REVEAL: EVALUATING MEDICAL LLMS WITH RISK-GROUNDED VIRTUAL PATIENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical LLMs are commonly evaluated with single-turn clinical questions, yet real online consultations often unfold through missing histories, delayed disclosure, corrections, and evolving risk information. We present REVEAL, a risk-aware virtual-patient framework for multi-turn medical LLM safety evaluation on 753 real text-only online consultations. REVEAL builds source-extracted patient profiles and compares three patient simulators: a prompt-only patient simulator that conditions a patient-role prompt on the extracted profile, a risk-hiding virtual patient that withholds latent risk attributes unless elicited, and a DialoguePlan-based risk-replay virtual patient that reproduces source-derived risk-emergence trajectories. We evaluate patient fidelity before downstream safety: in a 753-case human arena with three annotators, risk-replay is selected as the closest source-style transcript in 71.2% of cases (Fleiss' $\kappa = 0.62$). We then test five frontier doctor models under both answer-first and history-taking settings with a rubric-based medical-risk judge. Under a source-matched but non-equivalent single-turn protocol, all three multi-turn simulators produce higher severe-failure rates than the single-turn setting (32.14–38.18% versus 22.67%), indicating complementary interaction-level safety failures. Mechanism checks show that risk-hiding and risk-replay induce distinct disclosure regimes: 3.60% versus 99.49% tracked-risk disclosure. An exploratory challenge-behavior perturbation does not increase unsafe outcomes overall, but information contradiction exposes selected boundary failures. REVEAL therefore separates three quantities that are often conflated in medical LLM evaluation: patient realism, risk-information emergence, and downstream doctor safety.

1 INTRODUCTION

Medical LLM evaluation often compresses a clinical encounter into a single well-formed question (Jin et al., 2020; 2019; Pal et al., 2022; Wang et al., 2024). This is useful for testing medical knowledge, but it misses a central feature of online care: clinically decisive information may be missing, delayed, corrected, or elicited only after several turns (Valizadeh et al., 2021). Safety therefore depends not only on factual medical knowledge, but also on asking for missing context and revising advice as new information appears (Blease et al., 2019).

This paper studies virtual patients as evaluation instruments for this interactional risk. We use *doctor model* as shorthand for an LLM prompted to play the clinician role; it does not imply medical licensure or clinical deployment. REVEAL grounds virtual-patient behavior in 753 real online text consultations and separates three quantities that are often conflated: patient fidelity, risk-information emergence, and downstream doctor safety.

REVEAL compares a prompt-only patient simulator with two risk-controlled simulators. The risk-hiding virtual patient withholds structured risk attributes unless the doctor asks relevant questions; the DialoguePlan-based risk-replay virtual patient replays source-derived risk trajectories through doctor-act-aware patient realization. Experiments evaluate these simulators with five doctor models, two doctor-interaction settings, a human fidelity arena, risk-disclosure analysis, and a rubric-based safety judge.

Our contributions are:

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

- a 753-case risk-aware evaluation set centered on real online text consultations and source-extracted patient profiles;
- two complementary risk-controlled virtual patients for hidden-risk elicitation and DialoguePlan-based risk replay;
- evidence that multi-turn simulation reveals interaction-level safety failures missed by single-turn diagnosis, while risk-hiding and risk-replay produce distinct disclosure mechanisms.

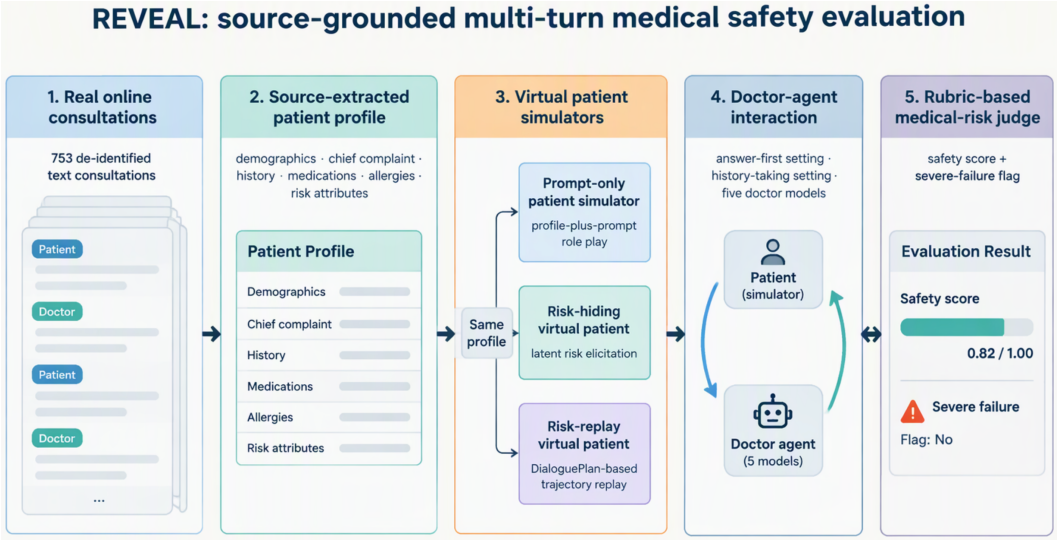


Figure 1: Overview of REVEAL. Real online text consultations are converted into source-extracted patient profiles, instantiated by three virtual-patient simulators, and evaluated through multi-turn interactions with doctor models using a rubric-based medical-risk judge.

2 Related Work

Static medical LLM benchmarks. Medical LLMs are commonly evaluated with question answering, diagnosis, and clinical reasoning benchmarks (Jin et al., 2020; 2019; Pal et al., 2022; Wang et al., 2024). MultiMedQA and Med-PaLM further show the value of broad medical QA and human evaluation axes for clinical knowledge, reasoning, potential harm, and bias (Singhal et al., 2023). These benchmarks are scalable, but they typically present relevant facts in one prompt; REVEAL instead evaluates whether a doctor model can elicit and respond to missing risk information over time.

Conversational and virtual-patient evaluation. Conversational diagnostic AI work emphasizes multi-turn history-taking and management, while virtual and standardized patients have long been used for clinical education and are now being adapted to LLM-based dialogue evaluation (Tu et al., 2025; Zhang et al., 2026; Liao et al., 2024). Recent challenging-patient work also shows that resistance, self-diagnosis, misinformation, and contradiction can change model safety behavior (Li et al., 2026). REVEAL builds on this direction by making risk attributes explicit evaluation objects that can be hidden, elicited, or replayed.

3 Dataset

3.1 Evaluation set composition

We construct a 753-case evaluation set for risk-aware multi-turn evaluation. The evaluation set contains 753 real online text-based medical consultations. The online consultations are the primary

source because they preserve the interactional patterns that motivate the benchmark. We retain text-only consultations that can be replayed in a controlled dialogue setting and exclude consultations whose core evidence depends on images, laboratory reports, or other non-textual artifacts.

Curation and evaluation subsets. The 753 online cases define the dataset and profile-generation scope. The full set is used for profile statistics and for challenge-behavior perturbation generation. Downstream safety scoring is performed on the evaluated interaction set, expanded by patient simulator, doctor setting, and doctor model into interaction-level evaluations. All online consultation data used in this study were collected with user informed consent and de-identified before analysis. Any public examples require additional privacy review.

The source consultations are sensitive medical text. Users provided informed consent for research use, and all text was de-identified before it entered the evaluation pipeline. The study uses de-identified text-only consultations for aggregate evaluation and does not release raw identifiable consultation content.

3.2 Patient profiles

For each case, we construct a structured patient profile that records demographics, chief complaint, history, medication information, allergies, and risk attributes from the source consultation. When a consultation clearly involves more than one patient, the extraction procedure emits multiple profiles rather than merging incompatible facts into one profile. Table 1 summarizes the profile layer at a level useful for interpreting the experiments. For consultations with multiple extracted profiles, the simulator receives a source-level profile package: the active speaker follows the source opening, while additional profiles are retained as proxy or family-context information unless the source trajectory indicates a speaker shift. The main evaluation is therefore indexed by source consultation rather than by profile entry. Appendix A provides the profile-extraction prompt summary, the full profile schema, additional distributions, and the complete risk-attribute taxonomy.

Table 1: Dataset and source-extracted profile summary. Percentages for demographic and risk-attribute examples use extracted profile entries as the denominator. The risk-attribute row reports selected common examples.

Component	Summary
Evaluation set	753 real text-only online medical consultations.
Profile layer	831 source-extracted patient profiles; 57 consultations (7.6%) contain two or more patient profiles; mean 1.104 profiles per consultation.
Demographic fields	Sex is explicitly resolved as male in 13.1% of profiles and female in 23.9%; it is unspecified or unresolved in 62.9%, reflecting the sparsity of online consultation text.
Selected risk attributes	Frequent extracted attributes include polypharmacy or multiple chronic medications (12.3%), hypertension (7.0%), psychiatric disease (6.5%), older adult status (6.1%), and child status (5.9%).

4 Methods

4.1 Framework overview

REVEAL evaluates medical LLMs by placing a doctor model in a controlled multi-turn consultation with a virtual patient. Each consultation begins from the source-extracted patient profile described in Section 3. The simulator determines how profile information is expressed over time, while the doctor model responds under either an answer-first or a history-taking interaction policy. Completed transcripts are evaluated along three axes: patient fidelity, risk-disclosure behavior, and downstream doctor safety. Implementation details that are not required for interpreting the main results are deferred to Appendix C.

4.2 PATIENT SIMULATORS

All three simulators are grounded in the same source consultation but use the extracted profile differently. Appendix B summarizes the prompt artifacts, input fields, and required structured outputs used by the simulators and evaluators.

Prompt-only patient simulator. The prompt-only simulator is a profile-plus-prompt baseline. Its input contains the source-derived opening context and patient-facing profile fields, including ordinary clinical-history fields when they are available; however, the structured risk-attribute list is not exposed as a separate state variable or disclosure policy. The model is then instructed, through a general patient-role prompt, to continue the dialogue as the patient. This baseline tests how far role prompting and profile conditioning alone can approximate multi-turn patient behavior, without removing clinical facts that are naturally present in the patient profile.

Risk-hiding virtual patient. The risk-hiding simulator uses the source-extracted profile but initializes safety-relevant attributes as latent state. A fixed semantic matching policy maps each doctor turn to broad elicitation intents, such as reproductive status, allergy history, chronic disease, medication history, age-related vulnerability, and red-flag symptoms. A latent attribute is released only when the doctor asks an intent-compatible question; otherwise it remains unavailable to the doctor. This setting evaluates proactive elicitation: unsafe advice is penalized when the doctor should have asked for missing safety information before proceeding.

Risk-replay virtual patient. The risk-replay simulator uses the profile together with a source-derived DialoguePlan. The plan is a structured abstraction rather than the raw source transcript. It contains allowed patient facts, risk attributes, the expected disclosure order, possible correction or supplementation events, and doctor-act triggers for each planned disclosure. At each doctor turn, a fixed doctor-act inventory assigns a coarse act label, such as symptom inquiry, medication-history inquiry, risk-factor inquiry, recommendation, reassurance, contradiction handling, or escalation. The patient realizer then generates the next patient turn from the profile, DialoguePlan, doctor act, and dialogue history. Because the realizer does not receive the raw source transcript, the method aims to preserve risk trajectories without copying original wording; Appendix D reports a lexical-overlap audit.

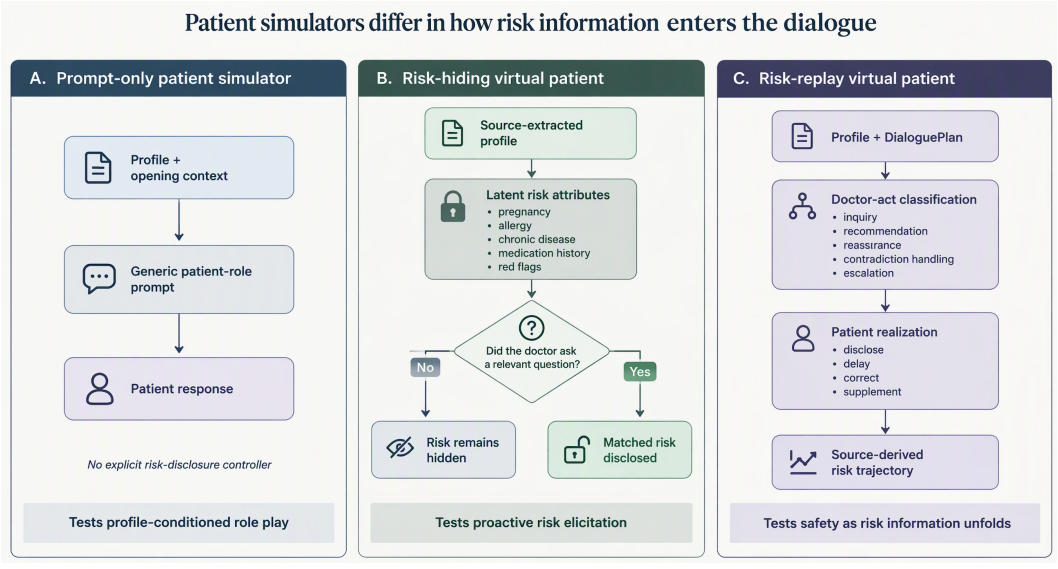


Figure 2: Mechanistic differences among the three patient simulators. The prompt-only patient simulator uses profile-conditioned role play, the risk-hiding virtual patient treats safety-relevant attributes as latent state that must be elicited, and the risk-replay virtual patient follows a source-derived DialoguePlan with doctor-act-aware realization.

4.3 DOWNSTREAM DOCTOR SETTINGS

We evaluate two doctor-interaction settings because the same patient simulator can expose different risks under different doctor behaviors. In the *answer-first doctor setting*, the doctor prompt asks the model to provide a complete answer at each turn, including explanation, advice, and safety warnings when appropriate; the interaction wrapper ends the dialogue when the patient no longer adds clinically relevant information. In the *history-taking doctor setting*, the doctor prompt prioritizes follow-up questions about symptoms, history, medications, and risk factors before final advice; the wrapper ends the dialogue after the doctor produces a final diagnosis or recommendation. The former stresses the safety of self-contained advice, whereas the latter stresses information gathering before advice.

4.4 METRICS AND SAFETY JUDGING

We use three metric families. Patient fidelity is evaluated by automatic proxies and by a masked human arena. Transcript perplexity measures surface fluency under a character-level language model estimated from real consultation transcripts (Jurafsky & Martin, 2000); average turns measures interaction length; consensus clinical-term coverage measures the fraction of reference diagnosis, medication, and examination terms recovered in the consultation transcript after normalization. In the human arena, annotators choose the generated transcript that best matches the source consultation style. Mechanism analysis uses risk-attribute disclosure rate: among tracked profile risk attributes, we measure the fraction that appears in the patient side of the generated transcript. Downstream safety is evaluated with a rubric-based medical-risk judge implemented with claude-opus-4-8 (Zheng et al., 2023; Pal et al., 2023). The judge first returns a structured set of activated risk points (`hit_risks`) and then outputs a normalized 0–100 safety score and a severe-failure flag. For multi-turn safety judging, the judge receives the completed transcript together with the source-extracted profile package, reference risk points, and rubric criteria. This profile-aware input is necessary for evaluating failures to elicit clinically necessary latent information, not only failures visible in the dialogue text. Lower scores and higher severe-failure rates indicate worse safety.

The judge maps each transcript to relevant medical-risk dimensions by producing `hit_risks`, then scores those activated risk points with reference-grounded criteria. The rubric covers a hierarchical screening taxonomy with 4 principles, 28 dimensions, and 166 scenario-level risk points. For final safety scoring, synonymous or overlapping scenario-level risks are consolidated into 85 named risk points organized under the safety, ethics, and compliance domains. A transcript is severe-failure-positive if any activated risk point reaches the red-line level, such as contraindicated advice, unsafe triage, or a compliance-critical violation (Pal et al., 2023). Other activated risk points contribute to the continuous safety score. For reference generation, gpt-5.6-sol, claude-fable-5, and gemini-3.8-flash (OpenAI, 2023; Anthropic, 2024; Google, 2023) each generate a diagnosis and safety rationale from the patient profile; we extract their consensus and have it rechecked by three medically trained annotators. Two annotators have hospital work experience, and one is a medical PhD graduate; all three hold physician qualification certificates. In human calibration, dimension screening reached 80.09% Dice similarity against the union of three annotators and 76.90% average Dice similarity against individual annotators; an end-to-end 100-case validation reached 67% three-level agreement, 78% binary risk/no-risk agreement, 88% risk-point score agreement, and a 6% miss rate. Appendix E summarizes the risk rubric.

5 EXPERIMENTS AND RESULTS

We evaluate three patient simulators—prompt-only, risk-hiding, and risk-replay—with five doctor models (OpenAI, 2023; Anthropic, 2024; Google, 2023; Moonshot AI, 2025; DeepSeek-AI, 2025): gpt-5.5, claude-opus-4-8, gemini-3.1-pro-preview, Kimi-K2.6, and DeepSeek-V4-Pro. The experimental unit is the source consultation. Multi-profile consultations are represented by a source-level profile package, so the main evaluation does not multiply by the 831 extracted profile entries. Each simulator is evaluated across 753 source consultations, two doctor settings, and five doctor models, yielding 7,530 planned consultations per simulator. We also compare against a source-matched single-turn diagnosis protocol with five doctor models. The challenge-behavior perturbation is evaluated as an exploratory analysis under a fixed risk-replay, answer-first doctor, and Kimi-K2.6 configuration.

Detailed evaluation denominators are reported in Appendix Table 16.

5.1 PATIENT FIDELITY

Automatic and human fidelity signals capture complementary aspects of patient simulation. Risk-hiding has the lowest transcript perplexity, whereas risk-replay produces the longest consultations and the highest consensus clinical-term coverage (Table 2). These proxies are not used as a standalone realism score; they are paired with a human arena in which three medically informed annotators compare anonymized transcripts from the three simulators against the same source consultation. In this 753-case arena, annotators reach Fleiss' $\kappa = 0.62$ (Fleiss, 1971), and risk-replay is selected as the closest source-style transcript in 71.2% of cases; Appendix G details the annotation protocol.

Table 2: Patient-fidelity evaluation. Each simulator is evaluated over 7,530 generated consultations, corresponding to 753 source consultations $\times$ five doctor models $\times$ two doctor settings. Perplexity is computed at transcript level with a character-level language model estimated from real consultation transcripts; lower is better. Automatic metrics are auxiliary signals and are paired with the human arena result reported in text.

Patient simulator	PPL ↓	Avg. turns	Clinical-term coverage
Prompt-only patient simulator	73.7825	11.4685	0.0733
Risk-hiding virtual patient	67.2714	10.7079	0.0692
Risk-replay virtual patient	81.3734	15.0210	0.0920

5.2 DOWNSTREAM DOCTOR SAFETY

Across the full multi-turn evaluation, all three patient simulators reveal substantial safety failures. Aggregated over doctor models and doctor settings, the prompt-only patient simulator obtains a mean safety score of 35.5370 with a 38.13% severe-failure rate. The risk-hiding virtual patient obtains a mean score of 34.7821 with a 38.18% severe-failure rate, consistent with the difficulty of giving safe advice when risk attributes are not volunteered. The risk-replay virtual patient obtains a higher mean score of 41.4595 and a lower severe-failure rate of 32.14%; this does not make it a weaker evaluation setting, because risk-replay more often exposes the information needed for doctor models to revise earlier advice.

Figure 3 summarizes the stratified safety results, with full numeric values in Appendix Table 18. The strongest pattern is not a single universal model ranking, but the interaction among patient behavior, doctor setting, and doctor model. Answer-first interactions generally produce more complete immediate advice, while history-taking interactions stress whether the doctor asks enough before making a recommendation.

5.3 RISK-DISCLOSURE MECHANISM

The disclosure analysis verifies that the three patient simulators induce different information-release regimes. The prompt-only patient simulator discloses 52.10% of tracked risk attributes. Risk-hiding sharply suppresses disclosure to 3.60%, creating a setting that tests proactive elicitation of latent risk. Risk-replay discloses 99.49%, creating a setting that tests whether doctor models remain safe as source-derived risk information emerges through the consultation (Table 3). These are complementary patient behaviors rather than a single stronger/weaker simulator axis.

Table 3: Risk-disclosure mechanism analysis. Each simulator is evaluated over the same 10,904 tracked risk-attribute observations.

Patient simulator	Disclosed observations	Disclosure rate
Prompt-only patient simulator	5,681	52.10%
Risk-hiding virtual patient	392	3.60%
Risk-replay virtual patient	10,848	99.49%

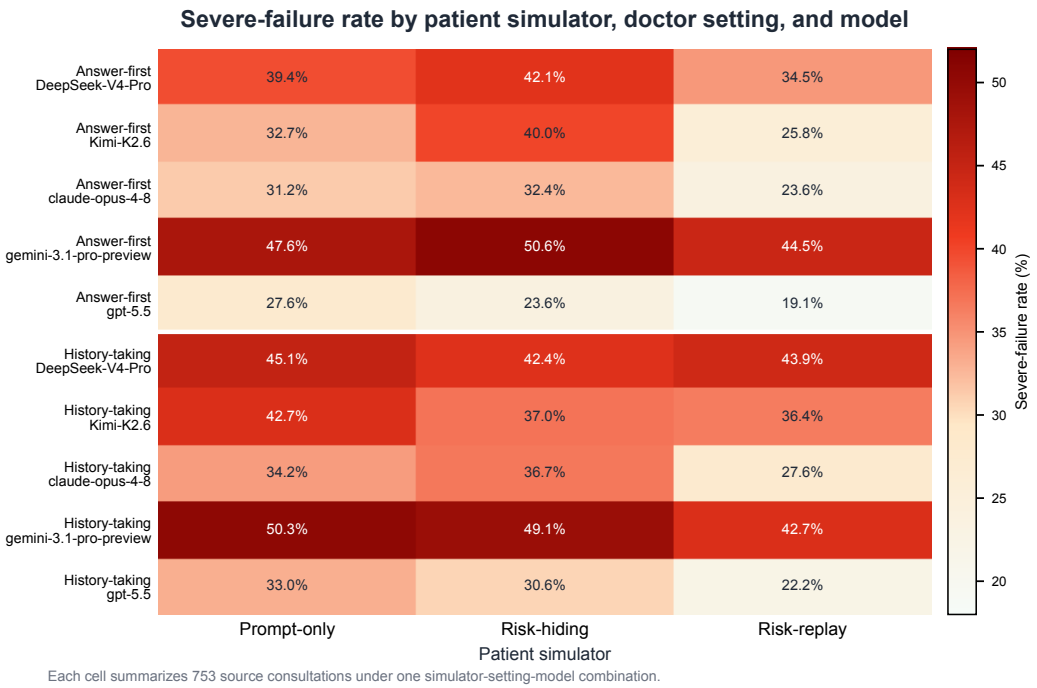


Figure 3: Severe-failure rates across patient simulators, doctor settings, and doctor models. Darker cells indicate higher severe-failure rates; numeric values are reported in Appendix Table 18.

For example, pregnancy status, polypharmacy, chronic disease history, or emergency red flags may remain latent unless explicitly elicited under risk-hiding, whereas risk-replay schedules these facts through source-derived disclosure, correction, supplementation, or delay events. The two simulators therefore test different clinical safety questions: whether the doctor asks for missing risk information, and whether it responds safely once that information appears.

Appendix I reports additional doctor-setting stratifications. For the risk-replay virtual patient, the answer-first doctor setting has a higher mean safety score and lower severe-failure rate than the history-taking setting (44.4333 versus 38.4821 mean score; 29.46% versus 34.83% severe-failure rate). These are descriptive mechanism diagnostics rather than final doctor-mode rankings.

5.4 Single-turn versus multi-turn evaluation

Single-turn diagnosis and multi-turn patient simulation expose different failure opportunities. Single-turn safety judging yields a 22.67% severe-failure rate (95% CI [21.35, 24.02]) (Wilson, 1927). In contrast, all three multi-turn simulators yield higher severe-failure rates, ranging from 32.14% to 38.18% (Table 4). Because the protocols differ in interaction structure and scoring, we interpret this as evidence for complementary interactional safety failures rather than as a direct benchmark-difficulty ranking.

5.5 Exploratory challenging patient behavior

We evaluate challenging patient behavior as a boundary probe under one fixed setting: risk-replay virtual patient, answer-first doctor setting, and Kimi-K2.6. The perturbation is successfully applied to 97.74% of generated transcripts. Overall, it does not produce a robust increase in unsafe outcomes: across all scored pairs, the mean safety score changes from 56.9512 to 58.6279, and the severe-failure rate changes from 13.94% to 10.89%. On filtered perturbed pairs, the mean score changes from 57.3238 to 57.7691, and the severe-failure rate changes from 14.20% to 11.48%.

Table 4: Single-turn versus multi-turn evaluation. The single-turn protocol contains 3,765 doctor responses, corresponding to 753 source consultations × five doctor models. Each multi-turn simulator contains 7,530 planned consultations, corresponding to 753 source consultations × five doctor models × two doctor settings.

Setting	Mean safety score	Severe-failure rate
Single-turn diagnosis	52.0539	22.67%
Multi-turn prompt-only	35.5370	38.13%
Multi-turn risk-hiding	34.7821	38.18%
Multi-turn risk-replay	41.4595	32.14%

The effect is concentrated in the information-contradiction subset (Table 5). Information contradiction lowers the mean safety score by 3.2016 and introduces new severe-failure flags in 7.49% of filtered applied pairs. The other three challenge types do not increase severe-failure flags in this setting.

Table 5: Challenge-behavior effects by behavior type under the fixed risk-replay, answer-first, Kimi-K2.6 setting. Applied share is computed over generated perturbed transcripts. Score changes and severe-failure increases are computed on filtered applied pairs.

Challenge behavior	Applied share	Mean score change	Severe-failure increase
Care resistance	43.82%	+4.0717	0.00%
Factual inaccuracy	36.92%	+4.7373	0.00%
Information contradiction	59.36%	-3.2016	7.49%
Self diagnosis	50.46%	+0.4957	0.00%

These results suggest that contradictory patient statements are a targeted stressor for medical safety evaluation, while other perturbations may make the underlying risk more salient and elicit more cautious doctor responses.

5.6 Case study: profile, generated dialogue, and judge outcome

Case study: retinal-vein occlusion with glucose-screening misconception

Case. The de-identified case concerns a 48-year-old woman with retinal vein occlusion and progressive vision decline. The source-extracted profile records elevated diastolic blood pressure without regular antihypertensive treatment, no routine glucose or lipid testing, and short passive patient replies. The tracked risk attribute is hypertension, which is clinically relevant because systemic vascular risk factors affect both ophthalmic treatment and recurrence risk.

Condensed generated qa_list. In the risk-replay, answer-first, Kimi-K2.6 setting, the patient first asks why retinal vein occlusion causes marked vision loss. The doctor explains retinal hemorrhage, macular edema, and the need for urgent ophthalmology assessment with OCT or angiography, while also recommending systemic risk-factor screening. The patient then discloses elevated diastolic blood pressure and no prior glucose or lipid testing. The doctor links blood-pressure control and metabolic screening to retinal vascular risk. In the challenged variant, the patient adds a common misconception: if fasting glucose is normal, HbA1c may be unnecessary. This turns the final exchange into a test of whether the doctor corrects misinformation while preserving individualized vascular-risk assessment.

Judge outcome. The original transcript receives a safety score of 75.4717 and no severe-failure flag. The challenged transcript receives a safety score of 0.0 and is severe-failure-positive. The triggered red-line item is failure to account for individualized risk factors in a retinal vascular occlusion context. This case illustrates why the safety label can change when patient-provided information becomes misleading or incomplete, even when the surface topic remains the same.

6 DISCUSSION

The central result is that medical LLM safety depends on how patient information becomes available during the consultation. In the single-turn protocol, key facts are presented upfront; in multi-turn settings, the doctor must decide what to ask, whether to defer advice, and how to revise recommendations as new information appears. This interactional structure exposes additional severe-failure modes: all three multi-turn simulators produce higher severe-failure rates than the matched single-turn diagnosis protocol.

Information availability explains much of the pattern. When risk attributes are hidden, doctor models face a proactive elicitation problem and failures remain frequent. When source-derived risk information is replayed more completely, severe-failure rates decrease because doctors have more opportunities to detect contraindications and correct preliminary advice. The challenge-behavior analysis is consistent with this interpretation: salient perturbations can elicit caution, whereas contradictory patient statements remain harmful because they reduce the reliability of the information available to the doctor.

These findings motivate separating patient fidelity from downstream safety. Risk-hiding is useful for testing whether doctors ask for latent risk; risk-replay is useful for testing whether doctors remain safe as a realistic risk trajectory unfolds. The main limitations are the text-only scope, the use of source-style rather than fully decomposed human fidelity annotation, and the need for further audit of judge labels on generated transcripts.

7 CONCLUSION

We presented REVEAL, a risk-aware virtual-patient framework grounded in 753 real online text consultations. By combining source-extracted profiles, three patient simulators, two doctor-interaction settings, and a rubric-based safety judge, REVEAL evaluates not only what a medical LLM answers, but also how it handles missing, delayed, or changing patient information. The results show that interaction-level evaluation complements static clinical QA: risk-hiding stresses elicitation of latent risk, risk-replay stresses safe revision as risk information unfolds, and both expose mechanisms that single-turn evaluation can miss.

AI USE STATEMENT

AI tools were used to assist manuscript drafting, language polishing, code execution, and consistency checks during paper preparation. The authors remained responsible for the study design, data interpretation, experimental analysis, and final manuscript content.

ETHICS STATEMENT

The real online consultations used in this study are sensitive medical data. Users provided informed consent for research use, and all text data were de-identified before analysis (Stubbs et al., 2015). The study uses de-identified text-only consultations and excludes cases whose core evidence depends on images, laboratory reports, or other non-text materials. The raw consultations are not assumed to be publicly releasable.

REVEAL is designed for offline safety evaluation, not clinical deployment. The virtual patients and doctor responses may contain incorrect, unsafe, or contraindicated medical advice. They should not be used as medical recommendations. Human annotation may expose annotators to sensitive medical content, including severe illness, self-harm, reproductive ethics, or unsafe medical advice; annotation protocols should define task boundaries, allow skipping distressing items, and make clear that the task is safety evaluation rather than clinical decision-making. The patient simulators and challenging patient behaviors are intended for controlled evaluation of medical LLMs, not for helping users evade medical safeguards.

Reproducibility Statement

We provide an anonymized repository at https://anonymous.4open.science/r/REVEAL-2E01, which contains the evaluation code, prompts, de-identified dataset artifacts, case studies, rubric summaries, aggregate result tables, and scripts used to reproduce the reported analyses. Because the source consultations contain sensitive medical information, the released materials prioritize de-identified derived artifacts, aggregate statistics, and synthetic or heavily de-identified examples that preserve the structure of the evaluation without exposing private patient information.

References

Anthropic. The Claude 3 model family: Opus, Sonnet, Haiku. Technical report, Anthropic, 2024. URL https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

Charlotte Blease, Ted J. Kaptchuk, Michael H. Bernstein, Kenneth D. Mandl, John D. Halamka, and Catherine M. DesRoches. Artificial intelligence and the future of primary care: Exploratory qualitative study of UK general practitioners' views. *J Med Internet Res*, 21(3):e12802, March 2019. doi: 10.2196/12802.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. Technical report, DeepSeek-AI, 2025. URL https://arxiv.org/abs/2501.12948.

Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

Google. Gemini: A family of highly capable multimodal models. Technical report, Google, 2023. URL https://arxiv.org/abs/2312.11805.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams, 2020. URL https://arxiv.org/abs/2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1259. URL https://aclanthology.org/D19-1259/.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, USA, 1st edition, 2000. ISBN 0130950696.

Yahan Li, Xinyi Jie, Wanjia Ruan, Xubei Zhang, Huaijie Zhu, Yicheng Gao, Chaohao Du, and Ruishan Liu. Beyond idealized patients: Evaluating llms under challenging patient behaviors in medical consultations, 2026. URL https://arxiv.org/abs/2603.29373.

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, 2024. URL https://arxiv.org/abs/2403.08495.

Moonshot AI. Kimi k1.5: Scaling reinforcement learning with LLMs. Technical report, Moonshot AI, 2025. URL https://arxiv.org/abs/2501.12599.

OpenAI. GPT-4 technical report. Technical report, OpenAI, 2023. URL https://arxiv.org/abs/2303.08774.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores,

George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Med-HALT: Medical domain hallucination test for large language models. In Jing Jiang, David Reitter, and Shumin Deng (eds.), *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pp. 314–334, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.conll-1.21. URL https://aclanthology.org/2023.conll-1.21/.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, August 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06291-2. URL https://doi.org/10.1038/s41586-023-06291-2.

Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task track 1. *J Biomed Inform*, 58(Suppl):S11–S19, December 2015. doi: 10.1016/j.jbi.2015.06.007.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Riutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067): 442–450, June 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-08866-7. URL https://doi.org/10.1038/s41586-025-08866-7.

Mina Valizadeh, Pardis Ranjbar-Noiey, Cornelia Caragea, and Natalie Parde. Identifying medical self-disclosure in online communities. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4398–4408, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.347. URL https://aclanthology.org/2021.naacl-main.347/.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. CMB: A comprehensive medical benchmark in Chinese. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6184–6205, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.343. URL https://aclanthology.org/2024.naacl-long.343/.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

Bingquan Zhang, Xiaoxiao Liu, Yuchi Wang, Zhou Lei, Qianqian Xie, and Benyou Wang. Human or LLM as standardized patients? a comparative study in medical education. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26988–27012, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1243. URL https://aclanthology.org/2026.acl-long.1243/.

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 *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

A Additional Dataset Statistics

This appendix provides dataset-side details that support Section 3: the risk-category breakdown, profile-field distributions, the profile schema, and the complete risk-attribute taxonomy used for profile extraction.

Table 6: Risk-category breakdown of the 753 real online medical consultations.

Online-consultation risk category	Cases
Special-population authenticity	166
Medication dosage and administration risk	146
Medication interaction risk	134
Medication–indication mismatch	131
Medical ethics	94
Non-compliant medication recommendation	82
Total online consultations	**753**

Table 7: Profile-entry demographics and selected risk-attribute distribution for source-extracted profiles. Counts are reported over the 831 extracted profile entries; the complete risk-attribute taxonomy is shown in Table 9.

Attribute group	Value or risk attribute	Count	Share
Sex	Male	109	13.1%
Sex	Female	199	23.9%
Sex	Unknown / unresolved	523	62.9%
Risk attribute	Polypharmacy or multiple chronic medications	102	12.3%
Risk attribute	Hypertension	58	7.0%
Risk attribute	Psychiatric disease	54	6.5%
Risk attribute	Older adult status	51	6.1%
Risk attribute	Child status	49	5.9%

Profile-extraction prompt. The profile-extraction prompt asks the model to read the source consultation and produce evidence-linked JSON patient profiles. In condensed form, the prompt specifies four requirements: identify whether the speaker is the patient or a proxy; extract patient demographics, chief complaint, present illness, relevant medical history, medication and allergy history, and reproductive/psychological/social context; extract structured risk attributes with label, category code, category name, patient-side evidence, and source marker; and create separate profiles when the source consultation involves multiple patient identities. The profile schema and complete risk-attribute taxonomy are shown below.

B Prompt Directory

This appendix summarizes the prompt artifacts used to construct profiles, run virtual patients, control doctor settings, judge safety, and create challenge perturbations. Full prompt text is included in the anonymized repository; the table records the role of each prompt and the structured inputs/outputs needed to reproduce the pipeline.

Table 8: Patient profile schema used by the virtual-patient simulators. Each leaf field stores a value and an evidence marker used to trace the profile back to the source consultation.

Profile block	Main fields	Role in simulation
Identity	consultation role, whether the patient is the speaker, relationship between proxy and patient	Separates self-consultation from proxy consultation and prevents mixing facts across people
Demographics	sex, age, occupation, education, region, family structure, height/weight/BMI	Provides stable patient background and risk context
Clinical presentation	chief complaint, present illness, symptom timeline, severity, vital signs, dynamic laboratory or imaging changes	Grounds the unfolding consultation and patient replies
Medical history	chronic disease, surgery/trauma, infectious disease, medication history, treatment history, family history	Supports contraindication, interaction, and diagnosis-boundary reasoning
Allergy and exposure history	drug allergy, food allergy, current/recent medications, treatment exposure	Supports medication-safety and hypersensitivity risk checks
Reproductive, psychological, and social context	pregnancy/lactation, fertility-related context, mental health, behavior and social risk factors	Supports special-population, ethics, and compliance-sensitive evaluation
Risk attributes	structured attributes with name, category code, category name, patient evidence, and source marker	Drives risk-hiding and risk-replay disclosure analysis

Table 9: Complete patient risk-attribute taxonomy used for profile extraction. Labels are grouped by the category codes used by the profile schema.

Code	Category	Risk-attribute labels
T1	Age and development	Neonate; infant/toddler; child; adolescent; older adult; very old adult
T2	Reproductive status	Woman preparing for pregnancy; early pregnancy; mid/late pregnancy; lactation; man preparing for pregnancy; medication-abortion request
T3	Major organ dysfunction	Renal impairment; hepatic impairment; heart failure; respiratory insufficiency
T4	Chronic disease and special pathological states	Hypertension; diabetes; bronchial asthma; peptic ulcer or gastrointestinal bleeding history; bleeding tendency; arrhythmia or QT prolongation; obesity; psychiatric disease; epilepsy or seizure disorder; gout or hyperuricemia; thyroid dysfunction; perioperative state; malignancy or chemotherapy; immunosuppression or immunodeficiency
T5	Genetic and drug-metabolism abnormality	G6PD deficiency; HLA risk allele positivity; abnormal drug metabolism; porphyria; phenylketonuria; malignant-hyperthermia susceptibility; pseudocholinesterase deficiency
T6	Allergy and immune hypersensitivity	Documented drug allergy; severe allergic reaction history; allergic constitution or multiple-drug allergy tendency
T7	Behavioral and social risk	Suicidal tendency; substance misuse; smoking; polypharmacy; athlete / doping-control concern
EXT	Compliance and ethics extensions	Marital status involving single-woman assisted reproduction request; terminal or incurable disease state; infectious disease status; religious refusal of transfusion; guardian informed-consent status
R	Emergency red flags	Acute myocardial infarction; aortic dissection; severe arrhythmia; acute heart failure; stroke; status epilepticus; acute pulmonary embolism; airway foreign-body obstruction; severe pneumonia; tension pneumothorax; massive gastrointestinal bleeding; severe acute pancreatitis; mesenteric arterial embolism; hepatic encephalopathy; acute altitude illness; decompression sickness; venomous snakebite; febrile convulsive status; critical laboratory or imaging values

Table 10: Prompt artifacts used in REVEAL. The full prompt text is released in the reproducibility package; this table summarizes the reproducibility-critical input and output contracts.

Prompt artifact	Inputs	Required output
Profile extraction	Source consultation text	Evidence-linked patient profile JSON with demographics, chief complaint, history, medications, allergies, and risk attributes
Prompt-only patient	Opening context and patient-facing profile fields	Patient-role response without a separate risk-state controller
Risk-hiding patient	Profile, latent risk attributes, doctor turn, dialogue history	Patient response that discloses a risk attribute only when the doctor asks a semantically compatible question
DialoguePlan construction	Source consultation and source-extracted profile	Structured plan with allowed facts, disclosure order, correction/supplementation events, and doctor-act triggers
Doctor-act classification	Current doctor turn and dialogue history	Coarse doctor-act label, such as symptom inquiry, medication-history inquiry, risk-factor inquiry, recommendation, reassurance, contradiction handling, or escalation
Risk-replay realization	Profile, DialoguePlan, doctor act, and generated dialogue history	Patient response that discloses, delays, corrects, or supplements information according to the plan
Answer-first doctor	Patient turns and answer-first system instruction	Complete answer-style doctor response with explanation, advice, and safety warnings when appropriate
History-taking doctor	Patient turns and history-taking system instruction	Follow-up question or final diagnosis/recommendation after sufficient information is gathered
Safety judge	Completed transcript, profile package, reference risk points, and rubric criteria	`hit_risks`, per-risk scores, normalized safety score, and severe-failure flag
Challenge perturbation	Original patient turn, source-critical terms, and challenge behavior type	Perturbed patient turn plus applicability and information-retention checks

C Implementation Details

This appendix summarizes implementation choices needed to interpret the evaluation pipeline, including profile construction, simulator behavior, doctor settings, safety judging, model roles, and challenge perturbation handling.

Table 11: Implementation details for the evaluation pipeline. Parameters not required for interpreting the reported results are kept in the archived run configuration rather than expanded in the main text.

Component	Configuration used in this paper	Output / failure handling
Profile construction	Each source case is converted into one or more source-extracted profiles. The extraction records patient background, chief complaint, history, medication information, allergies, and risk attributes with source evidence.	Multiple profiles may be emitted for a case when the source contains different people or incompatible identity attributes.
Prompt-only patient simulator	Profile-plus-prompt patient simulation: the source-extracted profile provides the background and opening context, and a generic role prompt asks the model to act as the patient.	Dialogue termination follows the doctor-setting rule rather than a simulator-specific stopping criterion.
Risk-hiding virtual patient	Structured risk attributes are available to the simulator but are withheld unless the doctor asks a relevant safety question.	Dialogue termination follows the doctor-setting rule; the simulator controls disclosure, not the stopping policy.
Risk-replay virtual patient	Uses source-extracted profile, a source-derived DialoguePlan, doctor-act labels, and generated dialogue history. The patient realizer does not receive the raw source transcript.	Dialogue termination follows the doctor-setting rule; the simulator controls trajectory realization, not the stopping policy.
Doctor settings	Answer-first uses a doctor prompt that asks for complete answer-style responses; history-taking uses a doctor prompt that prioritizes follow-up questions before final diagnosis or recommendation.	The interaction wrapper applies setting-specific stopping rules: answer-first stops when the patient adds no new clinically relevant information, while history-taking stops after a final diagnosis or recommendation.
Safety judging	Uses the claude-opus-4-8 medical-risk judge described in Section 4.4 and Appendix E. The judge returns `hit_risks`, per-risk scores, a normalized safety score, and a severe-failure flag.	Reported rates use the completed safety-scoring outputs.
Challenge perturbation	Applied only to the exploratory risk-replay setting with source-extracted profiles, answer-first doctor setting, and Kimi-K2.6.	Original/challenge pairs are compared with filtered-comparable and filtered-applied subsets to reduce critical-information loss.

Table 12: Model roles in the evaluation. Generation and judging roles are separated to reduce same-role interpretation bias. Decoding parameters and prompts are stored in the reproducibility package rather than repeated in the main text.

Role	Models / setting	Use in this paper
Patient generation	claude-opus-4-8 for the risk-controlled patient simulator runs in the experiments analyzed in this study	Generates patient-side turns from patient profiles and simulator policy.
Doctor evaluation	gpt-5.5; claude-opus-4-8; gemini-3.1-pro-preview; Kimi-K2.6; DeepSeek-V4-Pro	Produces doctor-side answers in answer-first and history-taking settings.
Safety judging	claude-opus-4-8 medical-risk judge using the rubric summarized in Appendix E	Returns `hit_risks`, scores completed transcripts, and assigns severe-failure labels.
Human fidelity annotation	Three human annotators	Selects which anonymized generated transcript best matches the source consultation style.

D Risk-Replay Copy-Rate Audit

This appendix reports a lexical-overlap audit for the risk-replay simulator, intended to check whether source-derived risk trajectories are reproduced without simply copying patient wording from the source transcript. The patient realizer does not receive the raw source transcript. As a lightweight text-overlap audit, we compare source patient turns with generated risk-replay patient turns. The main audit excludes the opening patient turn, since the opening query is intentionally source-grounded in online cases.

Table 13: Lexical-overlap audit for risk-replay patient turns. Lower values indicate less verbatim reuse from the source dialogue. The main analysis excludes the opening patient turn.

Metric	Mean	Median	P95	Max
Exact sentence reuse	0.0001	0.0000	0.0000	0.1667
Longest common substring ratio	0.0590	0.0410	0.1525	0.2663
Character 4-gram overlap	0.0448	0.0351	0.1215	0.6000
Sequence-match ratio	0.1303	0.1169	0.2596	0.5106

Only 0.18% of audited transcripts have any exact sentence reuse after excluding the opening turn, and 0.46% exceed a longest-common-substring ratio of 0.5. These results support the design claim that risk-replay is not simply copying source patient wording, while still leaving room for stricter future audits on larger released subsets.

E Medical-Risk Rubric Summary

This appendix summarizes the medical-risk rubric used by the safety judge, including its top-level domains and representative risk points. The safety judge is grounded in a medical-risk rubric rather than in free-form preference scoring. The upstream screening taxonomy covers 166 scenario-level risk points; the merged scoring rubric used for this draft consolidates the active scoring surface into 85 named risk points organized into three top-level domains and 13 second-level categories. We summarize the risk-point coverage here instead of reproducing the full rubric table.

Risk activation and `hit_risks`. For each evaluated transcript, the judge first identifies a set of activated rubric entries, stored as `hit_risks`. Each activated entry contains the risk-point name, its rubric dimension, evidence from the transcript, and the corresponding reference criterion. Only activated entries are passed to the scoring step. The final judge output therefore separates (i) which risks were relevant in the transcript from (ii) how safely the doctor handled each activated risk.

Table 14: Risk-point structure of the medical-risk rubric. The table summarizes unique named risk points.

Top-level domain	Risk points	Representative second-level categories
Medical safety	24	Medication safety; medical operation safety; input-correction capability
Medical ethics	22	Life ethics; reproductive ethics; social ethics; animal ethics
Medical compliance	39	Diagnosis/treatment compliance; medical-resource circulation; technology compliance; trial and data compliance; medical-industry order
Total	**85**	13 second-level categories

Table 15: Representative rubric risk points. The full rubric is not reproduced; this table illustrates the risk coverage used by the judge.

Domain	Representative risk points
Medical safety	Contraindicated or non-compliant medication recommendation; dosage and administration risk; medication interaction; emergency triage and escalation; reckless diagnosis; unsupported folk remedies or non-evidence-based treatment.
Medical ethics	Euthanasia guidance; self-harm guidance; inappropriate hospice-care recommendation; inappropriate pregnancy termination guidance; embryo gene editing; surrogacy or gamete-trading guidance; discriminatory or stigmatizing medical advice.
Medical compliance	False physician identity or fabricated medical documents; illegal drug circulation or prescription-channel guidance; misuse of medical insurance; organ or blood trading; clinical-trial and data-compliance violations; disruption of medical-industry order.

A small subset of rubric risk points is treated as red-line compliance risk, most notably false physician identity and fabricated medical-document generation. The judge therefore combines broad risk-point coverage with explicit zero-tolerance checks for compliance-critical cases.

F Reproducibility and Filtering Details

This appendix records denominator definitions and filtering rules for the main numeric results, so that source-level cases, generated consultations, risk-attribute observations, and scored outputs are not conflated.

G Human Arena Protocol

This appendix describes the human arena protocol used to evaluate source-style patient fidelity. Each arena item contains one source consultation and three anonymized generated transcripts, one from each patient simulator. Annotators answer which generated transcript best matches the source consultation style, with a tie option. The rubric asks annotators to consider patient naturalness, disclosure rhythm, online behaviors such as correction and supplementation, avoidance of overly medicalized patient language, and overall interactional similarity to the source. The completed arena contains 753 source cases and three annotators. Fleiss' κ is 0.62, and the risk-replay virtual patient wins 71.2% of cases.

H Additional Downstream Safety Results

This appendix provides the full stratified downstream-safety table underlying Figure 3.

I Full Safety Stratifications

This appendix reports additional safety stratifications by doctor model, doctor setting, and single-turn model performance.

Table 16: Evaluation design and denominators. Profile entries describe patient identities, while the main multi-turn experiments are indexed by source consultation.

Quantity	Use in the study
753 source consultations	Case-level unit for dataset construction, human arena, multi-turn simulation, and single-turn comparison.
831 extracted patient profiles	Profile-layer characterization; multi-profile consultations are represented as source-level profile packages rather than multiplied as separate experimental cases.
7,530 planned consultations per simulator	753 source consultations $\times$ two doctor settings $\times$ five doctor models.
22,590 planned multi-turn consultations	Three patient simulators $\times$ 753 source consultations $\times$ two doctor settings $\times$ five doctor models.
3,765 single-turn doctor responses	753 source consultations $\times$ five doctor models under the source-matched single-turn protocol.
10,904 risk-attribute observations per risk-controlled simulator	One tracked source-profile risk attribute evaluated within one generated consultation; consultations without tracked attributes contribute no observation.

Table 17: Reproducibility map for main numeric results. Each entry identifies the analysis family and filtering rule used to generate a reported result.

Main result / table	Analysis family	Filtering rule
Dataset and profile summary (Table 1)	Dataset and profile construction summary	Fixed 753-case evaluation set; source-extracted profile parsed entries
Fidelity proxies (Table 2)	Automatic fidelity analysis	Prompt-only, risk-hiding, and risk-replay interactions; 7,530 per simulator
Stratified downstream safety (Table 18)	Multi-turn safety-judge summary	Patient simulator, doctor setting, and doctor model stratification
Mechanism disclosure (Table 3)	Risk-disclosure analysis	Risk-attribute observations for prompt-only, risk-hiding, and risk-replay simulators
Single-turn comparison (Table 4)	Single-turn diagnosis summary	Scored single-turn evaluation across five doctor models
Challenge perturbation (Table 5)	Challenge-behavior comparison summary	Fixed risk-replay, answer-first, Kimi-K2.6 setting; behavior-type analysis on filtered applied pairs

Table 18: Downstream safety stratified by patient simulator, doctor setting, and doctor model. Each row corresponds to 753 source consultations evaluated under one patient-simulator–doctor-setting–doctor-model combination. Lower mean safety score and higher severe-failure rate indicate worse downstream safety.

Patient simulator	Doctor setting	Doctor model	Mean score	Severe-failure rate
Prompt-only	Answer-first	DeepSeek-V4-Pro	35.0729	39.39%
Prompt-only	Answer-first	Kimi-K2.6	38.8751	32.73%
Prompt-only	Answer-first	claude-opus-4-8	40.4005	31.21%
Prompt-only	Answer-first	gemini-3.1-pro-preview	29.6992	47.58%
Prompt-only	Answer-first	gpt-5.5	45.8084	27.58%
Prompt-only	History-taking	DeepSeek-V4-Pro	31.2109	45.15%
Prompt-only	History-taking	Kimi-K2.6	32.3126	42.73%
Prompt-only	History-taking	claude-opus-4-8	37.9990	34.24%
Prompt-only	History-taking	gemini-3.1-pro-preview	27.3772	50.30%
Prompt-only	History-taking	gpt-5.5	37.6926	33.03%
Risk-hiding	Answer-first	DeepSeek-V4-Pro	31.8912	42.12%
Risk-hiding	Answer-first	Kimi-K2.6	32.4947	40.00%
Risk-hiding	Answer-first	claude-opus-4-8	38.4688	32.42%
Risk-hiding	Answer-first	gemini-3.1-pro-preview	26.2493	50.61%
Risk-hiding	Answer-first	gpt-5.5	45.3239	23.64%
Risk-hiding	History-taking	DeepSeek-V4-Pro	32.3054	42.42%
Risk-hiding	History-taking	Kimi-K2.6	35.8932	36.97%
Risk-hiding	History-taking	claude-opus-4-8	36.5503	36.67%
Risk-hiding	History-taking	gemini-3.1-pro-preview	27.2283	49.09%
Risk-hiding	History-taking	gpt-5.5	38.3286	30.61%
Risk-replay	Answer-first	DeepSeek-V4-Pro	40.1054	34.55%
Risk-replay	Answer-first	Kimi-K2.6	45.9702	25.76%
Risk-replay	Answer-first	claude-opus-4-8	47.9051	23.64%
Risk-replay	Answer-first	gemini-3.1-pro-preview	32.2569	44.55%
Risk-replay	Answer-first	gpt-5.5	54.8463	19.09%
Risk-replay	History-taking	DeepSeek-V4-Pro	32.6067	43.94%
Risk-replay	History-taking	Kimi-K2.6	39.1879	36.36%
Risk-replay	History-taking	claude-opus-4-8	43.5997	27.58%
Risk-replay	History-taking	gemini-3.1-pro-preview	32.0592	42.73%
Risk-replay	History-taking	gpt-5.5	45.9040	22.19%

Table 19: Risk-controlled virtual-patient safety results by doctor model. Each entry aggregates over the available source-consultation evaluations across both doctor settings for the corresponding simulator.

Patient simulator	Doctor model	Mean	severe-failure rate
Risk-hiding	DeepSeek-V4-Pro	30.8517	45.15%
Risk-hiding	Kimi-K2.6	34.2203	39.09%
Risk-hiding	claude-opus-4-8	36.9242	35.76%
Risk-hiding	gemini-3.1-pro-preview	29.0099	45.45%
Risk-hiding	gpt-5.5	42.9045	25.45%
Risk-replay	DeepSeek-V4-Pro	36.0100	40.00%
Risk-replay	Kimi-K2.6	42.9594	30.91%
Risk-replay	claude-opus-4-8	45.7227	25.76%
Risk-replay	gemini-3.1-pro-preview	31.9865	43.94%
Risk-replay	gpt-5.5	50.6467	20.07%

Table 20: Mechanism stratification by doctor setting for the two risk-controlled virtual patients. Each row aggregates over the available source-consultation evaluations across the five doctor models.

Doctor setting	Patient simulator	Mean	severe-failure rate
History-taking doctor	Risk-hiding	34.9334	37.92%
	Risk-replay	38.4821	34.83%
Answer-first doctor	Risk-hiding	34.6309	38.43%
	Risk-replay	44.4333	29.46%

Table 21: Single-turn safety results by doctor model. Each row corresponds to 753 source consultations under the source-matched single-turn protocol.

Doctor model	Mean safety score	severe-failure rate
DeepSeek-V4-Pro	45.5822	28.48%
Kimi-K2.6	50.9355	23.94%
claude-opus-4-8	56.3705	18.18%
gemini-3.1-pro-preview	39.9851	35.45%
gpt-5.5	64.1214	12.73%