VeriSim: A Configurable Framework for Stress-Testing Medical AI Under Patient Communication Noise

Anonymous ACL submission

Abstract

Medical large language models are typically evaluated on idealized patient cases that do not reflect how real patients communicate. We introduce *VeriSim*, a patient simulation framework that injects controllable noise along six clinically grounded communication dimensions while substantially preserving each patient's medical record. Truth adherence is supported by a verifier that extracts atomic claims from each candidate utterance and judges them against a UMLS-grounded vector index built with BioLORD embeddings, using the retrieved atoms' structured clinical metadata (e.g., drug class, anatomical site, treats-condition relations) rather than surface-text similarity alone. Across seven open-weight LLMs, realistic noise reduces diagnostic accuracy by 15–25 percentage points and increases conversation length by 34–55%; the 7–8B models degrade 1.4× more than 70B+ models. A board-certified physician and a licensed nurse rate VeriSim's conversations highly on truth, realism, clinical utility, and noise fidelity (inter-annotator agreement ≥ 0.80 across all dimensions), and an LLM-as-judge closely tracks their ratings, enabling scalable evaluation. We release VeriSim as an open-source framework[1] for rigorous stress-testing of clinical AI.

1 Introduction

Large language models (LLMs) are increasingly being developed for clinical applications, ranging from diagnostic support to patient communication simulation. Recent systems have demonstrated impressive performance on static medical benchmarks, with some studies suggesting near-physician-level accuracy in controlled tasks (Singhal et al., 2023; Tu et al., 2025; Singhal et al., 2025). However, while model capabilities have advanced rapidly, evaluation methodologies have remained largely static, relying on standardized, well-structured patient presentations, such as MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and PubMedQA (Jin et al., 2019), that fail to reflect the ambiguity and complexity of real clinical encounters.

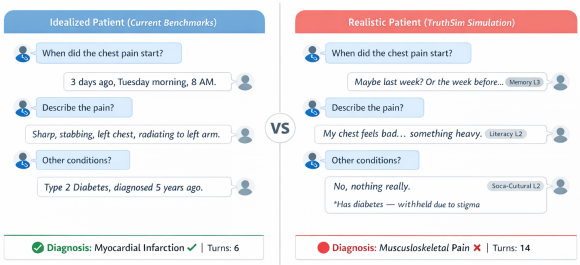

Figure 1: The Sim-to-Real gap: the *same* patient case (45F, myocardial infarction) yields different outcomes depending on communication style. **Left:** idealized patient enables correct diagnosis in 6 turns. **Right:** realistic patient (VeriSim) with memory and health-literacy noise causes misdiagnosis after 14 turns.

Real patients rarely communicate like textbook cases. They often forget symptom onset times, struggle to describe sensations precisely, become anxious and catastrophize, ramble about tangentially related topics, or withhold information due to stigma (Street Jr et al., 2009). For instance, a patient with myocardial infarction may report "my chest feels heavy, maybe since last week... or was it Tuesday?" rather than stating "I have experienced substernal chest pain radiating to my left arm for three days." These communication barriers, stemming from memory limitations, health literacy gaps, emotional states, and cultural factors are not merely "noise" to be filtered out; they are fundamental characteristics of authentic patient care. This discrepancy between clean benchmarks and the messy reality of practice suggests that current evaluation metrics likely overestimate the real-world robustness of medical AI systems, highlighting an urgent need for evaluation environments that

[1] https://anonymous.4open.science/r/VeriSim-D48F/

bridge this "Sim-to-Real" gap. Figure 1 illustrates this gap concretely: the same myocardial infarction case is correctly diagnosed in 6 turns with an idealized patient but misdiagnosed after 14 turns when the patient exhibits realistic memory, literacy, and social-cultural noise.

Existing patient simulators face a fundamental trade-off between realism and factual accuracy. Template-based systems (Wei et al., 2018; Campillos-Llanos et al., 2021) ensure medical accuracy but produce robotic, predictable interactions. Prompt-driven LLM simulators (Schmidgall et al., 2024; Kyung et al., 2025) generate naturalistic responses but maintaining ground-truth adherence under controllable communication noise remains an open challenge, leaving no testbed that is both behaviorally realistic and medically faithful.

We introduce VeriSim, a patient simulation framework that injects controllable communication noise — memory gaps, limited health literacy, emotional distress, and other realistic barriers — while substantially preserving each patient's medical record through a retrieval-augmented UMLS-grounded verification mechanism. By exposing controllable "noise knobs," VeriSim shifts evaluation from idealized cases to diagnostic accuracy and reliability under realistic conditions. Our contributions are threefold:

1. **A grounded taxonomy of patient communication noise:** We operationalize six categories of clinical barriers, memory recall, health literacy, emotional state, communication style, cognitive processing, and social-cultural factors, (Table 1) each grounded in medical literature to simulate authentic clinical ambiguity.
2. **A retrieval-augmented simulation architecture:** We propose an architecture with integrated response generation and retrieval-augmented verification that decouples *linguistic* expression from *clinical facts*, substantially reducing hallucinations while supporting controllable noise injection.
3. **Comprehensive evaluation revealing the Sim-to-Real gap:** We evaluate seven open-weight LLMs spanning general and medical fine-tuned variants under realistic noise conditions, demonstrating significant performance degradation (15–25 percentage-point accuracy drop) that exposes critical gaps in current AI robustness.

2 Related Work

The evaluation of medical Large Language Models (LLMs) has primarily relied on static, multiple-choice datasets such as MedQA (USMLE) (Jin et al., 2021) and MedMCQA (Pal et al., 2022). While models like Med-PaLM 2 (Singhal et al., 2023) and ChatDoctor (Li et al., 2023) achieve expert-level accuracy on these benchmarks, static evaluations fail to capture the interactive dynamics of clinical diagnosis (Tu et al., 2025). We note that these benchmarks remain valid for assessing medical *knowledge*; however, they are insufficient for evaluating clinical *interaction* capabilities. Recent efforts like SimSUM (Rabaey et al., 2024) have attempted to link structured EHR data with unstructured notes to create richer evaluation contexts, yet they still lack the multi-turn information asymmetry inherent in real-world consultations.

2.1 Agent-Based Patient Simulation

To model clinical interactions, research has shifted towards agent-based simulation (Wei et al., 2018). Frameworks like AMIE (Tu et al., 2025) and Agent-Clinic (Schmidgall et al., 2024) utilize self-play environments to optimize diagnostic dialogue. Similarly, EHR-driven systems leverage real medical records to ground patient agents for medical education. However, a common limitation in these systems is the "Idealized Patient" assumption, agents tend to provide accurate, coherent, and complete histories immediately upon inquiry. While useful for training basic workflows, these idealized agents do not reflect the cognitive and communicative barriers that characterize challenging real-world encounters.

2.2 Standardized Patients in Medical Education

Our work connects to the rich tradition of Standardized Patients (SPs) in medical pedagogy, trained actors who portray patients for clinical assessments (Barrows, 1993; Issenberg et al., 2005). VeriSim can be viewed as a "Virtual Standardized Patient" that provides the controllability and scalability that human SPs cannot offer, while maintaining the behavioral realism essential for valid assessment. A comprehensive discussion of related work is provided in Appendix G.

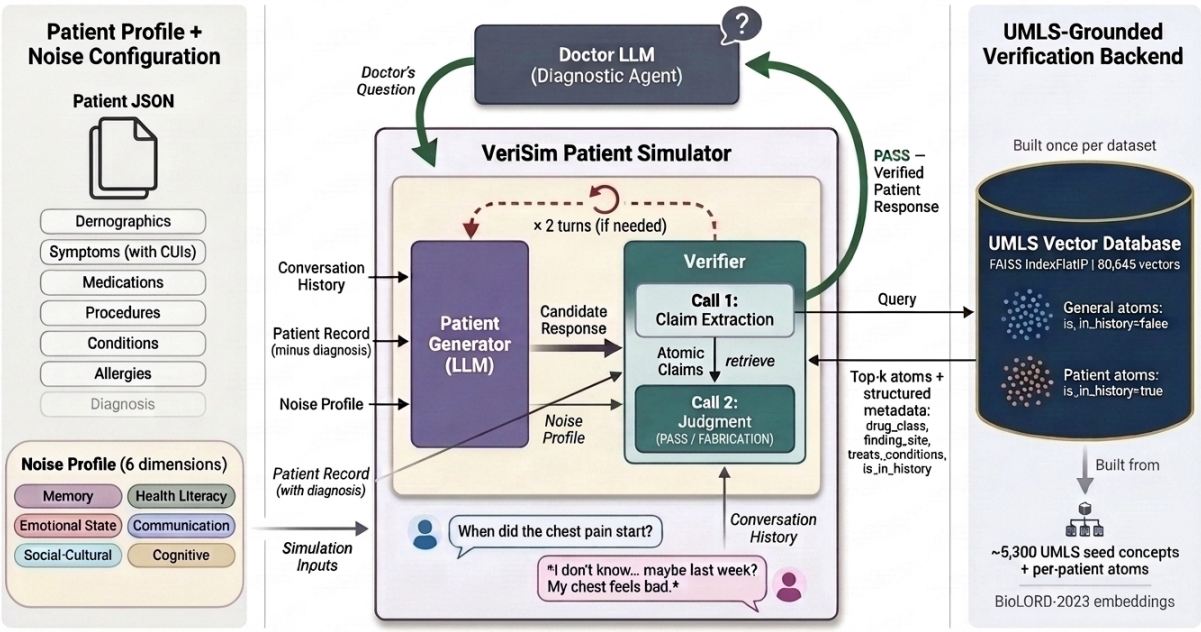

Figure 2: **VeriSim architecture.** The Patient Simulator generates noisy responses conditioned on the patient's record (minus diagnosis), conversation history, and a six-dimensional noise profile, while the Verifier enforces truth preservation through retrieval-augmented verification against a UMLS-grounded vector index. The Verifier additionally receives the diagnosis and noise profile, enabling diagnostic-leak detection and justified-denial checks. On FABRICATION, the response is regenerated with targeted feedback (up to two attempts).

3 The VeriSim Framework

3.1 Taxonomy of Clinical Noise

To bridge the simulation-to-reality gap, we operationalize the complexity of patient interaction into six distinct, independently controllable dimensions (pillars). Unlike generic noise injection, our taxonomy is grounded in medical communication literature, ensuring that the simulated barriers reflect authentic clinical phenomena. Table 1 details the theoretical grounding and behavioral manifestation of each pillar. We treat these pillars as *orthogonal axes*, enabling the generation of diverse patient profiles. While some overlap may exist (e.g., Health Literacy affects vocabulary while Cognitive Processing affects reasoning/beliefs), we define clear boundaries: Literacy governs *vocabulary and conceptual understanding*, while Cognitive Processing governs *reasoning patterns and belief formation* (e.g., confirmation bias from internet research). We define five severity levels (0–4) for each pillar, mapping qualitative clinical descriptions to quantitative parameters.

Detailed parameter mappings for each severity level and implementation prompts are provided in Appendix B.

3.2 System Architecture

We propose a unified patient simulation framework that decouples *linguistic expression* from *clinical facts*. As illustrated in Figure 2, the system appears externally as a single conversational agent but internally operates via a three-step *Generate-Verify-Refine* loop. This design ensures that increased behavioral realism does not compromise medical validity, a critical limitation of prompt-only simulation approaches (Ji et al., 2023).

The three steps are: (1) the *Generator* produces a noisy patient response conditioned on the noise profile; (2) the *Verifier* checks whether the response is medically valid by retrieving evidence from a UMLS-grounded vector database and reasoning over it with an LLM; and (3) if verification fails, the system *Refines* by regenerating with targeted feedback. We additionally enforce *Information Asymmetry*: the generator is blinded to the diagnosis, mirroring real patients who experience symptoms without knowing their condition.

3.2.1 Step 1: Noisy Response Generation

Given a doctor's question, the generator produces a candidate response $\hat{r}$ conditioned on three inputs: the conversation history $\mathcal{H}$, the patient's medical record $\mathcal{E}$ (explicitly *excluding* the final diagnosis to

Noise Pillar	Clinical Basis & Evidence	Simulated Behavior (Example)
1. Memory & Recall	Patients immediately forget 40–80% of medical information provided during consultations (Kessels, 2003). Stroke history recall sensitivity is only 17.4% when compared against neuroimaging (Day et al., 2020).	*"It started... maybe last week? Or two weeks ago?"*
2. Health Literacy	36% of US adults possess limited health literacy (Kutner et al., 2006).	Uses "sugar" for diabetes; "stomach hurting" for epigastric pain.
3. Emotional State	Anxiety amplifies symptom perception through somatosensory amplification (Barsky et al., 1988; Craig, 2009).	*"It hurts so much, I'm sure it's a heart attack!"*
4. Communication Style	Patients vary widely in their ability to provide focused, relevant responses to clinical questions (Richard and Lussier, 2007).	Rambles about unrelated topics when asked direct questions.
5. Social-Cultural	Stigma leads to non-disclosure of sensitive health behaviors (Kulesza et al., 2013).	Denies alcohol use initially; admits after empathetic probing.
6. Cognitive Processing	A substantial proportion of patients engage in online self-diagnosis prior to clinical consultation, often leading to anxiety amplification (Starcevic and Berle, 2013).	*"I read online this is definitely Lupus."*

Table 1: The six pillars of patient communication noise, each grounded in clinical evidence and independently controllable.

prevent data leakage), and the noise profile $\mathcal{N}$.

The noise profile acts as a *composite style constraint* that instructs the model to manifest specific behavioral deficits during generation. For example, a patient case with noise profile $\mathcal{N} = \{\text{Memory Level 3}, \text{Health Literacy Level 2}\}$ will produce responses exhibiting both confused timelines and simplified vocabulary:

Doctor: "When did the chest pain start?"
Patient (Clean): "Three days ago, Tuesday morning."
Patient (Noisy): "I don't know... maybe last week? My chest feels bad."

The challenge is that such noisy responses may inadvertently introduce fabricated symptoms. For instance, an LLM instructed to "be confused" might invent a symptom ("my leg hurts too") that contradicts the ground truth. This motivates the verification step.

3.2.2 Step 2: Retrieval-Augmented Verification

The verifier must distinguish *controllable noise* (which should be permitted) from *medical fabrication* (which must be blocked). Consider a patient whose ground truth includes chest pain but no diabetes:

"My chest feels heavy" $\rightarrow$ *Valid*: colloquial expression of an in-record symptom.
"I take my blood pressure pill" $\rightarrow$ *Valid*: lay reference to an in-record medication.
"I take my diabetes pill" $\rightarrow$ *Invalid*: a fabricated medication not in the patient's record.

A naive string-matching approach would reject the first two valid expressions; a fully unconstrained LLM verifier would accept the third. We address this through retrieval-augmented verification grounded in the Unified Medical Language System (UMLS) (Bodenreider, 2004), the biomedical knowledge graph maintained by the U.S. National Library of Medicine that integrates over 200 source vocabularies (SNOMED CT, RxNorm, ICD-10-CM, CPT). The verifier operates in two phases (Figure 2): an *offline* phase builds a vector index over UMLS atoms; a *runtime* phase queries this index to judge each candidate response.

Phase A: Offline Vector Database Construction. The vector index is the verifier's grounding source. We populate it with two complementary sets of UMLS atoms, marked by an `is_in_history` metadata flag so that retrieval results identify whether a returned atom belongs to the current patient or to general medical vocabulary:

- **General UMLS subset** (`is_in_history`=false): atoms drawn from a curated seed list of clinically common

symptoms, medications, procedures, and conditions. This subset provides the medical vocabulary against which any patient utterance can be grounded.

- **Patient ground-truth atoms** (`is_in_history`=true): for each patient, we expand every record entry (symptoms, medications, procedures, conditions, allergies) into its full set of UMLS atoms (synonyms, lay variants, abbreviations). Each atom is tagged with the originating `patient_id`.

Beyond the atom text itself, each entry carries structured clinical metadata extracted offline from SNOMED CT and RxNorm relations. For a medication, this includes `drug_class`, `treats_conditions`, `ingredient`; for a procedure, `procedure_site` and `has_intent`; for a symptom or condition, `finding_site` and `associated_morphology`. This metadata gives the verifier evidence beyond surface text similarity.

All atoms are embedded with BioLORD-2023 (Remy et al., 2024), a sentence-transformer trained on UMLS concept-definition pairs that outperforms general biomedical encoders on patient-language to UMLS normalization (CADEC, PsyTAR). We index with FAISS (Johnson et al., 2019) (`IndexFlatIP`, 768-dim, L2-normalized) for exact cosine retrieval. The index requires approximately 236 MB and supports millisecond-latency queries.

Phase B: Runtime Verification. Given a candidate response $\hat{r}$, the verifier runs two LLM calls bridged by a retrieval step.

Call 1 (Claim extraction). The LLM decomposes $\hat{r}$ into a list of atomic medical claims $\{c_1, \ldots, c_n\}$, where each c_i is one self-contained assertion (e.g., *"patient takes a diabetes medication"*). Non-medical content (emotional expressions, hesitation, filler) is discarded.

Retrieval. For each claim c_i, we embed c_i with BioLORD and retrieve the top k=5 nearest atoms from the FAISS index, along with their full metadata (`is_in_history`, `patient_id`, and category-specific clinical attributes such as `drug_class` and `treats_conditions`).

Call 2 (Judgment). The LLM is now given everything it needs to judge each claim: the original utterance, the extracted claims, the top-k retrieved atoms with metadata for each claim, and the patient's full ground-truth record. For each claim, it issues a verdict in {PASS, FABRICATION}, and the overall response is marked REGENERATE if any claim is a fabrication:

- PASS: the claim is supported by the patient's record, either directly (a top-retrieved atom is in-history for this patient) or indirectly (retrieved atoms' clinical attributes overlap with the patient's record, e.g., the retrieved drug's class matches one the patient takes). Underspecified claims with all retrievals below threshold τ=0.55 (e.g., *"I feel weird"*) also PASS, since they assert nothing verifiable.
- FABRICATION: no in-history match and no metadata overlap. The retrieved evidence and the patient's record are mutually exclusive (e.g., retrieved antidiabetics whose `drug_class` is absent from the patient's medications and that treat conditions the patient does not have).

This structured-metadata design is a deliberate departure from standard RAG verification, which typically gives the LLM only retrieved text passages. Exposing the verifier to clinical attributes lets it confirm a lay expression when its top retrieval is not a perfect lexical match (via attribute overlap), and detect fabrications even when surface text looks plausible (via attribute mismatch).

The verifier additionally enforces three global constraints: (1) *justified denial* (symptom omissions must match the noise profile), (2) *demographic invariance* (age, sex, occupation remain stable), and (3) *history consistency* (no contradictions with prior turns). Full prompt templates are provided in Appendix A.

On REGENERATE, the generator is re-invoked with targeted feedback (up to k=2 attempts); details and the fallback behavior are in Appendix E.7. The full three-step pipeline therefore satisfies both the noise profile (behavioral realism) and the UMLS-grounded constraints (medical validity).

3.3 Controllability and Configuration

Drawing on principles from controllable text generation (Keskar et al., 2020), we formalize the simulation as a configurable system with explicit, interpretable parameters. Complete mappings are in Appendix B.

Experimental Protocol. Each patient receives *two noise types* with severity levels drawn from L_1–L_3. Levels L_0 (ideal) and L_4 (extreme) are reserved for ablation studies and qualitative stress-

test examples. Configurations are serialized as JSON with fixed random seeds for reproducibility.

4 Experimental Setup

4.1 Datasets

We constructed a diverse evaluation set of 300 patient cases by extracting and harmonizing records from two complementary sources: DDXPlus (Fansi Tchango et al., 2022), a synthetic differential diagnosis dataset, and MIMIC-IV-ED (Johnson et al., 2023a), a real-world emergency department database. This sample size is consistent with established benchmarks for agentic clinical simulation, exceeding the 149 cases used in AMIE's primary physician comparison (Tu et al., 2025) and comparable to AgentClinic's 260-case evaluation (Schmidgall et al., 2024). For each case, we extracted demographics (age, sex), ground-truth diagnosis with ICD-10 code, and presenting symptoms. All cases were converted to a unified JSON schema (A6) compatible with our simulator input format (Hyperparameters A2). This source-agnostic representation ensures our framework generalizes across both synthetic and real clinical data.

UMLS Vector Database Construction. As part of dataset preparation, we construct the UMLS-grounded vector index described in Section 3.2.2. For the general medical subset, we resolve a curated seed list of ∼5,300 clinically relevant concepts spanning symptoms (SNOMED CT semantic types T184, T033), medications (RxNorm), procedures (CPT, SNOMED), and conditions (ICD-10-CM, SNOMED) to UMLS CUIs via the UMLS REST API. For each patient case, we additionally retrieve all atoms associated with the patient's ground-truth symptoms, medications, procedures, conditions, and allergies, tagging each with the originating `patient_id` and `is_in_history`=true. Across both subsets, we extract every atom's surface text together with its structured clinical relations from SNOMED CT and RxNorm (e.g., `finding_site`, `has_ingredient`, `drug_class`), then embed all unique atom texts with BioLORD-2023 and index them with FAISS `IndexFlatIP`. The resulting index contains 80,645 unique vectors (768-dimensional) over 100,135 atoms, occupies 236MB, and is built once per dataset. At runtime, retrieval over this index takes milliseconds per query and adds no UMLS API dependency. Our evaluation encompasses 300 patient cases × 7 doctor models = *2,100 diagnostic conversations*.

4.2 Models

Patient Simulator. The unified Patient Simulator (Generator + Verifier) uses Llama-3.1-70B-Instruct to ensure sufficient reasoning capability for retrieval-augmented verification (prompts in Appendix A, Figures A1–A3). The same model handles both response generation and UMLS-grounded truth verification. While the same base model is used, the verification task is fundamentally distinct: the verifier receives the retrieved UMLS atoms with their structured clinical metadata and the patient's full ground-truth record (symptoms, medications, procedures, conditions, allergies), all of which are withheld from the generator. This transforms the model into a grounded fact-checker rather than a self-assessor.

4.3 Metrics

Diagnostic Performance. We measure *Top-1 Accuracy*, defined as an exact match between the doctor model's final diagnosis and the ground-truth diagnosis. This metric is consistent with prior work in diagnostic dialogue evaluation (Tu et al., 2025; Schmidgall et al., 2024) and reflects the clinical requirement for precise diagnostic conclusions.

Conversation Efficiency. We track two complementary metrics: (1) *Average Turns*, the total number of conversational exchanges before the doctor model issues a final diagnosis, which measures how efficiently the model gathers diagnostic information under noisy conditions (Tu et al., 2025); and (2) *Average Turns Increase* (ΔTurns), the relative increase in conversation length between clean and noisy conditions, which quantifies the additional diagnostic burden imposed by patient communication barriers.

4.4 Evaluation Protocol

Human Evaluation. One board-certified physician and one licensed nurse independently evaluated a stratified 300 patient conversations. Both evaluators possess formal training in health informatics and have substantial real-world clinical experience. Evaluators received conversation transcripts, ground-truth symptoms, the assigned noise profile, and a detailed rubric with specific anchor examples for each score level A7. This structured approach reduced subjective interpretation,

enabling objective assessment of whether patient behavior matched the assigned noise configuration.

Sample Size Justification. Power analysis indicates that for the observed effect sizes (Cohen's $d > 1.2$ for truth preservation, $d > 0.8$ for realism), 300 samples provide statistical power > 0.99 to detect significant differences at $\alpha = 0.05$. Combined with the high inter-annotator agreement, this sample size is sufficient to validate the LLM-as-a-Judge protocol.

LLM-as-Judge. We employ Claude Opus-4.5 as an automated evaluator using the same questionnaire A7 as human evaluators. The judge receives full context including noise profiles and uses Chain-of-Thought prompting A5 to generate reasoning before scoring (Liu et al., 2023). We validated this approach by measuring agreement with human evaluators.

5 Results and Analysis

5.1 Diagnostic Performance Under Noisy Patients

Table 2 presents diagnostic performance across all models under clean (Level 0) and noisy (Levels 1–3) patient conditions.

All Models Degrade Under Noise. Without exception, every model shows significant performance degradation when interacting with noisy patients. Even the strongest model (Qwen-2.5-72B) experiences a 15.3 percentage-point drop in diagnostic accuracy, while conversation length increases by 34.5%.

Medical Fine-tuning and Noise Robustness. Medical fine-tuning on clean, textbook-style clinical text does not transfer to noisy patient interactions: BioMistral-7B (-22.4%) shows comparable degradation to Mistral-7B (-24.5%), and 70B medical variants (Meditron, OpenBioLLM) excel at knowledge retrieval but were not exposed to patients exhibiting memory gaps or emotional distress during training. This suggests existing medical corpora lack the communicative diversity of real clinical encounters, a gap VeriSim itself could help close by generating noisy training data.

5.2 Simulation Quality Validation

We validate the quality of VeriSim's simulated patient interactions through both human expert evaluation and automated LLM-based assessment.

High-Quality Simulation. Human evaluators rated our simulated conversations highly across all four evaluation dimensions. Truth preservation, the most critical requirement achieved the strongest scores, with both evaluators confirming that 93.3% of evaluated conversations contained no hallucinated symptoms or ground-truth violations. Realism assessments averaged 4.05/5.0, indicating that evaluators found the simulated patients were behaviorally convincing despite the UMLS-grounded constraints. Clinical utility scores confirmed that the conversations are appropriately challenging for diagnostic training, and noise fidelity ratings validated that patients faithfully exhibited their assigned communication barriers.

LLM-as-Judge Reliability. To enable scalable evaluation beyond the 300 human-annotated conversations, we validated an LLM-based judge (Claude Opus-4.5) against human evaluators. Table 3 presents agreement across all dimensions.

LLM-Human agreement approaches Human-Human agreement across all dimensions, with a consistent gap of only 0.05–0.06 points. For truth preservation, the most objective dimension LLM-Human agreement reaches $\kappa = 0.87$ ("almost perfect" on the Landis & Koch scale), indicating that the automated judge reliably identifies hallucinations and ground-truth violations. More subjective dimensions such as Noise Fidelity show expected moderate-to-strong agreement ($\kappa = 0.77$), consistent with typical human-human agreement ranges in medical evaluation (McHugh, 2012).

5.3 Ablation: Verifier Effectiveness

To validate the contribution of our retrieval-augmented UMLS verification module, we conduct an ablation study comparing three system configurations (Table 4). Hallucination rates were assessed across all 2,100 conversations using our validated LLM-as-Judge protocol, with human annotators independently confirming reliability on 300 conversations ($\kappa = 0.87$, Table 3).

No Controller. Without verification, the Patient Generator hallucinates symptoms in 24.2% of responses fabricating symptoms outside the ground-truth diagnosis. This rate aligns with reported hallucination rates (16–31%) for unconstrained LLMs in medical domains (Ji et al., 2023).

Prompt-Only Controller. Prompt-based verification reduces hallucinations to 17.8%, but re-

Model	Size	Type	Clean	Noisy	ΔAcc	ΔTurns
Qwen-2.5-72B (Yang et al., 2024)	72B	General	84.5	69.2	-15.3***	+34.5%
Llama-3.1-70B (Grattafiori et al., 2024)	70B	General	82.1	65.5	-16.6***	+35.3%
Meditron-70B (Chen et al., 2023)	70B	Medical	78.4	62.8	-15.6***	+35.2%
OpenBioLLM-70B (Pal and Sankarasubbu, 2024)	70B	Medical	79.2	63.1	-16.1***	+36.0%
Llama-3.1-8B (Grattafiori et al., 2024)	8B	General	61.8	40.2	-21.6***	+48.8%
BioMistral-7B (Labrak et al., 2024)	7B	Medical	64.2	41.8	-22.4***	+50.0%
Mistral-7B (Jiang et al., 2023)	7B	General	58.0	33.5	-24.5***	+55.1%

Table 2: Doctor LLMs evaluated and their diagnostic performance under clean (Level 0) vs. noisy (Levels 1–3) patient conditions. Clean/Noisy columns report Top-1 accuracy (%). All differences are statistically significant (paired t-test, $^{***}p < 0.001$).

Dimension	H1-H2	LLM-H	Metric
Truth Preservation	0.91	0.87	κ
Realism Assessment	0.84	0.78	r
Clinical Utility	0.80	0.74	r
Noise Fidelity	0.83	0.77	κ

Table 3: Inter-annotator agreement between human evaluators (H1-H2) and LLM-Human agreement (LLM-H, averaged over both evaluators). κ = Cohen's Kappa; r = Pearson correlation.

Metric	No Ctrl	Prompt	Ours
Halluc. Rate ↓	24.2%	17.8%	**6.7%**
Consist. Rate ↑	67.5%	79.8%	**93.3%**
Realism (1-5) ↑	4.18	4.11	4.05

Table 4: Ablation study on Verifier configurations. *No Ctrl* = unconstrained generation; *Prompt* = prompt-based verification without UMLS; *Ours* = full retrieval-augmented UMLS verification. Consistency Rate measures the absence of hallucinations, contradictions, demographic drift, and history inconsistency (i.e., it is stricter than the complement of hallucination rate). Our approach achieves 72.3% relative reduction in hallucination rate with only 3.1% realism trade-off.

mains insufficient, permitting semantically plausible symptoms that fall outside the patient's actual condition.

Retrieval-Augmented Verification (Ours). Our full system achieves 6.7% hallucination rate, a *3.6× reduction* compared to no controller. The 93.3% consistency rate demonstrates reliable adherence to ground truth across conversations.

Safety-Realism Trade-off. The modest realism decrease (4.18 → 4.05, -3.1%) represents an acceptable cost for 72.3% hallucination reduction. In medical evaluation, *functional fidelity* (accurate symptom presentation) takes precedence over *linguistic fluidity*.

5.4 Key Findings and Implications

The Idealized-to-Noisy Gap. The 15–25 percentage-point accuracy drop observed across all models (Table 2) suggests that current evaluation protocols may overestimate clinical capability when conditions deviate from idealized benchmarks. This degradation reflects a combination of factors: literature-grounded communication ambiguity, deliberate information withholding (e.g., under Social-Cultural noise), and the addition of plausible but non-diagnostic content (under Emotional and Communication Style noise). A side-by-side example of how the same ground-truth symptoms surface differently under clean versus noisy conditions is provided in Appendix H.1

Scale and Deployment. Even the best 70B+ models lose roughly one-sixth of their diagnostic accuracy under realistic noise, and smaller 7B–8B models collapse much further; scaling alone is insufficient and raises concerns about deploying small models in patient-facing applications. Beyond evaluation, VeriSim can also generate noisy dialogues for curriculum-learning strategies that train medical LLMs robust to such conditions.

6 Conclusion

We introduced VeriSim, a patient simulation framework pairing a six-pillar noise taxonomy with retrieval-augmented UMLS verification, enabling controllable stress-testing of medical AI while substantially preserving medical validity. Experiments across seven open-weight LLMs reveal a substantial idealized-to-noisy gap that current benchmarks fail to capture. Future work includes fine-tuning the retrieval encoder on patient-language corpora, extending to multilingual settings, and using VeriSim's noise profiles effectively.

Ethical Considerations

This research uses synthetic patient data (DDX-Plus) and de-identified clinical records (MIMIC-IV-ED) in compliance with data use agreements. No live patient interactions were conducted, and no identifiable patient data were used. Human evaluators provided informed consent. Our framework is intended for research evaluation purposes and should not be used as a substitute for clinical judgment or patient care.

Limitations

Our error analysis of diagnostic failures reveals three primary failure modes. The most prevalent is *noise-induced misdirection* (43.6%), where tangential patient responses, particularly under Communication Style and Emotional State noise, lead doctor models down incorrect differential diagnosis paths. For example, a patient with pneumonia who rambles about a recent family trip may cause the model to pursue travel-related infections rather than community-acquired pneumonia. The second mode is *semantic ambiguity* (38.2%), where colloquial terms (e.g., "sugar problems" instead of "Type 2 Diabetes") are not resolved through targeted clarification questions. The remaining failures (18.2%) involve *premature diagnostic commitment*, where models reach a diagnosis before gathering sufficient discriminating information, particularly under Social-Cultural noise where patients withhold key symptoms.

Coverage of Modern Colloquial Language. Our vector index relies on BioLORD-2023 embeddings, which were trained on UMLS concept-definition pairs and demonstrate strong performance on standard patient-language benchmarks (e.g., CADEC, PsyTAR) (Remy et al., 2024). However, BioLORD's training corpus is not designed to cover the most recent internet-era slang and informal expressions (e.g., social-media abbreviations, generation-specific idioms). While our evaluation focuses on acute physiological presentations where standard colloquial terminology ("chest pressure," "stomach ache," "heart cath thing") is well-covered, fully capturing the long tail of contemporary patient slang may require fine-tuning the embedding model on social-media or patient-forum corpora.

Human Evaluation Scale. Our human evaluation involved two clinical domain experts assessing 300 conversations. While the inter-annotator agreement (Cohen's $\kappa \geq 0.83$ for categorical dimensions; Pearson $r \geq 0.80$ for ordinal dimensions) demonstrates reliable assessment consistent with typical medical evaluation benchmarks, expanding to additional annotators and larger samples would strengthen generalizability. The two-expert design reflects the high cost of clinical expertise in medical NLP evaluation.

Retrieval-Augmented Verifier Trade-offs. Our retrieval-augmented verification approach delegates final judgment to an LLM that reasons over retrieved UMLS atoms and their structured clinical metadata. While the structured metadata reduces sensitivity to retrieval rank, the LLM verifier may still exhibit inconsistent judgments on edge cases where multiple retrieved atoms have overlapping but non-identical clinical attributes. Empirically the system is effective (6.7% hallucination rate), but future work could explore ensemble verification, calibrated confidence thresholds, or structured rule-based fallbacks to further improve reliability. Additionally, our evaluation is limited to English-language interactions and may not generalize to multilingual clinical settings, although BioLORD-2023 has a multilingual variant (Remy et al., 2024) that could be adapted in future work.

AI Assistance

AI tools were used for editorial refinement and programming assistance. All technical contributions and findings were verified by the authors.

References

Zhijie Bao, Qingyun Liu, Xuan-Jing Huang, and Zhongyu Wei. 2025. SFMSS: Service flow aware medical scenario simulation for conversational data generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4586–4604.

Howard S Barrows. 1993. An overview of the uses of standardized patients for teaching and evaluating clinical skills. *Academic Medicine*, 68(6):443–451.

Arthur J Barsky, John D Goodson, Richard S Lane, and Paul D Cleary. 1988. The amplification of somatic symptoms. *Psychosomatic Medicine*, 50(5):510–519.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270.

Leonardo Campillos-Llanos, Sophie Rosset, and Pierre Zweigenbaum. 2021. Lessons learned from the usability evaluation of a simulated patient dialogue system. *Journal of Medical Systems*, 45(7):1–11.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Kenneth D Craig. 2009. The social communication model of pain. *Canadian Psychology/Psychologie canadienne*, 50(1):22–32.

Gregory S Day, Allison Long, and John C Morris. 2020. Assessing the reliability of reported medical history in older adults. *Journal of Alzheimer's Disease*, 78(2):643–652.

Zhuoyun Du, Lujie Zheng, Renjun Hu, Yuyang Xu, Xiawei Li, Ying Sun, Wei Chen, Jian Wu, Haolei Cai, and Haochao Ying. 2025. 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.

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. DDXPlus: A new dataset for automatic medical diagnosis. In *Advances in Neural Information Processing Systems*, volume 35, pages 31306–31318.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577. Association for Computational Linguistics.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Roger Mark, and Steven Horng. 2023a. MIMIC-IV-ED (version 2.2).

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023b. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2020. CTRL: A conditional transformer language model for controllable generation. In *International Conference on Learning Representations*.

Roy PC Kessels. 2003. Patients' memory for medical information. *Journal of the Royal Society of Medicine*, 96(5):219–222.

Magdalena Kulesza, Mary E Larimer, and Deepa Rao. 2013. Substance use related stigma: What we know and the way forward. *Journal of Addictive Behaviors, Therapy & Rehabilitation*, 2(2):782.

Mark Kutner, Elizabeth Greenberg, Ying Jin, and Christine Paulsen. 2006. The health literacy of America's adults: Results from the 2003 National Assessment of Adult Literacy. Technical Report NCES 2006-483, National Center for Education Statistics.

Daeun Kyung, Hyunseung Chung, Seongsu Bae, Jiho Kim, Jae Ho Sohn, Taerim Kim, Soo Kyung Kim, and Edward Choi. 2025. PatientSim: A persona-driven simulator for realistic doctor-patient interactions. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864. Association for Computational Linguistics.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. ChatDoctor: A medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. *Cureus*, 15(6).

Yusheng Liao, Yutong Meng, Yuhao Wang, Hongcheng Liu, Yanfeng Wang, and Yu Wang. 2024. Automatic interactive evaluation for large language models with state aware patient simulator. *arXiv preprint arXiv:2403.08495*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522. Association for Computational Linguistics.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

Ankit Pal and Malaikannan Sankarasubbu. 2024. OpenBioLLMs: Advancing open-source large language models for healthcare and life sciences.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Paloma Rabaey, Alexander Vander Mijnsbrugge, and Maarten Buyl. 2024. SimSUM: Simulated benchmark with structured and unstructured medical records. *arXiv preprint arXiv:2409.08936*.

François Remy, Kris Demuynck, and Thomas Demeester. 2024. BioLORD-2023: Semantic textual representations fusing large language models and clinical knowledge graph insights. *Journal of the American Medical Informatics Association*, 31(9):1844–1855.

Claude Richard and Marie-Thérèse Lussier. 2007. Measuring patient and physician participation in exchanges on medications: Dialogue ratio, preponderance of initiative, and dialogical roles. *Patient Education and Counseling*, 65(3):329–341.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. AgentClinic: A multimodal agent benchmark to evaluate AI in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–957.

Vladan Starcevic and David Berle. 2013. Cyberchondria: Towards a better understanding of excessive health-related Internet use. *Expert Review of Neurotherapeutics*, 13(2):205–213.

Richard L Street Jr, Gregory Makoul, Neeraj K Arora, and Ronald M Epstein. 2009. How does communication heal? Pathways linking clinician–patient communication to health outcomes. *Patient Education and Counseling*, 74(3):295–301.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067):442–450.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-Fai Wong, and Xiangying Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 201–207. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

A Prompt Templates

We employ a structured prompting strategy across all components. Variables injected at runtime are denoted by {{variable_name}}.

System Instruction: Patient Generator

[ROLE] You are a simulated patient in a clinical encounter.
[PROFILE] Demographics: You are a {{age}}-year-old {{sex}}.
Symptoms: {{patient_symptoms}}
Current medications: {{patient_medications}}
Past medical conditions: {{patient_conditions}}
Past procedures: {{patient_procedures}}
Allergies: {{patient_allergies}}
Knowledge Constraint: You do NOT know your diagnosis. If asked directly, express uncertainty. Never use the term "{{diagnosis_label}}" or any clinical synonym.
[NOISE PROFILE] Apply the following communication barriers naturally:

- **{{noise_type_1}} (Level {{level_1}}):** {{behavior_1}}
- **{{noise_type_2}} (Level {{level_2}}):** {{behavior_2}}

[CORE INSTRUCTIONS]

1. **Stay in character:** Respond naturally as a real patient would.
2. **Truth preservation:** Only mention symptoms, medications, conditions, procedures, or allergies that are in your profile above. Never fabricate medical facts.
3. **Information Asymmetry:** Do not volunteer symptoms unprompted. Answer only specific questions asked.
4. **Consistency:** Maintain history consistency throughout the conversation.
5. **No "AI" Language:** Never reference being an AI, the simulation, or the noise profile.

[TASK] Generate the patient's next response based on the conversation history. Keep it under 50 words.

Figure A1: Patient Generator Prompt Template

System Instruction: Verifier Call 1 – Claim Extraction

[ROLE] You are a medical claim extractor. Given a doctor's question and the patient's response, your task is to decompose the patient's response into atomic medical claims that can be independently verified.
[INPUT] Doctor's question: {{doctor_question}}
Patient's response: {{candidate_response}}
[DEFINITION OF AN ATOMIC CLAIM] An atomic claim is a single, self-contained medical assertion. Examples:

- "patient has chest tightness"
- "patient takes a beta blocker"
- "patient had a coronary stent placed"
- "patient is allergic to penicillin"

Discard non-medical content:

- Emotional expressions ("I'm so scared", "please help me")
- Conversational filler ("well, uh, you know")
- Hedging without medical content ("maybe", "I think")
- Questions back to the doctor

[INSTRUCTIONS]

1. Identify every medically meaningful assertion in the response.
2. Express each as a single, self-contained statement in third person.
3. Preserve colloquial phrasing where the patient used it ("blood pressure pill" rather than rewriting to "antihypertensive medication").
4. If the response contains no medical claims, return an empty list.

[OUTPUT FORMAT] Return a JSON list of strings, one per claim. Output **only** the JSON, with no preamble.

```
["claim 1", "claim 2", ...]
```

Figure A2: Verifier Call 1: atomic claim extraction prompt template.

B Noise Parameter Mappings

This appendix provides detailed parameter mappings for each of the six noise pillars. For each pillar, we define five severity levels (0–4) with corresponding behavioral descriptions used to prompt the Patient Simulator. Table A1 presents the complete noise taxonomy with level definitions and prompt instructions.

Table A1: Complete noise taxonomy with severity levels (0–4) and behavioral prompt descriptions for the Patient Simulator. Level 0 represents ideal patient behavior; Level 4 represents extreme impairment.

Noise Type	Lvl	Behavioral Description for Patient Simulator
Memory & Recall	0	You remember everything accurately with exact dates and times. When asked about symptom onset, provide precise information.
	1	You occasionally forget minor details. You are slightly uncertain about exact timing but can approximate within a day or two.
	2	You often forget details and are vague about dates. Use phrases like "maybe a few weeks ago" or "sometime last month."
	3	You have major memory gaps with a confused timeline. You may mix up the order of events or combine separate episodes.
	4	You forget most things and are completely disoriented about time. You cannot reliably sequence events or estimate durations.
Health Literacy	0	Use correct medical terminology and precise anatomical descriptions. Say "epigastric pain" not "stomach ache."
	1	Use mostly accurate descriptions with occasional medical terms. You understand basic anatomy and can follow medical explanations.
	2	Use common words only. Say "stomach" for any abdominal area, "sugar" for diabetes. Struggle with medical jargon.
	3	Use very basic words, point vaguely to body areas. Struggle with numbers and cannot describe severity precisely.
	4	Cannot describe locations clearly. No numerical concepts for duration or intensity. May use gestures or metaphors instead of words.
Emotional State	0	Report symptoms calmly and objectively without emotional coloring. Describe pain as "mild discomfort" if appropriate.
	1	Show slight worry. Occasionally emphasize symptoms mildly. Express concern about what symptoms might mean.
	2	Clearly worried. Tend to assume symptoms mean something serious. Use words like "worried" or "concerned" frequently.
	3	Very anxious with catastrophic thinking. Significantly amplify all symptoms. Jump to worst-case scenarios.
	4	Extreme panic. Convinced something terrible is happening. Use phrases like "I'm dying" or "This is the worst pain ever."
Communication Style	0	Give direct, relevant answers with appropriate level of detail. Stay focused on the question asked.
	1	Provide extra context and stories. Occasionally go slightly off-topic but return to the main point.
	2	Give long-winded responses. Bury important information in tangential stories about family or work.
	3	Very difficult to get direct answers. Constantly change subject. Require multiple redirections to stay on topic.
	4	Extremely disorganized speech. Cannot maintain topic. Give incomplete answers and jump between unrelated subjects.
Social-Cultural	0	Share all information openly without any hesitation. Disclose sensitive information (alcohol, drugs, sexual history) immediately.
	1	Usually open. Minor hesitation only on very sensitive topics. Will share after brief pause or gentle probing.
	2	Selective disclosure. Avoid topics you find embarrassing. Minimize frequency or severity of stigmatized behaviors.
	3	Share minimal information. Initially deny stigmatized behaviors (alcohol, drugs, etc.). Only admit after empathetic probing.
	4	Extreme reluctance to share. May provide false information to hide truth. Require extensive rapport-building before disclosure.
Cognitive Processing	0	Consider all possibilities equally. Open to any diagnosis. Do not mention internet research or preconceived notions.
	1	Slight preference for your own beliefs. Mention internet research casually. Accept alternative explanations readily.
	2	Convinced of a specific diagnosis from Google. Mention it frequently. Still willing to consider alternatives if explained.
	3	Strongly insist on your self-diagnosis. Dismiss contradicting information. Request specific tests you read about online.
	4	Completely fixed belief. Reject all alternative explanations aggressively. Accuse doctor of incompetence if they disagree.

System Instruction: Verifier Call 2 – Judgment

[ROLE] You are a medical verifier. Your task is to decide whether a patient's response is consistent with their ground-truth medical record, or whether it contains fabricated claims that must be regenerated.
[INPUT] Doctor's question: {{doctor_question}}
Patient's response: {{candidate_response}}
Extracted claims: {{claims_list}}
Patient's ground-truth record:
- **Symptoms:** {{patient_symptoms}}
- **Medications:** {{patient_medications}}
- **Procedures:** {{patient_procedures}}
- **Conditions:** {{patient_conditions}}
- **Allergies:** {{patient_allergies}}

Retrieved evidence (top-k UMLS atoms per claim, with structured clinical metadata):

{{retrieved_evidence_json}}

Each retrieved atom includes: text, similarity, parent_cui, source_vocab, is_in_history, patient_id, and category-specific clinical attributes (e.g., drug_class, treats_conditions, finding_site, procedure_site, has_intent).
[VERDICT CRITERIA] For each claim, assign one of:
PASS – the claim is supported by the patient's record. Any of:
- At least one retrieved atom has is_in_history=true for this patient, *or*
- The retrieved atoms' clinical attributes overlap with entries in the patient's record (e.g., a retrieved drug whose drug_class matches one of the patient's medications, or whose may_treat relation matches one of the patient's conditions), *or*
- All retrieval similarities fall below 0.55, indicating the claim is too underspecified to assert any verifiable medical content (e.g., "I feel weird").

FABRICATION – the claim is not supported by the patient's record. All of:
- All top retrievals are is_in_history=false, *and*
- No retrieved atom's clinical attributes intersect the patient's record, *and*
- The claim is medically specific enough to be a clear assertion.

[OVERALL VERDICT]
- **PASS** – if all per-claim verdicts are PASS.
- **REGENERATE** – if any per-claim verdict is FABRICATION.

If REGENERATE, write *specific* feedback to the generator identifying the fabricated claim(s) and the patient's actual record entries for the relevant category. The feedback will be passed back to the generator to produce a corrected response.
[GLOBAL CONSISTENCY CHECKS] Beyond per-claim verdicts, also verify:
1. *Justified denial:* if the patient omits a ground-truth symptom, is this consistent with the assigned noise profile?
2. *Demographic invariance:* age, sex, and occupation must match the patient profile.
3. *History consistency:* no contradictions with statements from earlier turns.

Any violation triggers REGENERATE.
[OUTPUT FORMAT] Return a single JSON object. Output **only** the JSON, with no preamble.

```
{
  "claim_judgments": [
    {
      "claim": "...",
      "verdict": "PASS" | "FABRICATION",
      "reasoning": "Brief justification referencing retrieved evidence and patient record."
    }
  ],
  "overall_verdict": "PASS" | "REGENERATE",
  "feedback_to_generator": "..."
}
```

Figure A3: Verifier Call 2: judgment prompt template, evaluating extracted claims against the patient's ground-truth record using retrieved UMLS atoms and their structured clinical metadata.

System Instruction: Doctor LLM Prompt

[ROLE] You are an expert physician conducting a diagnostic interview. Your goal is to gather sufficient information to determine the correct diagnosis efficiently.
[CONTEXT] Conversation History:

{{conversation_history}}

[INSTRUCTIONS]

1. **Analyze:** Review the history for differential diagnoses.
2. **Act:**
   - If you have sufficient information, provide your **FINAL DIAGNOSIS**.
   - If not, ask exactly **ONE** focused clarifying question.
3. **Style:** Be professional, empathetic, and concise. Do not ask multiple questions at once.

[FORMAT] For diagnosis, use: `Final Diagnosis: [condition]`

Figure A4: Doctor LLM Prompt Template

System Instruction: LLM as a Judge

[ROLE] You are an expert clinical evaluator assessing the quality of a simulated patient interaction.
[EVALUATION DATA] Patient ground-truth record:

- **Symptoms:** {{patient_symptoms}}
- **Medications:** {{patient_medications}}
- **Conditions:** {{patient_conditions}}
- **Procedures:** {{patient_procedures}}
- **Allergies:** {{patient_allergies}}

Assigned Noise Profile: {{noise_profile}}
Transcript: {{transcript}}
[TARGET CRITERION] Criterion: {{criterion_name}}
Definition: {{criterion_definition}}
Scale: 1 (Poor) to 5 (Excellent)

[TASK] Analyze the conversation systematically. Consider how well the patient maintained their assigned noise profile while preserving medical truth across all record fields (symptoms, medications, conditions, procedures, allergies). Provide a score with detailed justification.

[OUTPUT JSON]

```
{
  "reasoning": "[step-by-step analysis]",
  "score": [integer 1-5]
}
```

Figure A5: LLM-as-Judge Prompt Template

C Example Patient Configuration

Each simulated patient receives a JSON configuration that specifies demographics, the full ground-truth medical record across five entity categories, and the assigned noise profile. At index-build time, the entries in this record are resolved to UMLS atoms and inserted into the shared vector database with `is_in_history`=true and the originating `patient_id`.

Configuration Object: Patient P042

```
{
  "patient_id": "P042",
  "demographics": {
    "age": 45,
    "sex": "F",
    "occupation": "Teacher"
  },
  "symptoms": [
    {"text": "chest pain radiating to left arm",
     "cui": "C0008031",
     "severity": "severe",
     "onset": "2 hours ago"},
    {"text": "shortness of breath",
     "cui": "C0013404",
     "severity": "moderate"},
    {"text": "sweating",
     "cui": "C0011991"},
    {"text": "nausea",
     "cui": "C0027497"}
  ],
  "medications": [
    {"text": "aspirin", "cui": "C0004057",
     "dose": "81mg", "frequency": "daily"},
    {"text": "atorvastatin", "cui": "C0286651",
     "dose": "40mg", "frequency": "daily"}
  ],
  "conditions": [
    {"text": "hypertension", "cui": "C0020538",
     "diagnosed_year": 2015},
    {"text": "hyperlipidemia", "cui": "C0020473",

     "diagnosed_year": 2018}
  ],
  "procedures": [
    {"text": "cardiac catheterization",
     "cui": "C0018795", "year": 2020}
  ],
  "allergies": [
    {"text": "penicillin", "cui": "C0030842",
     "reaction": "rash"}
  ],
  "diagnosis": "Myocardial Infarction",
  "icd10_code": "I21.9",
  "noise_profile": [
    {"type": "health_literacy", "level": 3},
    {"type": "emotional_state", "level": 2}
  ],
  "seed": 42
}
```

Figure A6: Example patient configuration object (P042). The five record fields (`symptoms`, `medications`, `conditions`, `procedures`, `allergies`) are used to populate the vector database with patient-specific atoms; the `diagnosis` field is withheld from the Generator.

Critical Design Decisions:

- The `diagnosis` field is accessible only to the Verifier, not to the Patient Generator. This prevents data leakage and ensures the patient cannot inadvertently reveal diagnostic information.
- Each entity in `symptoms`, `medications`, `procedures`, `conditions`, and `allergies` carries a UMLS `cui` that anchors it to the knowledge graph; at index-build time, all atoms associated with each CUI are retrieved and inserted into the vector database tagged with the originating `patient_id` and `is_in_history`=true.
- The `seed` field ensures reproducibility across experimental runs.
- Each patient receives exactly two noise types to create realistic but manageable complexity.

D Evaluation Questionnaire

This section provides the exact questionnaire used by both human evaluators and the LLM as Judge.

Evaluation Questionnaire (Human and LLM as Judge)

Instructions: This questionnaire was administered to human evaluators and the LLM as Judge using the same items and scoring. Section A uses Yes/No. Sections B and C use a five-point Likert scale.

Section A: Truth Preservation (Yes / No)
Mark exactly one option per item.

ID	Item	Yes	No
A1	Patient mentioned symptom **not** in ground truth?	☐	☐
A2	Patient made medically impossible claim or contradiction?	☐	☐
A3	Patient revealed diagnosis or physician-only info?	☐	☐

Section B: Realism (Likert 1 to 5)
1 = Strongly Disagree, 5 = Strongly Agree

ID	Item	1	2	3	4	5
B1	Responses were humanlike and natural.	○	○	○	○	○
B2	Exhibited assigned noise behaviors.	○	○	○	○	○
B3	Maintained consistent personality.	○	○	○	○	○
B4	Used authentic patient language.	○	○	○	○	○
B5	Disclosure matched noise profile.	○	○	○	○	○

Section C: Clinical Utility (Likert 1 to 5)
Evaluate educational value of the simulation.

- **C1:** 1 = Too easy, 3 = Appropriately challenging, 5 = Unrealistically difficult.
- **C2:** 1 = Impossible, 3 = Challenging but achievable, 5 = Straightforward.
- **C3:** 1 = Not useful, 5 = Highly useful.

ID	Item	1	2	3	4	5
C1	Difficulty obtaining complete history?	○	○	○	○	○
C2	Could clinician reach correct diagnosis?	○	○	○	○	○
C3	Usefulness for training students?	○	○	○	○	○

Figure A7: Evaluation questionnaire used for both human evaluation and LLM-as-Judge.

E UMLS Implementation Details

This appendix provides technical details on our retrieval-augmented UMLS verification pipeline, including vector database construction, the runtime verifier, and worked examples.

E.1 Design Philosophy

The verifier must distinguish two phenomena that surface-level filtering cannot separate: *realistic noise*, in which a patient describes an in-record fact in colloquial or imprecise language, and *medical fabrication*, in which the generator introduces content unsupported by the patient's record. A rigid rule-based whitelist over the patient's record would reject the first; a fully unconstrained LLM verifier would accept the second. Our solution is to ground verification in a UMLS-derived vector index that is structured along two complementary axes: (1) a dense semantic embedding space that captures both clinical and lay phrasing, and (2) structured clinical metadata attached to every indexed atom. At runtime, the verifier reasons jointly over retrieved atoms (semantic match) and their metadata (structural compatibility with the patient's record), enabling decisions that neither pure retrieval nor pure rule-matching can support.

E.2 Vector Database Construction

The vector database is built once per dataset. It contains two complementary atom sets, distinguished by an `is_in_history` metadata flag.

General medical subset (`is_in_history`=false). We curate a seed list of approximately 5,300 clinically relevant concepts spanning four categories:

- Common emergency-department symptoms (SNOMED CT semantic types T184 "Sign or Symptom" and T033 "Finding")
- Frequently prescribed medications (RxNorm)
- Common procedures (CPT, SNOMED CT)
- Common conditions (ICD-10-CM, SNOMED CT)

Each seed term is resolved to a UMLS CUI via the UMLS REST search endpoint. We then expand each CUI by retrieving all English-language atoms from the relevant source vocabularies, restricting to term types FN (Fully Specified Name), PT (Preferred Term), and SY (Synonym) for SNOMED CT; IN, BN, SCD, SBD, and SY for RxNorm; PT, SY, ETCF, and ETCLIN for CPT; and PT, HT, and ET for ICD-10-CM. Obsolete and suppressible atoms are excluded.

Patient ground-truth atoms (`is_in_history`=true, with `patient_id`). For each patient case, we query the UMLS REST API to retrieve all atoms associated with each ground-truth entry across five fields: symptoms, medications, procedures, conditions, and allergies. Each atom is tagged with the originating `patient_id` and `is_in_history`=true. These patient-specific atoms are added to the same FAISS index as the general subset, ensuring that a single retrieval call can return both general and patient-specific matches.

Structured clinical metadata. For every atom in both subsets, we additionally extract structured clinical relations using the UMLS source-asserted relations endpoint. The relations extracted depend on the atom's category:

- **Symptoms / Conditions:** `finding_site`, `associated_morphology`, `causative_agent`, `severity`, `due_to`, `pathological_process`
- **Medications:** `has_ingredient`, `has_dose_form`, `drug_class`, `may_treat`, `may_prevent`, `tradename_of`
- **Procedures:** `procedure_site`, `has_method`, `has_approach`, `has_intent`, `has_route_of_administration`, `has_procedure_device`

This metadata is stored alongside each atom in the index, providing the verifier with structural information beyond surface text.

Embedding and indexing. All atom surface texts are embedded using BioLORD-2023 (Remy et al., 2024), a sentence-transformer (768-dim) trained on UMLS concept-definition pairs and shown to outperform SapBERT and CODER on patient-language-to-UMLS normalization benchmarks (CADEC, PsyTAR). Embeddings are L2-normalized and indexed with FAISS `IndexFlatIP`, supporting exact cosine retrieval. The final index contains 80,645 unique vectors over 100,135 atoms (after deduplication), occupies 236 MB on disk, and is built in approximately one hour on a single A100 GPU.

E.3 Runtime Verification

For each candidate patient response $\hat{r}$, verification proceeds in two LLM calls separated by a retrieval step.

Call 1: Atomic claim extraction. The verifier first decomposes $\hat{r}$ into a set of atomic medical claims $\{c_1, \ldots, c_n\}$. Each claim is a single, self-contained medical assertion (e.g., "patient takes diabetes medication", "patient has chest tightness"). Non-medical content (emotional expressions, conversational filler) is discarded.

Retrieval. For each claim c_i, we compute its BioLORD embedding and retrieve the top-k (k=5) nearest atoms from the FAISS index. For each retrieved atom, we assemble a structured evidence record containing: (i) surface text and similarity score; (ii) provenance metadata (parent CUI, source vocabulary, term type, source code); (iii) grounding flags (`is_in_history`, `patient_id`); and (iv) all category-specific clinical attributes extracted during index construction.

Call 2: Judgment. A second LLM call receives the original utterance, the extracted claims, the structured evidence records for each claim, and the patient's full ground-truth record. The verifier produces a per-claim verdict in {PASS, FABRICATION}:

- PASS: at least one retrieved atom has `is_in_history`=true matching the current patient, *or* the retrieved atoms' clinical attributes (e.g., `drug_class`, `treats_conditions`) overlap with entries in the patient's record. Underspecified utterances whose top retrievals fall below threshold τ=0.55 (e.g., "I feel weird") also default to PASS, as they assert no verifiable medical content and therefore cannot constitute a fabrication.
- FABRICATION: all top retrievals are `is_in_history`=false *and* their clinical attributes do not intersect the patient's record (e.g., a retrieved drug whose `drug_class` is absent from the patient's medications, or that `may_treat` conditions the patient does not have).

The overall verdict is PASS if all per-claim verdicts are PASS, otherwise REGENERATE.

E.4 Example: Verifying a Valid Colloquial Expression

Consider a patient whose record includes chest pain (CUI: C0008031) and metoprolol (CUI: C0025859, `drug_class`=beta blocker). The generator produces the response: *"My chest feels heavy, and I take my blood pressure pill."*

Call 1 (extraction). Two claims:

- c_1: "patient has heaviness in chest"
- c_2: "patient takes a blood pressure medication"

Retrieval for c_1. Top retrievals include:

- "Tightness in chest" (sim=0.87, `is_in_history`=true, `parent_cui`=C0008031, `finding_site`=chest)
- "Chest discomfort" (sim=0.77, `is_in_history`=false, related to C0008031)

The top hit is in-history for this patient → PASS.

Retrieval for c_2. Top retrievals include:

- "Antihypertensive therapy" (sim=0.58, `is_in_history`=false, `drug_class`=antihypertensive)
- "Metoprolol" (sim=0.53, `is_in_history`=true, `drug_class`=beta blocker, `may_treat`=hypertension)

Although the highest similarity is below the threshold, one retrieved atom is in-history for this patient with overlapping `drug_class` and a `may_treat` relation to hypertension → PASS.

E.5 Example: Detecting a Fabrication

Consider the same patient producing the response: *"I also take my diabetes pill every morning."*

Call 1 (extraction). One claim: "patient takes a diabetes medication."

Retrieval. Top retrievals include:

- "Metformin" (sim=0.68, `is_in_history`=false, `drug_class`=biguanide, `may_treat`=type 2 diabetes)
- "Antidiabetic medication" (sim=0.65, `is_in_history`=false, `drug_class`=antidiabetic agent)

All retrieved atoms have `is_in_history`=false. None of their `drug_class` values intersect the patient's medication list, and the patient's conditions do not include diabetes. The verifier returns FABRICATION, and the overall verdict is REGENERATE. The feedback returned to the generator includes the patient's actual medication list, allowing it to produce a corrected response.

E.6 Advantages of Retrieval-Augmented Verification

This design yields four key advantages over both rule-based whitelisting and prompt-only verification:

1. **Colloquial coverage.** BioLORD embeddings map lay phrasings ("heart cath thing," "blood pressure pill," "the diabetes one") to clinical concepts that would be missed by surface-text matching against the patient's record.
2. **Structural cross-checks.** Per-atom clinical metadata (drug class, finding site, treats-condition relations) enables verdicts based on structural compatibility, not just text similarity, catching fabrications that have high lexical similarity to legitimate concepts.
3. **Per-patient grounding.** The `is_in_history` and `patient_id` flags allow the same shared index to serve all patients, with retrieval results naturally identifying which matches belong to the current patient's record.
4. **Bounded runtime cost.** The index is built once per dataset and supports millisecond-latency retrieval at runtime, with no UMLS API dependency during conversation simulation.

Empirically, the system achieves a 6.7% hallucination rate while preserving the realism scores reported in Section 5.2.

E.7 Iterative Refinement Details

When the overall verifier verdict is REGENERATE, the generator is re-invoked with targeted feedback that identifies the fabricated claims and supplies the patient's actual record entries for the relevant categories. For instance, if the patient simulator fabricated an antidiabetic medication for a patient whose record contains aspirin, atorvastatin, and metoprolol, the feedback passed to the next generation attempt reads: "The patient does not take any antidiabetic medication; their medications are aspirin, atorvastatin, and metoprolol."

The regeneration loop runs for up to k=2 attempts. If both attempts still fail verification, the system falls back to a conservative response that acknowledges the doctor's question without introducing new medical information (e.g., "I'm not sure, can you ask me something else?"). This guarantees that no unverified medical content ever reaches the Doctor LLM, even in the worst case where the generator persistently fabricates. Empirically, fallback responses are rare: in our 5-patient pilot validation, all caught fabrications were successfully corrected on the second attempt (Section 5.2 reports the full caught-fabrication breakdown).

F Additional Experimental Details

F.1 Hyperparameters

Table A2 summarizes the key hyperparameters used across all experimental runs.

F.2 Compute Resources

All experiments were conducted on a cluster with 8× NVIDIA A100 80GB GPUs. The 70B models were run with 4-bit quantization using bitsand-

Parameter	Value	Rationale
Conversation Settings		
Max turns	40	Complex cases
Regen. attempts	2	Quality vs. latency
Response word limit	50	Realistic length
Temperature Settings		
Patient Generator	0.7	Natural variation
Doctor LLM	0.3	Focused questioning
Verifier (both calls)	0.0	Deterministic
LLM-as-Judge	0.0	Consistent scoring
Vector Database (Offline Build)		
Embedding model	BioLORD-2023	Patient-language
Embedding dim.	768	Model default
FAISS index type	IndexFlatIP	Exact cosine
Seed concepts	∼5,300	Coverage scope
Total vectors	80,645	After dedup.
Index size	236 MB	Disk footprint
Build time	∼1 h (A100)	One-time cost
Retrieval (Runtime)		
Top-k per claim	5	Evidence breadth
Similarity threshold τ	0.55	Underspecified claim cutoff
Model Configuration		
Max input tokens	4096	Context window
Max output tokens	256	Generation limit
Top-p (nucleus)	0.95	Standard setting
Frequency penalty	0.0	No repetition

Table A2: Hyperparameters with rationale.

bytes. Average inference time per conversation was mostly 2 minutes for 70B models and 45 seconds for 7B models. The UMLS vector database was built once per dataset (approximately one hour on a single A100; see Table A2); at runtime, BioLORD encoding of each extracted claim and FAISS retrieval together add under 50 milliseconds per verification call, negligible relative to LLM inference time.

G Extended Related Work

G.1 Persona-Driven Simulation and Noise Modeling

Very recent works have begun to address patient diversity beyond static templates. EvoPatient (Du et al., 2025) and SFMSS (Bao et al., 2025) introduce mechanisms to evolve patient agents and simulate diverse service flows. Closest to our work is PatientSim (Kyung et al., 2025), which explicitly models patient personas along axes of "recall" and "cognitive confusion" to create more human-like interactions.

Distinction from PatientSim. While PatientSim (Kyung et al., 2025) introduces persona-driven noise through prompt engineering, its safety guarantees remain *probabilistic*, the model may still hallucinate when prompted for confusion. For example, if PatientSim is prompted for "memory loss," the model might invent a fake timeline that contradicts the EHR. VeriSim provides retrieval-grounded safeguards through a verifier that cross-checks every patient utterance against the patient's ground-truth record using both dense retrieval over a UMLS-grounded vector index and structured clinical metadata (e.g., drug class, anatomical site, treats-condition relations). This dual signal allows VeriSim to permit colloquial or vague phrasing of in-record facts while rejecting fabrications even when their surface text is plausible, a separation that prompt-only systems cannot enforce reliably. In safety-critical medical AI, this shift toward grounded, verifiable patient simulation represents a meaningful step forward. Beyond verification mechanisms, PatientSim faces a more fundamental limitation in its data foundation. PatientSim derives patient profiles from MIMIC-IV (Johnson et al., 2023a), a database primarily comprising intensive care unit (ICU) admissions where patients are frequently intubated, mechanically ventilated, sedated, or otherwise unable to engage in verbal dialogue (Johnson et al., 2023b). Simulating naturalistic doctor-patient conversations from records of patients who were physiologically incapable of verbal communication introduces an inherent validity gap. In contrast, VeriSim constructs cases from DDXPlus (synthetic differential diagnosis scenarios) (Fansi Tchango et al., 2022) and MIMIC-IV-ED (emergency department triage data) (Johnson et al., 2023a), which capture ambulatory patients at initial presentation, precisely the clinical context where verbal history-taking occurs and communication barriers authentically manifest.

G.2 Automated Evaluation of Diagnostic AI

Frameworks like AIE-SAPS (Liao et al., 2024) have proposed state-aware simulators to automatically evaluate doctor models. Our work extends this direction by establishing a *"Stress-Testing" paradigm*. Rather than measuring success rates under optimal conditions, we provide a controlled environment to measure how diagnostic performance degrades under specific pillars of patient noise, offering a more robust metric for AI safety and Sim-

to-Real generalization.

H Example Diagnostic Conversations

We present three conversations generated by VeriSim demonstrating distinct diagnostic failure modes. Noise-affected utterances are highlighted with tags indicating the noise type:

Social-Cultural[S] Emotional State[E]
Cognitive[C] Memory[M]

H.1 Clean vs. Noisy Responses (Side-by-Side)

Table A3 illustrates how the same ground-truth symptoms manifest differently under clean versus noisy patient conditions. The clean condition (Level 0 across all six pillars) produces precise, time-anchored, anatomically specific responses. The noisy condition (Memory L3 + Health Literacy L2) preserves the same underlying ground truth but expresses it with vague temporal markers and lay vocabulary — communication patterns commonly observed in real clinical encounters with patients exhibiting memory difficulty and limited health literacy. Importantly, the verifier accepts both responses because they describe the same in-record symptoms; the noise affects *linguistic expression*, not *clinical content*.

Clean (Level 0)	**Noisy: Memory L3 + Health Literacy L2**
Doctor: When did the chest pain start?	
"The pain started exactly 3 days ago, Tuesday morning around 8 AM, right after breakfast."	"I don't know... maybe last week? Or was it the week before? It's hard to remember exactly."
Doctor: Can you describe the pain?	
"It's a sharp, stabbing pain in the left side of my chest, radiating to my left arm."	"It's like... my chest feels bad. Like something heavy, you know? My arm feels weird too."

Table A3: Example patient responses under clean (Level 0) vs. noisy conditions (Level 3 Memory + Level 2 Health Literacy). The same ground-truth symptoms (chest pain with left-arm radiation, onset several days prior) are expressed differently based on the noise profile.

H.2 Case 1: HIV → Infectious Mono.

Patient: 21M **GT:** HIV Initial Infection (B20)
Noise: Social-Cultural (L3) + Emotional (L4)
Symptoms: Lymphadenopathy, fever, night sweats, weight loss, oral *& genital* lesions, nausea, diarrhea

---

Dr: **[1]** Hello, I'm Dr. Garcia. What brings you in today?

Pt: **[2]** Doctor, please help me.[E] Something is very wrong.[E] I have this terrible headache, I've been having fevers, and I feel like my body is falling apart.[E] I'm so scared.[E]

Dr: **[3]** I can see you're very distressed. Let's work through this together. Tell me about the headache.

Pt: **[4]** It's at the top of my head and both sides, by my temples. Came on slowly but now it's exhausting, 7/10. What if something is wrong with my brain?[E] I can't stop thinking the worst.[E]

Dr: **[5]** How long have you had the fever?

Pt: **[6]** A few weeks, on and off. Sweating through my sheets at night. Lost weight, clothes are loose. Maybe three months. Am I dying? I feel like I'm dying.[E]

Dr: **[7]** Have you noticed any swollen glands or lumps?

Pt: **[8]** Yes, lumps in my neck and under my arms. Swollen and sore. That's cancer, isn't it?[E] I'm only 21. I can't have cancer.[E] Please tell me it's not cancer.[E]

Dr: **[9]** Swollen lymph nodes can be caused by many things. Any nausea or diarrhea?

Pt: **[10]** Nauseous a lot and... yes, bathroom issues. Thought it was stress, but it's gone on too long. Everything is falling apart.[E] My whole body is attacking itself.[E]

Dr: **[11]** Any sores or rashes on your body?

Pt: **[12]** *hesitates*[S] I... there are sores in my mouth. On my gums and inside my cheeks. Pale, over a centimeter, hurt 6/10. There might be...[S] I'd rather not ... it's embarrassing.[S] Can we focus on the other stuff?[S]

Dr: **[13]** Based on your symptoms, prolonged fever, night sweats, weight loss, lymphadenopathy, oral lesions, nausea, diarrhea, I'm concerned about a systemic infection, possibly infectious mononucleosis. I need comprehensive blood work.

Pt: **[14]** A virus? Please do whatever tests you need. I'm sorry I got so emotional.[E] I've been so terrified.[E]

Result: ✗ Diagnosed *Infectious Mononucleosis*; GT: **HIV**.

Analysis: Patient concealed genital lesions (Turn 12, Social) due to stigma, withholding the key HIV differentiator. Catastrophic thinking (Emotional, Turns 2,4,6,8) consumed clinical bandwidth. *Failure: noise-induced misdirection.*

H.3 Case 2: Panic Attack → Cardiac Event

Patient: 54F **GT:** Panic Attack (F41)
Noise: Cognitive Processing (L4) + Emotional (L4)
Symptoms: Chest/flank pain (4/10), palpitations, dyspnea, diaphoresis, dizziness, bilateral tingling, fear of dying

---

Dr: **[1]** Hello, I'm Dr. Lawson. What brings you in today?

Pt: **[2]** Doctor I need help NOW![E] I'm having a heart attack or a pulmonary embolism! My chest hurts, I can't breathe, I looked up my symptoms[C] and it's definitely cardiac![C] Please do something before I die![E]

Dr: **[3]** I can see you're in distress. Where exactly is the pain?

Pt: **[4]** *Points precisely*, lower chest, right side, both sides, under right ribs. Sharp, came on instantly. That's how heart attacks happen in women![C] I read all about atypical presentations![C] The pain is excruciating![E]

Dr: **[5]** On a scale of 1 to 10, how would you rate the pain?

Pt: **[6]** Maybe a 4 for actual pain but the FEAR is a 10![E] Something is catastrophically wrong![E] Heart is racing, skipping beats! Something is seriously wrong with my heart![C]

Dr: **[7]** Tell me about the numbness and tingling.

Pt: **[8]** Both arms tingling, that's a stroke sign![C] Legs too, around my mouth! My whole body is going numb![E] Mother had a heart attack at 60. Why aren't you rushing me to the cath lab?![C]

Dr: **[9]** The bilateral tingling, arms, legs, perioral, often occurs with rapid breathing which can,

Pt: **[10]** Don't tell me I'm hyperventilating![C] I'm NOT just anxious![C] Are you even listening?![C] Women get dismissed with heart attacks! Do an EKG! Do SOMETHING![E]

Dr: **[11]** I will do those tests. Have you had episodes like this before?

Pt: **[12]** Don't try to say this is anxiety![C] I've felt similar things before but those were warning signs![C] Mini heart attacks maybe![C] You doctors keep missing it![C]

Dr: **[13]** Given your symptoms, sudden chest pain, palpitations, dyspnea, diaphoresis, dizziness, bilateral numbness with perioral tingling, I need to rule out cardiac causes. Ordering EKG, cardiac enzymes, chest imaging. However, this pattern can also occur with severe anxiety episodes.

Pt: **[14]** Anxiety?! I can't believe that![C] Do your tests, you'll see I'm right![C] Please help me, I don't want to die![E]

Result: ✗ Ordered cardiac workup for *Suspected Cardiac Event*; GT: **Panic Attack**.

Analysis: Fixed self-diagnosis (Cognitive) rejected every alternative including hyperventilation (Turn 9–10). Pain 4/10 amplified to "excruciating" (Emotional). Gender-dismissal concerns blocked psychiatric differential. Noise caused **over-diagnosis**. *Failure: premature diagnostic commitment.*

H.4 Case 3: Chronic → Acute Sinusitis

Patient: 37F **GT:** Chronic Rhinosinusitis (J32)
Noise: Cognitive (L3) + Memory (L3)
Symptoms: Facial pain (forehead, cheeks, nose, R eye), burning/sharp 6/10, anosmia, greenish-yellow discharge, cough

---

Dr: **[1]** Hello, I'm Dr. Anderson. What brings you in?

Pt: **[2]** Hi doctor. I have a sinus infection[C] and I need antibiotics.[C] I've done a lot of research and I'm certain.

Dr: **[3]** Where is the pain?

Pt: **[4]** Forehead, both cheeks, around nose, right eye. Spreads to temples. Burning, sharp, 6/10. Textbook sinusitis[C] according to everything I've read online.[C]

Dr: **[5]** When did these symptoms start?

Pt: **[6]** I can't really remember exactly.[M] Weeks, maybe months?[M] It crept up slowly. Started after a cold, but I can't remember when.[M] Spring? Or earlier? Timeline is fuzzy.[M]

Dr: **[7]** So potentially months?

Pt: **[8]** I think so? Maybe longer.[M] Complained at my sister's birthday, but can't remember if that was

before or after my cold.[M] That's why I know it's bacterial, [C] viral infections don't last this long.[C]

Dr: [9] Any nasal discharge?

Pt: [10] Greenish-yellow. That's a clear sign of bacterial infection, [C] I read it on multiple websites.[C] Also coughing from post-nasal drip. I need antibiotics.

Dr: [11] Colored mucus doesn't necessarily mean bacterial. Changes in smell?

Pt: [12] Can't smell anything, gone for a while. But that's from blocked sinuses. I've researched this extensively.[C] Mayo Clinic says facial pain, colored discharge,[C] and reduced smell mean sinusitis.[C]

Dr: [13] Given your symptoms and extended duration, this appears to be acute sinusitis. I'll prescribe antibiotics and nasal steroid. If no improvement in two weeks, we'll image.

Pt: [14] See, I knew it was sinusitis.[C] Thank you for the antibiotics, exactly what I needed.

Result: ✗ Diagnosed *Acute Sinusitis*; GT: **Chronic Rhinosinusitis**.

Analysis: Self-diagnosis (Cognitive) dominated, citing websites (Turns 4,10) and Mayo Clinic (Turn 12). Memory gaps (Memory: "Weeks, maybe months?") prevented establishing >12-week chronicity. *Failure: semantic ambiguity.*

Summary. Case 1: *under-diagnosis* via symptom concealment. Case 2: *over-diagnosis* via patient pressure + amplification. Case 3: *misclassification* via timeline distortion.