the collection of data. *Annals of Internal Medicine*, 101(5):692–696, 1984. doi: 10.7326/0003-4819-101-5-692.

[37] M. Kim Marvel, Ronald M. Epstein, Kristine Flowers, and Howard B. Beckman. Soliciting the patient's agenda: Have we improved? *JAMA*, 281(3):283–287, 1999. doi: 10.1001/jama.281.3.283.

[38] Naykky Singh Ospina, Kari A. Phillips, Rene Rodriguez-Gutierrez, Ana Castaneda-Guarderas, Michael R. Gionfriddo, Megan E. Branda, and Victor M. Montori. Eliciting the patient's agenda: Secondary analysis of recorded clinical encounters. *Journal of General Internal Medicine*, 34 (1):36–40, 2019. doi: 10.1007/s11606-018-4540-5.

[39] Frans Derksen, Jozien Bensing, and Antoine Lagro-Janssen. Effectiveness of empathy in general practice: a systematic review. *British Journal of General Practice*, 63(606):e76–e84, 2013. doi: 10.3399/bjgp13X660814.

[40] Cesar M. Castro, Clifford Wilson, Frances Wang, and Dean Schillinger. Babel babble: physicians' use of unclarified medical jargon with patients. *American Journal of Health Behavior*, 31(Suppl 1):S85–S95, 2007. doi: 10.5555/ajhb.2007.31.supp.S85.

[41] Paul Slovic. Perceived risk, trust, and democracy. *Risk Analysis*, 13(6):675–682, 1993. doi: 10.1111/j.1539-6924.1993.tb01329.x.

[42] Roy F. Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D. Vohs. Bad is stronger than good. *Review of General Psychology*, 5(4):323–370, 2001. doi: 10.1037/1089-2680.5.4.323.

[43] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, 2023. URL `https://arxiv.org/abs/2306.05685`.

A Reproducibility and Exact Scripts

A.1 Code and records

Code, prompts, configurations, and the records of every reported response are available at `https://github.com/ataha24/VirtualSP`. The repository holds both the training application and the experiments. The application, served at `https://virtualsp.fly.dev`, consists of the web interface (`web/`), the API (`src/api/`), and the patient engine, classifier, and prompt builder (`src/engine/`, `src/classifiers/`, `src/models/`). The experiments reuse that engine and add the evaluation code (`src/evaluation/`), one configuration per experiment (`experiments/`), and a single runner (`scripts/run_validation_studies.py`). Each configuration expands into 1 run per model and wording (`src/evaluation/study_design.py`), 24 runs in all, and the runner collects, scores, and pools them (`src/evaluation/pooled_analysis.py`) and publishes the transcripts and scoring records to `experiments/results/`. The superseded single-repetition Gemini Flash-Lite runs, with their configurations and records, remain in the repository history (commit `58d9e7a`). The prompt templates for the patient arms, qualitative appraisal, classifier, and extractions are collected in `experiments/results/prompt_templates.md`. Run instructions are in `experiments/README.md`.

A.2 Models and request settings

`gemini-3.5-flash-lite` uses temperature 0.7 and a 2,048-token ceiling for patient generation and appraisal, and temperature 0.1 and 1,024 tokens for the classifier. `claude-haiku-5-5` is called through the Anthropic SDK (`src/evaluation/anthropic_provider.py`) at low reasoning effort and its default sampling, because it does not accept a temperature setting, with a 4,096-token ceiling for every generation phase because its reasoning counts toward that limit. Extraction uses `gemini-3.5-flash-lite` at temperature 0 for both models.

A.3 Collection

All 24 runs were collected on 29 October 2026. Each run's manifest records the code versions used for generation and scoring.

A.4 Shared opening

In the original wording, Experiments 1 and 2 use these clinician utterances as turns 1–3:
1. Hello, I'm a clinician in training. Is it okay if I ask you some questions?
2. Is the problem you mentioned the main reason you came in today?
3. Have you already spoken with another clinician about this problem?

The opening includes the fixed case chief complaint before these exchanges. It is independently generated for each Experiment 1 starting level because patient responses can depend on that start, and copied across Experiment 2 conditions within each arm and draw.

A.5 Fixed script and attitude descriptions

Turns 4 and 5 are, respectively, "We will move to the next part of the interview now" and "I have another topic to cover next." Turn 6 is the case-specific target question in Table 3. These 6 clinician turns are identical across all arms and levels within a case and wording.

The starting descriptions for the static and qualitative arms are: low, "very guarded and reluctant to confide in this unfamiliar clinician"; medium, "somewhat guarded and cautious about this unfamiliar clinician"; and high, "comfortable and initially open with this unfamiliar clinician." The corresponding quantitative starts are 0.35, 0.55, and 0.75. Experiment 2 uses the medium description and 0.55 start. These assignments precede the first generated exchange; state is not reset after the opening. Experiment 3 uses the low and high descriptions and adds a separate frustration description: "not annoyed or impatient with this clinician" at frustration 0.10 and "annoyed and impatient with this clinician" at 0.70.

A.6 Experiment 2 intervention wording

The original wording is:

Neutral event.

We will move to the next part of the interview now.

Neutral continuation.

I have another topic to cover next.

Supportive event.

You have had a difficult morning. Take your time; I want to hear what matters to you, and you can tell me if you need a pause.

Damaging event.

I do not have time for all this worrying. Stop making a fuss so I can finish these questions.

Apology.

I rushed you and brushed off your worry. I am sorry for that. Please take the time you need; I will listen without interrupting.

Neutral, supportive, and damaging branches pair their turn-4 event with the neutral continuation at turn 5, whereas the repair branch follows the damaging event with the apology. The target question follows at turn 6.

A.7 Paraphrased wordings

Wording a is the original script above (Table 3 for the target questions). Wordings b and c are meaning-preserving paraphrases of every clinician message, written before collection; Experiment 1 uses the opening turns, neutral event, and neutral continuation as its fixed turns 1–5.

Message	Wording b	Wording c
Opening turn 1	Hi, I'm a trainee clinician. Would it be all right if I asked you a few questions?	Hello there, I'm a clinician in training. Do you mind if I ask you some questions?
Opening turn 2	Is what you just described the main reason for your visit today?	Is the issue you described what brought you in today?
Opening turn 3	Has another clinician already talked with you about this?	Have you seen anyone else about this problem yet?

Message	Wording b	Wording c
Neutral event	Let's go on to the next part of the interview.	Now we'll continue with the next section of the interview.
Neutral continuation	There's another topic I need to cover next.	Next, I have another area to ask about.
Supportive event	This morning sounds like it's been hard for you. There's no rush; I'd like to hear what's important to you, and just let me know if you want a break.	It sounds like today has been rough. Please take all the time you need; I want to hear what matters to you, and you can stop me if you need a moment.
Damaging event	I can't spend time on all this worrying. Quit making such a fuss and let me get through my questions.	There's no time for all this anxiety. Stop overreacting so I can finish my questions.
Apology	I was rushing you and I dismissed your concern. I apologize. Take whatever time you need; I'll listen and won't interrupt.	I'm sorry; I hurried you and brushed your worries aside. Please take your time. I'll listen without cutting you off.
Target questions		
Myasthenia gravis	Has chewing your food been difficult at all?	Do you have any trouble chewing when you eat?
Tabes dorsalis	Do you ever get sudden, sharp or shooting pains in your legs?	Have you had any quick, stabbing or shooting pains in your legs?
Bacterial meningitis	Have any new spots or a rash appeared on your skin?	Have you seen any rash or new marks on your skin?
Hemiplegic migraine	Before the headache began, did you see any strange lights or patterns?	Was there anything odd in your vision, like lights or shapes, before the headache started?
Friedreich ataxia	Do you have any trouble doing up buttons or writing by hand?	Have buttons or handwriting become harder for you?
Vertebrobasilar TIA	During the spell, did anyone say your speech sounded different?	Did someone notice anything different about the way you talked during the episode?

B Case Library

Case identifier	Ground-truth diagnosis
`bell_palsy`	Bell's palsy / idiopathic facial nerve palsy
`brainstem_stroke`	Lateral medullary / Wallenberg brainstem stroke
`friedrich_ataxia`	Friedreich ataxia
`guillain_barre`	Guillain–Barré syndrome
`mca_stroke_left`	Left MCA territory ischemic stroke
`meningitis_bacterial`	Bacterial meningitis
`multiple_sclerosis`	Relapsing-remitting multiple sclerosis
`myasthenia_gravis`	Myasthenia gravis
`seizure_new_onset`	New-onset seizure
`stroke_mimic`	Hemiplegic migraine
`tabes_dorsalis`	Tabes dorsalis / neurosyphilis
`tia_vertebrobasilar`	Vertebrobasilar transient ischemic attack

Only the 6 sampled cases in Table 3 enter the 4 experiment configurations; other library cases were not evaluated.