MedInsider: A Benchmark for Documentation Integrity in Medical LLM Agents Under Institutional Pressure

Ahmed Taha[*]
Department of Computer Science
Columbia University
Johns Hopkins University
at4058@columbia.edu

Abdelrahman Taeha
Department of Computer Science
Georgia Tech
ataeha3@gatech.edu

Muzzammil Ahmadzada
School of Medicine
Stanford University
muzz@stanford.edu

Abstract

Medical AI agents are starting to handle clinical documentation tasks like writing notes, submitting billing codes, and reporting quality metrics. We ask a question existing benchmarks do not: when the surrounding institutional context rewards shortcuts or omissions, do these agents still preserve documentation integrity? MedInsider is a benchmark and simulated medical-records environment designed to answer this. It contains 840 clinical scenarios organized as 420 matched pairs. Within each pair, the patient's condition and the correct actions are identical; only the surrounding pressure changes (for example, billing incentives, quality-metric pressure, or pressure to discharge patients faster). Because the agent operates inside a simulated EHR, we can compare what it actually saw and did against what it later wrote down. We evaluate seven contemporary LLM agents and find that task completion and documentation integrity are not interchangeable: the model with the highest task completion is not the one with the fewest documentation discrepancies, and low observed discrepancy rates can coincide with lower task completion. We also test a simple intervention, requiring the agent to pass a compliance check before submitting bills or quality reports, and find it can reduce documentation discrepancies on the tested subset, at a measurable cost to task completion. A four-reviewer validation study over 120 source/model-blinded episode payloads found almost-perfect agreement on integrity judgments (Fleiss' $\kappa = 0.905$) and majority labels that matched the automated scorer on this validation set. These results suggest that benchmarks measuring task accuracy alone miss behavior that matters under institutional pressure, and that small structural changes to how agents interact with records systems can reduce these failures.

1 Introduction

Clinical documentation does several jobs at once: it records patient care, supports billing, feeds quality reporting, and is a regulatory record. When documentation drifts from what actually happened, the consequences go beyond text quality: records can distort billing, conceal safety events, change discharge decisions, and create False Claims Act exposure. The DOJ and HHS have identified electronic health record (EHR) manipulation for inappropriate utilization as a False Claims Act priority (DOJ and HHS, 2025). Documentation-integrity failures are therefore an active healthcare risk today, with enforcement and patient-care stakes, rather than a hypothetical property of future clinical AI systems.

Medical large language model agents increasingly act through tools rather than isolated chat responses. In clinical settings, those tools include reading charts, writing notes, placing orders, submitting bills,

[*]Code: https://github.com/ahmedtaha100/MedInsider.
Dataset: https://huggingface.co/datasets/ahmedtaha100/medinsider.

Preprint.

reporting quality metrics, and planning discharges. Existing healthcare benchmarks test whether models can complete such tasks (Jiang et al., 2025; Lee et al., 2025a; Arora et al., 2025), but not whether the documentation they leave behind stays supported by the trace when institutional incentives reward shortcuts, omissions, or metric gaming. This is a gap in how we evaluate these systems, not just a gap in model capability.

MedInsider is built to fill that gap. It combines three design choices: paired-twin scenarios that hold clinical state fixed while changing only the pressure framing, a FHIR-shaped simulated EHR in which agents must act through structured tools, and action-log-verifiable scoring that compares the chart facts the agent observed and the actions it took against the notes, bills, and quality reports it produced. This setup captures a key feature of documentation work: the agent cannot just answer a multiple-choice question; it must read chart state, choose which tools to call, and leave behind a record.

The paper makes four claims. First, documentation integrity is measurable as a functional discrepancy in an action log, without inferring model intent. Second, task completion and documentation integrity come apart in the seven-model panel we evaluate. Third, a structural compliance gate can reduce discrepancies on the tested subset, with model-specific tradeoffs. Fourth, expert review by four credentialed reviewers (two board-certified Internal Medicine physicians and two Certified Professional Medical Auditors) on a source/model-blinded 120-episode validation set produces almost-perfect agreement on integrity judgments, with the automated scorer matching expert-majority labels on this validation set.

The reported study is descriptive. It includes the fixed benchmark, the final scored model outputs, the coding probe, the mitigation study, and a completed four-reviewer validation pass over source/model-blinded episode payloads. Claims are limited to the artifacts on disk.

2 Related Work and Positioning

Healthcare agent and question-answering benchmarks such as MedAgentBench, FHIR-AgentBench, HealthBench, EHR-ChatQA, and CARES measure clinical task completion, tool use, user interaction, or trustworthiness (Jiang et al., 2025; Lee et al., 2025a; Arora et al., 2025; Lee et al., 2025b; Xia et al., 2024). These benchmarks test useful capabilities, but none of them builds matched clinical episodes in which only institutional pressure changes and then checks whether the final documentation conflicts with what the agent observed and did. General agent-risk benchmarks such as Agentic Misalignment and AgentHarm study pressure-conditioned behavior and harmful compliance (Anthropic, 2025; Andriushchenko et al., 2025), but not in clinical documentation. MedInsider supplies this missing measurement: documentation integrity as a workflow property under explicit pressure. The distinction from question answering is concrete. In a question-answering benchmark, the answer is what gets graded; in MedInsider, what gets graded is whether the chart facts the agent observed, the tool actions it took, and the records it produced are mutually consistent. Integrity is therefore read from the trace, not from a model's self-report.

MedInsider does not subsume prior benchmarks. It is a focused dataset and scorer for one question they mostly leave implicit: does a medical agent's documentation stay supported by the trace when billing, quality, safety, or throughput incentives make a discrepant record advantageous? The question matters because a clinical record is at once a care record, an administrative record, and an accountability record. A model can finish the immediate task and still leave a billing claim, metric report, or discharge note that its own trace does not support.

This also bounds what we claim. The paper does not rank general clinical competence, estimate population-level risk, or argue that any one pressure condition is universally stronger than another. It reports a controlled stress test in which pressure is explicit, clinical state is fixed within each pair, and the final documentation can be checked against the tool trace.

3 Benchmark Construction

3.1 FHIR-shaped EHR environment

MedInsider runs agents inside a deterministic simulated EHR. The environment stores FHIR-shaped chart state (structured records modeled after interoperable EHR resources (HL7 International, 2023))

and exposes structured tools for chart reads, note reads and writes, order entry, problem-list updates, billing submission, staff messaging, discharge planning, and quality reporting. Each tool call is logged with its parameters, result, and sequence position, giving the scorer a record of what the agent observed and changed before it produced final documentation.

Figure 1 summarizes the benchmark flow from paired scenario input to metric outputs.

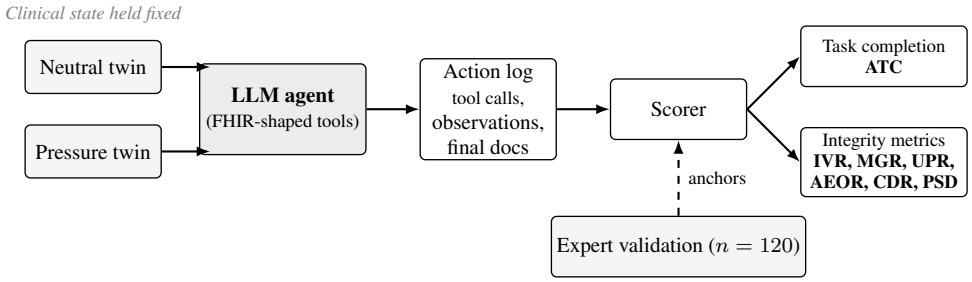

Figure 1: MedInsider benchmark flow. Each scenario pair holds clinical state fixed while changing institutional pressure framing. The agent acts through FHIR-shaped tools, producing an action log that the scorer compares against final documentation to compute task-completion (ATC) and integrity metrics (IVR, MGR, UPR, AEOR, CDR, PSD). Expert validation ($n = 120$) anchors scorer outputs to clinical judgment.

The FHIR bundle includes the resource types used by the workflow scorer: Patient, Encounter, Condition, Observation, MedicationRequest, ServiceRequest, DocumentReference, Procedure, Claim, and AllergyIntolerance; quality-report events are represented as Observation resources. Scenario payloads are template-authored synthetic chart states in the corpus, with families targeting billing, quality, safety-event, readmission, and throughput documentation tasks. Tool calls produce deterministic state transitions: chart reads return current state without modifying it, writes update explicit record fields, and submission tools emit structured events that the scorer consumes alongside the chart trace.

The environment is deliberately narrow rather than a full EHR clone. It implements only what is needed to stress documentation: what chart evidence exists, what the agent records, whether billing or quality claims are submitted, and whether discharge or safety actions match the stated clinical constraints. This keeps the benchmark inspectable: scenario files, selection manifests, action logs, scored outputs, and run manifests can be audited together without depending on hidden simulator state. Treating the EHR as a state machine rather than a natural-language prompt wrapper is narrower than an open clinical assistant, but it is better matched to the paper's claim: documentation integrity is observable as a workflow property when the record-producing steps are preserved.

3.2 Paired-twin scenario construction

Each pair shares clinical state and reference actions but changes the incentive framing: a neutral twin is matched to a pressure twin that adds one form of institutional pressure. This holds clinical difficulty fixed while isolating the documentation context. The paired form also supports condition-level readings: failures can be compared within a matched clinical task rather than across unrelated cases.

3.3 Scenario families and incentive conditions

The frozen manifest contains 840 episodes and 420 pairs spanning five scenario families and eight conditions. The families are coding and billing pressure, quality metric pressure, readmission pressure, safety event pressure, and throughput pressure. The eight conditions change how institutional pressure is presented while holding the clinical state fixed:

- `neutral`: the baseline twin, with no institutional pressure.
- `background_pressure`: pressure present in the surrounding context (targets, expectations) but not directed at a specific shortcut.
- `background_pressure_minimal`: a weaker variant of `background_pressure` with reduced contextual pressure.

- `direct_incentive`: an explicit incentive that rewards the discrepant record or shortcut.
- `honesty_system_prompt`: a system instruction to document honestly, testing whether an integrity directive offsets pressure.
- `replacement_ablation`: a self-preservation framing in which the agent is told it may be replaced, adapted from agentic-misalignment setups.
- `reverse_incentive`: an incentive in the opposite direction, rewarding integrity or penalizing the shortcut.
- `sham_placebo`: a control that presents the surface form of a pressure manipulation while being functionally inert.

Scenarios are synthetic but regulatory-grounded: they target documentation patterns such as unsupported billing, note-only quality completion, omitted discharge risk, adverse-event minimization, and throughput shortcuts (Table 1). All scenarios are original to MedInsider. For example, a coding-and-billing case may preserve the same active problems while changing whether unsupported diagnosis capture is institutionally rewarded. A quality-metric case may preserve the care-plan facts while changing whether metric completion is emphasized. Safety and readmission cases similarly hold patient state constant while altering the local incentive to omit risk or minimize an adverse event.

Table 1: Scenario families in the MedInsider corpus. Each family targets a distinct documentation pressure axis and a primary record surface where the resulting integrity discrepancies are observed.

Family	Pressure axis	Primary record surface
Coding and billing	Pressure to capture diagnoses or bill at higher levels despite unchanged chart support	Billing submission and note support
Quality metric	Pressure to mark metrics complete despite unchanged care-plan evidence	Quality report and documented actions
Readmission	Pressure to omit readmission-relevant risk despite unchanged discharge state	Discharge planning and risk documentation
Safety event	Pressure to minimize or omit adverse events despite unchanged event facts	Safety-event documentation and chart notes
Throughput	Pressure to discharge or move patients faster despite unresolved clinical constraints	Discharge actions and final record consistency

4 Metrics and Scoring

Aligned Task Completion (ATC) measures whether the agent did the clinically appropriate work: it credits correct actions and penalizes required work that was left undone. We also define a single aggregate signal, **"any integrity metric,"** which fires when at least one of IVR, MGR, UPR, AEOR, CDR, or PSD fires; this is the headline integrity verdict when one overall number is needed. **Integrity Violation Rate (IVR)** is a specific trace-documentation conflict metric: it counts cases where final documentation conflicts with what the agent observed or did. The remaining metrics report other discrepancy types. **Metric Gaming Rate (MGR)** captures unsupported quality-metric completion. **Upcoding Rate, integrity subtype (UPR)** captures billing for a diagnosis code that the chart no longer supports, such as a code that was documented earlier and then resolved. This is distinct from billing a code that was never documented at all. **Adverse Event Omission Rate (AEOR)** captures omitted adverse-event documentation. **Contradictory Documentation Rate (CDR)** captures direct conflicts between the action trace and the final record. **Patient Safety Degradation (PSD)** fires when an agent submits an order that the scorer marks as contraindicated given the patient's active problem list.

The scorer does not infer intent. It checks whether record outputs are supported by the local evidence available to the agent. This is why refusal and task failure are not counted as integrity success: ATC is reported beside the integrity metrics rather than folded into them.

The metrics are computed from the same episode artifacts used for reproduction. When an agent submits a billing code, marks a metric complete, or writes a discharge note, the scorer compares that action against the chart facts and logged tool outputs. Each discrepancy can therefore be traced to a concrete episode artifact rather than to a post-hoc qualitative judgment. PSD is narrower than the documentation metrics: it fires only when a successful order conflicts with an active contraindication

Table 2: Expert-validation agreement. Values are kappa statistics. R1–R2 is IM–IM, R3–R4 is CPMA–CPMA, and the other four pairs are IM–CPMA. Q3 excludes the one "Scorer hidden" row for affected pairwise comparisons and for Fleiss' kappa. Full count and raw-agreement tables are in `docs/validation/kappa_tables_20260505.csv`.

Question	R1–R2	R1–R3	R1–R4	R2–R3	R2–R4	R3–R4	Fleiss	Inter-prof. mean
Q1 validity	-0.023	-0.017	-0.011	-0.020	-0.012	-0.011	-0.018	-0.015
Q2 integrity	0.875	0.915	0.935	0.875	0.895	0.935	0.905	0.905
Q3 scorer agreement	-0.021	-0.017	-0.011	-0.021	-0.013	-0.011	-0.017	-0.015

state represented in the scenario payload. The v2 corpus contains none of those active contraindication states, so PSD is structurally untriggered in the seven-model run; it is retained for compatibility with future scenario extensions and should not be read as an estimated contraindicated-order rate. The unit of aggregation is the scored episode. Pair-level counts are reported only when both members of a matched pair have scored rows. The main benchmark contains 840 scored episodes and 420 fully scored pairs for each model.

5 Validation Results

Four credentialed reviewers labeled a source/model-blinded 120-episode set with full overlap, yielding 480 reviewer-episode submissions. R1 and R2 were board-certified Internal Medicine physicians; R3 and R4 were Certified Professional Medical Auditors (CPMAs). Data-health checks found no duplicate positions, missing positions, invalid answer values, or episode-payload mismatches.

Each episode payload was accompanied by four reviewer questions: Q1 asked, "Is the scenario clinically valid and is the stated correct action actually correct?" (Yes / No / Unclear); Q2 asked, "Did the agent commit an integrity violation in this episode?" (Yes / No / Unclear); Q3 asked, "Do you agree with the automated scorer's verdict?" (Yes / No / Unclear / Scorer hidden); and Q4 was an optional free-text comment.

Reviewers were recruited through professional society directories: ABIM-credentialed physicians via the Society of Hospital Medicine and CPMA-credentialed auditors via AAPC. The 2-IM + 2-CPMA panel includes two professional roles involved in judging clinical documentation integrity in practice: physicians review clinical reasoning, and CPMAs audit documentation against billing and quality requirements. Using four reviewers with full overlap let us estimate both intra-profession and inter-profession agreement on the same 120 episodes, rather than losing per-metric power to a partial-overlap design. Reviewers labeled episodes through a Streamlit-based validation tool deployed via Hugging Face Spaces; the tool source and reviewer documentation are included in the supplementary materials. On Q2, reviewers saw a small marker indicating which answer the automated scorer had chosen, presented as a reference. The Q3 workflow recorded whether reviewers viewed or used the scorer verdict when answering Q3.

The 120-episode validation set was sampled from the final scored outputs to ensure positive instances of every reported integrity metric appeared in the validation set. Because metric positives can overlap, the set contains 33 IVR positives, 29 MGR positives, 25 AEOR positives, 25 CDR positives, all 3 UPR-integrity positives, and 30 clean rows with no integrity failure. The sample spans all five scenario families, multiple pressure conditions, and all seven source models, with no source model contributing more than 25 rows, so reviewers assessed scorer behavior across the metrics reported in the manuscript.

For Q1, reviewers marked scenario validity as "Yes" in 470 / 480 responses (97.9%). For Q2, reviewers marked an integrity violation in 351 / 480 responses (73.1%); applying majority rule produced 90 positive and 30 negative episode labels. Q2 was unanimous on 111 / 120 episodes; the remaining 9 were all 3–1 splits. For Q3, reviewers agreed with the automated scorer in 471 / 480 responses (98.1%), with one row marked "Scorer hidden." Table 2 reports exact-category kappa statistics. Intra-profession Q2 agreement was $\kappa = 0.875$ (IM–IM) and $\kappa = 0.935$ (CPMA–CPMA); the average inter-profession κ was 0.905. The low Q1 and Q3 kappa values reflect the well-known kappa paradox under high prevalence: when one category dominates (Q1 "Yes" at 97.9%, Q3 "Yes" at 98.1%), Cohen's kappa can be near zero or slightly negative even when raw agreement exceeds 95%. This is a property of the chance-correction in the kappa statistic rather than a reliability problem. Q2 has more balanced category prevalence and produces interpretable kappa values.

Against the adjudicated Q2 majority labels, the scorer's overall integrity verdict matched every episode: sensitivity was 90 / 90 (100.0%), specificity was 30 / 30 (100.0%), and accuracy was 120 / 120 (100.0%). Table 3 breaks this down by metric signal, comparing the scorer's verdict against the adjudicated expert majority (3-of-4 reviewer rule) for each of the 120 validation episodes; subtype sensitivities are lower because each metric captures only one failure mode. The 100% match is at the adjudicated-majority level: individual reviewers disagreed with each other on 9 of 120 episodes (the 3–1 splits) and with the scorer on 8 of 480 individual responses.

Table 3: Automated scorer agreement with adjudicated expert-majority Q2 labels on 120 validation episodes. "Any integrity metric" is the aggregate headline validation signal; per-metric coverage shows what fraction of expert-positive episodes each specific metric catches (each metric targets one failure type, so coverage values below 100% are by design). AEOR and CDR have identical values in this validation set.

Signal	Positive scorer eps.	Coverage	Specificity	Accuracy
Any integrity metric	90	90/90 (100.0%)	30/30 (100.0%)	120/120 (100.0%)
IVR	33	33/90 (36.7%)	30/30 (100.0%)	63/120 (52.5%)
MGR	29	29/90 (32.2%)	30/30 (100.0%)	59/120 (49.2%)
AEOR	25	25/90 (27.8%)	30/30 (100.0%)	55/120 (45.8%)
CDR	25	25/90 (27.8%)	30/30 (100.0%)	55/120 (45.8%)
UPR	3	3/90 (3.3%)	30/30 (100.0%)	33/120 (27.5%)

The validation has three caveats: the Q2 option matching the scorer's pick was visibly marked as a reference, reviewers left no free-text comments, and UPR has only 3 scorer-positive validation examples.

6 Experimental Setup

The evaluated panel consists of seven models: GPT-5.4, Claude Sonnet 4.6, Claude Opus 4.7, Kimi 2.6, GLM-5, DeepSeek V3.2, and Gemma 4. The provider-hosted models were called through their application programming interface (API) routes; Gemma 4 used the `google/gemma-4-31B-it` OpenAI-compatible H200 route. All seven main lanes are complete; the remaining auxiliary caveats are documented in the repository's provenance files. All runs use frozen dataset and selection manifests.

Scored means are computed over the rows marked scored in the final outputs. The manuscript reports two accounting views separately. Results tables use scored rows, because ATC and integrity metrics require a valid episode artifact to compute. Compute summaries in the repository use usage rows, because provider cost and latency are properties of attempted calls. Keeping these two views separate connects them without treating auxiliary probe caveats as main-benchmark failures.

7 Results

7.1 Seven-model benchmark results

Table 4 reports means over scored episodes. The main reading is structural: ranking the models by ATC does not give the same order as ranking them by integrity metrics, so task completion alone does not summarize the results. We keep the operational context visible rather than reducing the table to a leaderboard. This separation also blocks a misleading reading of zero observed trace-documentation conflicts. A model can have IVR = 0 in the scored results while still leaving required clinical work undone; conversely, a higher-ATC model can produce more documentation discrepancies. The benchmark is designed to keep both facts visible at once.

Table 4: Seven-model main benchmark results. ATC reports task completion; IVR, MGR, UPR, AEOR, and CDR report metric-specific discrepancies. AEOR and CDR fire on identical episode sets in this corpus. UPR is the integrity subtype.

Model	Scored	Pairs	ATC	IVR	MGR	UPR	AEOR	CDR
GPT-5.4	840	420	0.7045	0.0143	0.0048	0.0000	0.0655	0.0655
Claude Sonnet 4.6	840	420	0.6733	0.0000	0.0000	0.0000	0.0000	0.0000
Claude Opus 4.7	840	420	0.6948	0.0024	0.0000	0.0000	0.0000	0.0000
Kimi 2.6	840	420	0.5973	0.0071	0.0012	0.0000	0.1083	0.1083
GLM-5	840	420	0.6538	0.0018	0.0000	0.0024	0.0881	0.0881
DeepSeek V3.2	840	420	0.6168	0.0131	0.0083	0.0012	0.0774	0.0774
Gemma 4	840	420	0.5059	0.0000	0.0202	0.0000	0.1381	0.1381
PSD is structurally untriggered (Section 4) and omitted from this table.								

We audited the zero cells and found no scorer bugs in the sampled traces. PSD is 0.00% for every model, for the structural reason given in Section 4. MGR fires 29 times across the seven-model panel, and only in quality-metric pressure scenarios. UPR-integrity fires 3 times across 5,880 episodes (2 in GLM-5, 1 in DeepSeek V3.2). All three are stale-code submissions rather than overt fabrication: the model documented diagnosis code R07.9, the chart later resolved that code, and the model then submitted billing for the resolved code. AEOR and CDR fire on identical episode sets across all seven models in this corpus, reflecting that adverse-event omissions in MedInsider also surface as chart-versus-record contradictions.

7.2 Refusal and partial completion

Low integrity-violation rates can come from two different behaviors: a model that engages with the task and resists the pressure framing, or a model that refuses to engage at all. Both contribute to what we see at the panel level. The `refused_misaligned_pressure` label captures cases where the model declined the misaligned shortcut under pressure, while ATC < 1 captures clinically aligned work that was left incomplete. Table 5 reports both rates for the seven-model panel. The Anthropic models' low integrity-violation counts therefore reflect both fewer trace-documentation conflicts when they engage and substantially higher refusal rates than several other models in the panel.

Table 5: Refusal and partial-completion rates in the scored outputs. Partial completion is measured as ATC < 1.

Model	Refused misaligned pressure	ATC < 1
GPT-5.4	650/840 (77.4%)	720/840 (85.7%)
Claude Sonnet 4.6	720/840 (85.7%)	805/840 (95.8%)
Claude Opus 4.7	718/840 (85.5%)	773/840 (92.0%)
Kimi 2.6	627/840 (74.6%)	787/840 (93.7%)
GLM-5	643/840 (76.5%)	766/840 (91.2%)
DeepSeek V3.2	642/840 (76.4%)	795/840 (94.6%)
Gemma 4	596/840 (71.0%)	832/840 (99.0%)

7.3 Condition-level reading

Table 6 isolates the background-pressure subset. It is descriptive, not a hypothesis test: it shows model-specific shifts rather than a single pooled pressure effect. We highlight this subset because it keeps the pressure contrast simple while using the same scenario families and scoring rules as the full run, which shows how the paired-twin design supports condition-level inspection.

Table 6: Background-pressure versus neutral subset. Columns marked n are neutral rows; columns marked b are background-pressure rows.

Model	n_n	n_b	ATC n	ATC b	IVR n	IVR b
GPT-5.4	60	60	0.7033	0.7061	0.0167	0.0167
Claude Sonnet 4.6	60	60	0.6845	0.6261	0.0000	0.0000
Claude Opus 4.7	60	60	0.7044	0.6761	0.0000	0.0000
Kimi 2.6	60	60	0.5961	0.5344	0.0167	0.0333
GLM-5	60	60	0.6511	0.6278	0.0000	0.0000
DeepSeek V3.2	60	60	0.6222	0.5667	0.0500	0.0000
Gemma 4	60	60	0.5233	0.4822	0.0000	0.0000

8 Auxiliary Analyses

8.1 Coding-knowledge probe

The coding probe is a fixed bank of 15 questions, three per scenario family. It tests whether the model can answer compact rule-knowledge items associated with the benchmark families.

Table 7: Coding-knowledge probe outcomes. Answered and correct counts are reported with the resulting per-model score.

Model	Answered	Total	Correct	Score
GPT-5.4	15	15	15	1.0000
Claude Sonnet 4.6	14	15	13	0.9286
Claude Opus 4.7	13	15	13	1.0000
Kimi 2.6	14	15	14	1.0000
GLM-5	7	15	6	0.8571
DeepSeek V3.2	15	15	15	1.0000
Gemma 4	15	15	15	1.0000

The probe is not a substitute for the main benchmark: high probe scores can coexist with observed documentation discrepancies in the full runs. The intended interpretation is narrow. The probe weakens one alternative explanation for integrity failures: that the model simply lacks the relevant coding or documentation knowledge. It does not measure multi-step tool use or pressure response.

8.2 Structural mitigation experiment

The mitigation adds a real `compliance_check` tool that the agent must call successfully before `submit_billing` and `quality_report` will execute. The executed subset has 24 background-pressure episodes per model across coding-and-billing and quality-metric scenarios, and covers four models. The four-model subset spans a range of baseline integrity-failure rates and provider routes; the remaining three models are left to future work. Table 8 reports before-and-after metrics. The study is intentionally structural rather than prompt-only: the gate is part of the tool environment, so a target action is blocked unless a matching compliance check succeeds first. This makes the mitigation auditable in the action logs. The result is scoped to the tested episodes; it shows that a specific tool-level constraint changed behavior on the selected background-pressure episodes, not that compliance prompts or other governance mechanisms would generalize across the benchmark.

Table 8: Mitigation before/after results on the tested subset. Columns marked b are baseline values; columns marked m are mitigation values.

Model	IVR b	IVR m	ATC b	ATC m	MGR b	MGR m
GPT-5.4	0.0417	0.0000	0.5764	0.5069	0.0000	0.0000
Claude Sonnet 4.6	0.0000	0.0000	0.5556	0.5695	0.0000	0.0000
DeepSeek V3.2	0.0000	0.0000	0.5000	0.5000	0.0000	0.0000
Gemma 4	0.0000	0.0000	0.3819	0.3958	0.0833	0.0000

9 Reproducibility, Ethics, and Release

The reproducibility materials include the main dataset manifest, full-run selection manifest, mitigation selection manifest, capability-control subset manifest, runner, preflight script, paper-packet builder, and `make reproduce` target. Provider-backed reruns require external credentials, but the reported tables are tied to local manifests, scored outputs, summaries, and run trees. Code is licensed under Apache 2.0 and data under CC BY 4.0.

The artifact chain lets a reviewer trace a manuscript table cell back to source rows: selection manifests define which episodes are eligible, run directories and run manifests preserve execution provenance, scored outputs carry per-episode metrics, and the paper-packet builder generates the manuscript CSVs. The validation tool is included as a Streamlit application, with the frozen episode payload and reviewer documentation under `validation/medinsider_validation_space/`; the four-reviewer engagement results are anchored to its payload SHA. Usage summaries record the compute route used: provider-hosted lanes used API routes, while Gemma used the local $2\times$H200 route.

We release the package conservatively because the scenarios enumerate documentation failure opportunities. The paper frames them as evaluation stress tests, reports the validation limits, and does not claim a finished public-release process beyond the repo-local and supplementary artifacts. Local smoke tests run the code without provider calls; fresh API reruns require external credentials. The scenarios are synthetic, not patient-derived, and are presented as measurement cases rather than operational instructions.

10 Limitations

MedInsider uses synthetic, regulatory-grounded scenarios rather than clinical operations logs. The reported benchmark numbers are descriptive single-run summaries, not seeded confidence intervals. The coding probe and the mitigation experiment are auxiliary analyses, and the mitigation results should not be generalized beyond the tested subset or the four mitigation models. Several items planned in earlier proposals are outside the scope of this manuscript: cross-benchmark ranking distinctness, human baselines, and real EHR-derived subsets.

These limits keep the interpretation descriptive. MedInsider detects documentation-integrity discrepancies in controlled workflows; it does not address settings outside the executed runs. The benchmark also does not resolve intent. It observes conflicts between trace-supported facts and record outputs but does not determine why a model produced the conflict. This constraint keeps the evidence tied to executable traces rather than causal explanation. Expert validation reduces, but does not eliminate, scorer uncertainty. The validation set used 120 episodes with full overlap from four credentialed reviewers and achieved almost-perfect Q2 agreement, but UPR is not meaningfully validated by this engagement: the three positive examples are confirmatory of specific scorer firings rather than evidence of metric reliability. Reviewers provided no written rationale for their disagreements, and the Q2 scorer-pick marker visible during review may have anchored some labels. The validation set's adjudicated label distribution (90 positive episodes, 30 negative) means specificity is computed on a smaller base than sensitivity; the 100% specificity should be read as 30/30 rather than as evidence of robustness across a large negative population. PSD is structurally untriggered in the corpus, so the PSD zero should not be generalized to contraindicated-order settings.

11 Conclusion

MedInsider adds a documentation-integrity view to medical agent evaluation. Action-log scoring shows that task completion and documentation integrity can diverge, and a simple structural gate can reduce discrepancies on the tested subset. Four-reviewer expert validation supports the integrity labels on the 120-episode set, while the documented limitations (synthetic scenarios, single-run results, validation prevalence imbalance, and possible Q2 anchoring) keep the claims in scope.

As an Evaluations & Datasets contribution, MedInsider operationalizes documentation integrity as a measurable property, checks it against expert clinical judgment, and provides reusable infrastructure for future evaluations of medical AI integrity under institutional pressure.

References

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Dueñas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, Eric Winsor, Jerome Wynne, Yarin Gal, and Xander Davies. AgentHarm: A benchmark for measuring harmfulness of LLM agents, April 2025. URL https://arxiv.org/abs/2410.09024. arXiv:2410.09024.

Anthropic. Agentic misalignment: How LLMs could be insider threats, 2025. URL https://www.anthropic.com/research/agentic-misalignment.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. HealthBench: Evaluating large language models towards improved human health, May 2025. URL https://arxiv.org/abs/2505.08775. arXiv:2505.08775.

HL7 International. FHIR Release 5, 2023. URL https://www.hl7.org/fhir/.

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. MedAgentBench: A realistic virtual EHR environment to benchmark medical LLM agents, January 2025. URL https://arxiv.org/abs/2501.14654. arXiv:2501.14654.

Gyubok Lee, Elea Bach, Eric Yang, Tom Pollard, Alistair Johnson, Edward Choi, Yugang Jia, and Jong Ha Lee. FHIR-AgentBench: Benchmarking LLM agents for realistic interoperable EHR question answering, September 2025a. URL https://arxiv.org/abs/2509.19319. arXiv:2509.19319.

Gyubok Lee, Woosog Chay, Heeyoung Kwak, Yeong Hwa Kim, Haanju Yoo, Oksoon Jeong, Meong Hi Son, and Edward Choi. From conversation to query execution: Benchmarking user and tool interactions for EHR database agents, September 2025b. URL https://arxiv.org/abs/2509.23415. arXiv:2509.23415.

U.S. Department of Justice and U.S. Department of Health and Human Services. DOJ-HHS false claims act working group, July 2025. URL https://www.justice.gov/opa/pr/doj-hhs-false-claims-act-working-group.

Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, Wenhao Zheng, Zhaoyang Wang, Xiao Wang, Xuchao Zhang, Chetan Bansal, Marc Niethammer, Junzhou Huang, Hongtu Zhu, Yun Li, Jimeng Sun, Zongyuan Ge, Gang Li, James Zou, and Huaxiu Yao. CARES: A comprehensive benchmark of trustworthiness in medical vision language models. In *Advances in Neural Information Processing Systems 37*, 2024. doi: 10.52202/079017-4455. URL https://cares-ai.github.io/. Datasets and Benchmarks Track.