# REPAIRING CLINICAL-AGENT PLANS WITH DATA-INDUCED PROCESS CONSTRAINTS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Tool-using clinical agents can place laboratory, imaging, and medication orders, yet their free-text plans make the care process difficult to verify. Hand-crafted checklists can support auditing, but they transfer poorly across settings and risk circular evaluation. Here, we induce machine-checkable process constraints from a held-out split of MIMIC-IV admissions. The method applies Wilson-bounded prevalence mining to identify required actions, clinical concepts, and temporal precedence, incorporating modality-ordered payloads. A deterministic minimal-edit operator then repairs structured plans using four atomic edits and disease-agnostic logic. Across four cohorts, repair raised hard-constraint satisfaction from 0.55–0.82 to 0.91–1.00 (appendicitis, pulmonary embolism, urinary tract infection, and cholecystitis). By contrast, Reflexion, self-consistency, and an LLM critic recovered little of these gains. Replacing the induced specifications with size-matched donor-cohort specifications reduced macro satisfaction by 0.115 and inserted appendectomies into 38 of 40 pulmonary embolism trajectories. LLM-generated specifications covered less than half of the mined requirements and reduced satisfaction by up to 0.307. Induction requires only tens of admissions and is robust to most threshold settings. These results indicate that verifiable clinical agent behavior can be derived from observed care processes rather than from another language model.

## 1 INTRODUCTION

A tool-using clinical agent may take a history, order laboratory tests and imaging, initiate treatment and reach the correct diagnosis Jiang et al. (2025); Zakka et al. (2024). Yet diagnostic accuracy alone does not establish that the care process was appropriate. When plans are expressed in free text, it is difficult to verify whether required imaging was obtained, whether treatment followed the evidence that justified it, or whether a pathway intended for another condition was applied Singhal et al. (2023); Tu et al. (2025); Hager et al. (2024). Such ordering errors can directly affect care: antibiotics given before blood cultures can reduce diagnostic yield, and procedures initiated without appropriate confirmation can expose patients to avoidable harm. The plan—not only the final diagnosis—must therefore be auditable. Two common approaches are insufficient. Self-reflection, self-consistency and LLM-based critics keep verification within the same model class that generated the error, despite evidence that models rarely self-correct without external feedback (Shinn et al., 2023; Wang et al., 2022; Madaan et al., 2023; Huang et al., 2024). Hand-authored checklists provide an external standard but require disease-specific engineering and can make evaluation circular when the same rules define both the intervention and its success criterion.

We instead induce process constraints from observed care. Using MIMIC-IV admissions disjoint from evaluation (Johnson et al., 2023), we derive a declarative specification for each cohort: required tool actions and clinical concepts, diagnostic-to-therapeutic temporal precedences, and modal order payloads that translate abstract requirements into concrete ordersFerber et al. (2026). A deterministic operator then repairs each structured plan using four atomic edits chosen to minimize edit cost. The repair algorithm is identical across cohorts; the only hand-authored clinical component is a disease-agnostic map from raw order strings to normalized concepts, which encodes no disease-specific requirements (Figure 1).



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

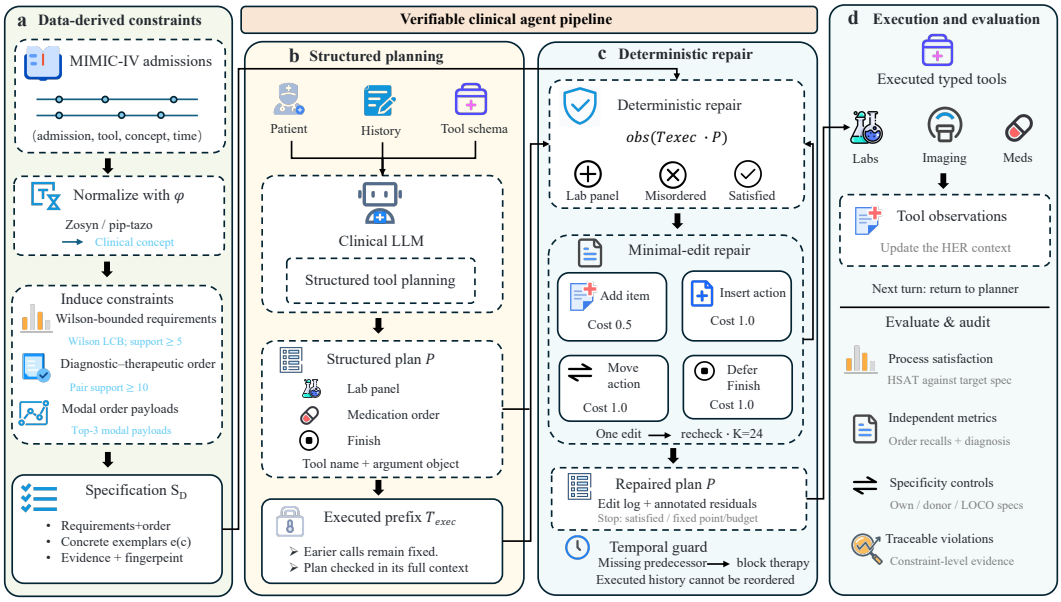


Figure 1: **Overview of the proposed framework.** (**a**) Process constraints and order exemplars are induced from clinical records. (**b**) The clinical agent generates a structured plan. (**c**) A deterministic operator verifies and repairs the plan using four atomic edits while preserving executed history. (**d**) Repaired plans are executed and evaluated for constraint satisfaction and independent clinical benchmark metrics.

Our evaluation separates specification compliance from independent order recall and secondary medication outcomes. Archived cohort-routed experiments compare repair with plain planning and a no-specification critic, with reflection and self-consistency added for appendicitis. We also swap donor-cohort or model-authored specifications into the repair step while keeping the scoring specification fixed, which asks whether the induced content matters beyond a second planning pass.

**Contributions.** We contribute (i) induced specifications that couple checkable constraints with executable order payloads, so they are directly actionable; (ii) a deterministic repair operator that respects mined diagnostic-to-therapeutic staging and terminates under an explicit budget; and (iii) fixed-target controls that separate the value of induced content from that of the repair mechanism. The resulting standard describes observed practice, not necessarily appropriate care for every patient.

## 2 Related Work

**Clinical agents.** Medical agents combine specialist collaboration, adaptive recruitment and simulated encounters (Tang et al., 2024; Kim et al., 2024; Schmidgall et al., 2024). EHR-based systems extend this setting to executable actions (Shi et al., 2024; Jiang et al., 2025; Zakka et al., 2024). Our host environment evaluates agents against orders documented during an admission (Ferber et al., 2026). We retain that environment and its recall measures, while representing requirements and temporal relations as explicit constraints.

**Self-correction and constrained generation.** Reflexion uses verbal feedback, self-consistency aggregates sampled reasoning, and ReAct interleaves reasoning with actions (Shinn et al., 2023; Wang et al., 2022; Yao et al., 2022). External tool feedback can improve revision (Gou et al., 2024), whereas intrinsic correction can fail (Huang et al., 2024). Learned verification and constrained decoding address complementary problems of candidate selection and output validity (Cobbe et al., 2021; Poesia et al., 2022; Beurer-Kellner et al., 2023; Scholak et al., 2021; Geng et al., 2023). Our checker instead evaluates semantic requirements and precedences over tool calls.

**Process discovery and runtime constraints.** Declarative process discovery already mines temporal constraints from event logs, including methods that prune redundant or trivially satisfied rules (Maggi et al., 2012). Healthcare process mining provides the application context (Rojas et al., 2016), so constraint discovery alone is not what is new here. AgentSpec provides a DSL for runtime enforcement and studies model-generated rules (Wang et al., 2026). Agent-C combines temporal specifications, SMT checking and constrained generation to enforce compliant actions (Kamath et al., 2025). What we study instead is the joint induction of cohort requirements and executable order payloads, together with fixed-target specificity and failure analyses. Unlike enforcement at action generation, our repaired plans remain subject to downstream filtering and model execution; we do not claim equivalent runtime guarantees.

# 3 METHOD

## 3.1 PLANS AND OBSERVATIONS

Let $\tau_{\text{exec}}$ denote executed tool calls and $P = (a_1, \ldots, a_m)$ the remaining plan, with action $a_i = (T_i, x_i)$ comprising a tool and its arguments. Repair modifies only $P$. Tools accept list-valued laboratory, urine, microbiology and medication orders, or individual radiology and procedure requests.

A shared normalisation map $\varphi$ maps clinical strings to concepts in six namespaces: `med`, `lab`, `urine`, `micro`, `rad` and `proc`. It also associates concepts with host tools. The vocabulary is authored once across cohorts and encodes terminology rather than disease-specific requirements. Administrative artefacts, universal admission routines and home-medication reconciliation are excluded by recorded filtering metadata.

Write $c' \preceq c$ when $c'$ equals or specialises $c$, and let $\kappa(a)$ extract concepts from an action's arguments. The observation comprises the tools used, $\mathcal{T}_\tau$, and first-occurrence indices:

$$\sigma_\tau(c) = \min\{i : \exists c' \in \kappa(a_i),\ c' \preceq c\}, \qquad \min \emptyset := \infty. \tag{1}$$

Verification uses $\text{obs}(\tau_{\text{exec}} \cdot P)$, so an executed occurrence takes precedence over a later planned occurrence.

## 3.2 CONSTRAINT SEMANTICS AND SCORING

A cohort specification $\mathcal{S}_D$ contains typed requirements, avoidances and precedences:

$$\begin{aligned} \tau \models \texttt{require_action}(T) &\iff T \in \mathcal{T}_\tau, \\ \tau \models \texttt{require_concept}(c) &\iff \sigma_\tau(c) < \infty, \\ \tau \models \texttt{avoid_concept}(c) &\iff \sigma_\tau(c) = \infty, \\ \tau \models \texttt{precede}(b, a) &\iff \sigma_\tau(b) \le \sigma_\tau(a) \quad \text{if both occur.} \end{aligned} \tag{2}$$

Requirements and avoidances always apply. A precedence applies only when both endpoints occur; missing endpoints are assessed through separate requirements. Concepts in the same action satisfy non-strict precedence. Each constraint records hard/soft severity, data/guideline provenance and available supporting evidence.

For the nonempty set of applicable hard constraints $\mathcal{A}_H(\tau)$, define

$$\text{HSAT}(\tau; \mathcal{S}_D) = \frac{\sum_{\gamma \in \mathcal{A}_H(\tau)} \mathbf{1}[\tau \models \gamma]}{|\mathcal{A}_H(\tau)|}. \tag{3}$$

A content fingerprint identifies each specification. Given the vocabulary and checker version, its score and violation trace are deterministic.

## 3.3 INDUCTION FROM ADMISSIONS

Induction excludes every frozen evaluation identifier and records split hashes. Each of the $N$ remaining admissions becomes events $(h, T, c, t)$ using the host benchmark's 24-hour extraction window. For admission-level support $n$ and $\hat{p} = n/N$, the Wilson lower bound is

$$\underline{p}(n, N) = \frac{\hat{p} + z^2/(2N) - z\sqrt{\hat{p}(1-\hat{p})/N + z^2/(4N^2)}}{1 + z^2/N}, \qquad z = 1.96. \tag{4}$$

Concept severity is assigned by

$$\mathrm{sev}(c) = \begin{cases} \textsc{hard}, & n_c \geq 5,\ \underline{p}(n_c, N) \geq 0.50, \\ \textsc{soft}, & n_c \geq 5,\ 0.20 \leq \underline{p}(n_c, N) < 0.50, \\ \text{discarded}, & \text{otherwise.} \end{cases} \tag{5}$$

Tool requirements use thresholds $0.60$ and $0.25$ with the same support floor.

For diagnostic $b$ and therapeutic $a$, let $N_{ba}$ count admissions containing both, and $k_{ba}$ those recording $b$ before $a$. Retain the pair when

$$N_{ba} \geq 10, \qquad \underline{p}(k_{ba}, N_{ba}) \geq 0.80; \tag{6}$$

label it hard at a bound of at least $0.90$. Date-only procedures use day-level comparisons. Predecessors must already be hard diagnostic requirements; therapeutic endpoints use their most general retained concept level. Appendix B gives the deduplication and tool-injection rules.

Guideline assertions receive provenance annotations. Negated concepts are excluded from requirement promotion; those with empirical prevalence $\hat{p}_c \leq 0.10$ become soft avoidance candidates. This channel produced no avoidance and changed no hard concept requirement in the evaluated four-cohort ablation (Appendix L).

For retained concept $c$, let $\mathcal{E}_c$ collect observed payload occurrences and $X_c$ their distinct values. The modal exemplar is

$$e(c) \in \arg\max_{x \in X_c} \sum_{x' \in \mathcal{E}_c} \mathbf{1}[x' = x]. \tag{7}$$

Up to three frequent payloads retain executable fields such as drug, dose, route, laboratory label or imaging region. They provide population-derived order arguments, not patient-specific prescribing rules.

### 3.4 Deterministic greedy repair

Repair uses four edits: ADD_ITEM appends an exemplar to a compatible list-valued action at cost $0.5$; INSERT_ACTION, MOVE_ACTION and DEFER_FINISH each cost $1.0$. The operator scans hard violations in fixed specification order, emits at most one admissible edit, and recomputes the observation. It prefers appending over inserting.

A therapeutic injection requires all mined predecessors to occur in the executed prefix or current plan. Subsequent precedence checks can move an unexecuted predecessor before its successor. Executed calls are immutable; existing avoided orders cannot be deleted. Repair stops when no hard violation remains, a full scan produces no edit, or $K = 24$ iterations have elapsed. A final check removes a premature `Finish` when an unblocked requirement remains open.

The budget guarantees termination. Adding or moving orders preserves satisfied presence requirements, subject to the stated `Finish` condition. Neither aggregate-score monotonicity nor global minimum edit cost is guaranteed (Appendix C). Residuals distinguish immutable history, missing exemplars, blocking and budget exhaustion. Optional item alignment is evaluated separately (§5.4).

### 3.5 Interpretation

Repair and scoring use the same specification, so improved HSAT measures compliance rather than independent clinical quality. We therefore also report case-specific order recall. Evaluation-case orders are not repair targets; requirements and exemplars come from disjoint admissions. Cohort-level prevalence nevertheless cannot determine whether an order is appropriate for an individual patient.

## 4 Experimental setup

**Data and model.** We use MIMIC-IV through the host benchmark's extraction (Johnson et al., 2023; Ferber et al., 2026). Eight cohorts contain 1,103 induction and 527 held-out admissions, with 15–497 induction admissions per cohort (Appendix D). Live evaluation uses 170 admissions: appendicitis, pulmonary embolism, urinary tract infection and cholecystitis. Lists are excluded from induction. The additional four cohorts support offline specificity analyses. Every arm uses `deepseek-v4-flash`; the same backbone supplies the host's medication judge.

Table 1: Live satisfaction and paired gains over the critic. Intervals are 95% bootstrap intervals from the archived multiple-comparison analysis. All gains survive Holm correction.

| Cohort | Plain | Critic | Ours | Δ vs critic (95% CI) |
|---|---|---|---|---|
| Appendicitis | 0.824 | 0.848 | 0.917 | $+0.069$ [+0.040, +0.097] |
| Pulmonary embolism | 0.550 | 0.561 | 1.000 | $+0.439$ [+0.410, +0.471] |
| Urinary tract infection | 0.756 | 0.748 | 0.991 | $+0.244$ [+0.204, +0.288] |
| Cholecystitis | 0.737 | 0.759 | 0.986 | $+0.227$ [+0.190, +0.267] |

**Comparison arms.** VERIFIABLE PLAIN, abbreviated *plain*, uses structured planning without repair. CONSTRAINT, abbreviated *ours*, adds the induced specification and deterministic repair. REFLEXION retains up to four within-episode verbal lessons; cases are not replayed. SELF CONSISTENCY samples three plans at temperature $0.7$ and votes over tool–concept pairs. The archived LLM CRITIC receives truncated versions of patient context, executed calls, tool schemas and the draft plan, but no induced specification. It proposes additions and whether to defer finishing. These controls therefore measure the combined value of induced evidence and deterministic repair, not either ingredient on its own. Appendicitis includes these five arms and a separate CONSTRAINT ALIGNED arm; the other live cohorts compare plain, critic and ours. Appendix A distinguishes these experiments from legacy configurations.

**Outcomes and inference.** The archived primary outcome is deterministic issued-call hard satisfaction. External-to-specification outcomes are string-matched blood, urine, radiology, procedure and microbiology recall. The archived INDEPENDENT family also contains diagnosis match, retained for statistical provenance but not interpreted as diagnosis-blind evidence. Secondary medication outcomes comprise LLM-judged name and class recall, and a deterministic, hand-authored medication-bundle score. The bundle measures keyword-based coverage of four appendicitis medication categories, not patient-specific appropriateness. Contrasts pair admissions with both applicable measurements; denominators differ when evaluation is incomplete or a metric is inapplicable. Marginal arm means therefore need not subtract to the paired delta.

Intervals use 5,000 admission-level bootstrap resamples. Two-sided paired sign-flip tests enumerate all sign assignments for at most 15 nonzero differences and otherwise use 20,000 Monte Carlo draws. Holm and Benjamini–Hochberg corrections apply separately to primary (8), independent (48), other checker-derived (24) and secondary medication (18) tests. The last family retains the archived label JUDGED, although its two bundle contrasts are deterministic. We keep these archived families and report family-wise significance separately from FDR-only findings. The 98-test paired analysis and aggregate arm summaries accompany the manuscript as CSV and JSON files.

**Offline replay.** Specificity and component analyses repair saved plain trajectories while holding the original model output fixed. Existing executed calls cannot be reordered. Replay isolates changes to repair inputs but is not a theoretical lower bound on live performance, which can also change subsequent planning. We distinguish replay from live results throughout.

# 5 RESULTS

## 5.1 INDUCED REPAIR IMPROVES TOOL-CALL COMPLIANCE

Induced repair increased appendicitis hard satisfaction from $0.824$ to $0.917$, a paired gain of $0.093$ (95% CI [+0.063, +0.123]; 37 wins, 10 losses, 3 ties). Open hard violations fell from $5.46$ to $3.10$ per case (Table 4). Reflexion and self-consistency did not improve plain planning: their paired deltas were $-0.015$ [-0.046, +0.014] and $-0.003$ [-0.042, +0.033]. The critic gained $0.024$ [-0.001, +0.050]. These intervals cover the backbone and adaptation we actually ran, not self-correction methods in general.

Across the other three cohorts, repair reached $0.986$–$1.000$ (Table 1). All eight primary contrasts against plain or critic survived Holm correction ($p = 0.0004$). Against the critic, wins/losses/ties were 35/12/3, 40/0/0, 40/0/0 and 39/1/0 for appendicitis, pulmonary embolism, urinary tract infection and cholecystitis, respectively.

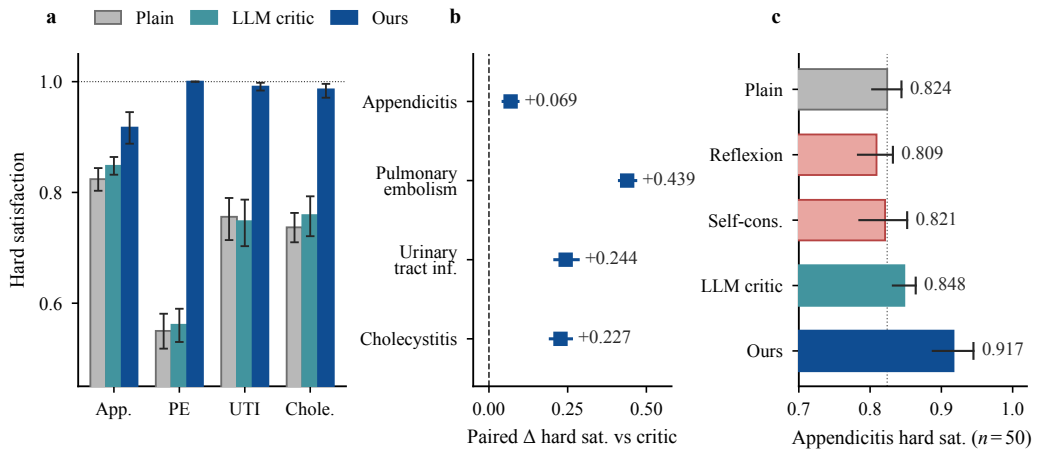


Figure 2: **Archived issued-call hard satisfaction.** Scores do not require successful tool returns; cohort labels select repair specifications. (**a**) Means and 95% admission-bootstrap intervals for four cohorts ($n = 50$ appendicitis; $n = 40$ otherwise). (**b**) Paired gains over the critic; all four survive Holm correction in the eight-test primary family. (**c**) Appendicitis comparison with reflection and self-consistency. Plain denotes structured planning without repair.

**Independent outcomes.** Blood recall improved in all four cohorts against both comparators (Holm $p = 0.0024$), with gains of 0.146–0.192 over plain. Urinary tract infection microbiology recall improved by 0.144 (Holm $p = 0.019$). The archived pulmonary embolism diagnosis contrast was $+0.417$ (Holm $p = 0.0024$), but cohort-label access prevents interpreting it as independent diagnostic improvement. Appendicitis and cholecystitis microbiology recall, and cholecystitis procedure recall, passed FDR but not Holm correction. Urine recall showed no consistent improvement. Taken together, these outcomes point to better coverage of some documented orders, not uniform improvement across the benchmark; the full paired analysis accompanies the manuscript.

### 5.2 Requirement coverage and precedence in pulmonary embolism

The pulmonary embolism specification contains 21 hard constraints, all actionable through exemplars or injectable descendants. Plain planning scored 0.550, with 8.43 open hard violations per case. Repair added missing intravenous heparin, oral anticoagulation, dextrose-containing fluids and a white-cell differential (Figure 3c). These are mined features of this cohort's records, not universal treatment requirements.

Offline replay reached 0.996 with 6.78 edits on average. The three residual cases violated CT-before-fluid precedence because fluids had already executed. Live repair reached 1.000, which tells us the specification was satisfiable inside the loop, not that the model produced a complete care plan on its own (Appendix F).

### 5.3 Specificity of induced requirements and payloads

Replacing own-cohort with size-matched donor specifications reduced macro replay satisfaction by 0.115. For pulmonary embolism, an appendicitis donor inserted appendectomies into $38/40$ trajectories and reduced satisfaction from 0.996 to 0.719. Leave-one-cohort-out induction reached 0.802. Appendicitis did not separate: own, donor and pooled specifications scored 0.774, 0.772 and 0.771. Frozen-history regressions and overlapping abdominal pathways limit this control's interpretation for appendicitis.

On reconstructed held-out care, the own specification ranked first in seven of eight rows and six of eight columns. It exceeded the mean of other cohorts in every column (Figure 4a; Appendix E). Exploratory cohort identification achieved 0.391 accuracy, above uniform guessing (0.125) but below

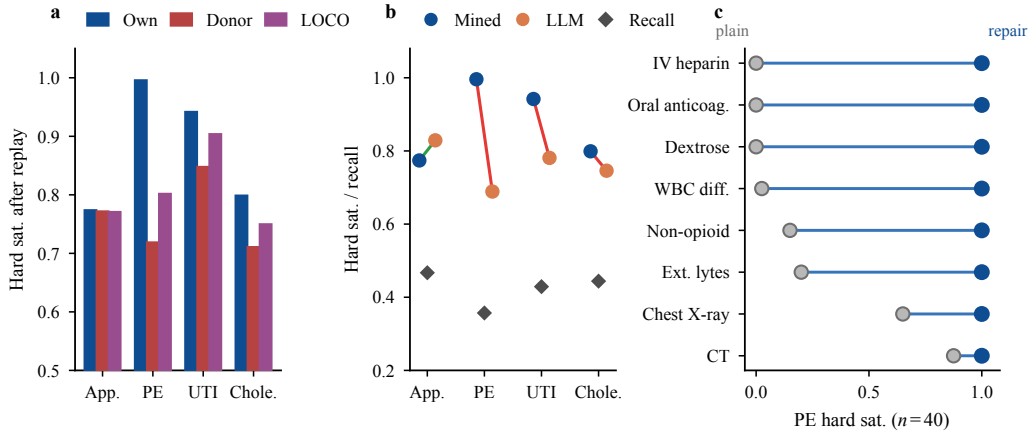


Figure 3: **Specification content changes replay outcomes.** (**a**) Repair driven by own-cohort, size-matched donor or leave-one-cohort-out specifications, always scored against the own-cohort specification. (**b**) Mined versus LLM-authored specifications; circles show recovery of mined hard concepts. (**c**) Selected pulmonary embolism requirements before and after live repair ($n = 40$).

the majority-class baseline (0.503); macro accuracy was approximately 0.572. Discrimination is real but uneven.

**Payload and precedence ablations.** Table 2 isolates repair components on 25 pulmonary embolism replay cases. Removing exemplars returns satisfaction to the plain baseline; LLM-written exemplars recover part of the gain but require more edits. This supports the need for executable payloads, not their uniquely data-derived origin. Removing precedences has no measured effect here, not evidence that they are unnecessary online (Appendix G).

Table 2: **Repair-component ablations.** Offline replay on 25 pulmonary embolism trajectories, scored against the unchanged original specification. Intervals are case-bootstrap 95% CIs. No exemplar removes induced payloads; LLM exemplar substitutes model-authored orders; soft only repairs originally soft constraints; no precedence removes ordering constraints from repair only.

| Configuration | HSAT (95% CI) | Δ vs plain | Mean edits |
|---|---|---|---|
| Plain (no repair) | 0.537 [0.496, 0.576] | — | — |
| Full repair | 0.998 [0.994, 1.000] | +0.461 | 6.72 |
| No precedence | 0.998 [0.994, 1.000] | +0.461 | 6.72 |
| LLM exemplar | 0.817 [0.802, 0.834] | +0.280 | 7.72 |
| Soft only | 0.629 [0.602, 0.656] | +0.092 | 4.76 |
| No exemplar | 0.537 [0.496, 0.576] | +0.000 | 1.00 |

**LLM-authored specifications.** Given the same 71-concept vocabulary and tool bindings, the model recovered 0.36–0.47 of mined hard concepts, with precision 0.78–1.00. Keeping induced payloads but replacing requirements reduced replay satisfaction by 0.053–0.307 on three cohorts. Appendicitis instead improved from 0.774 to 0.829, where fewer injections can avoid activating violated precedences. Model-written payloads further reduced pulmonary embolism satisfaction from 0.689 to 0.555 (Table 11). This latter condition uses no induction-derived payloads; its evaluation still uses MIMIC-derived trajectories and specifications.

### 5.4 Sensitivity, failures and measurement

For pulmonary embolism, 25 induction admissions recovered 0.857 of full-split hard concepts and replay satisfaction 0.912; 50 admissions reached 0.963. Recovery was not strictly monotone, nor were subsampled requirements always a subset of the full specification. Figure 4 shows the curves

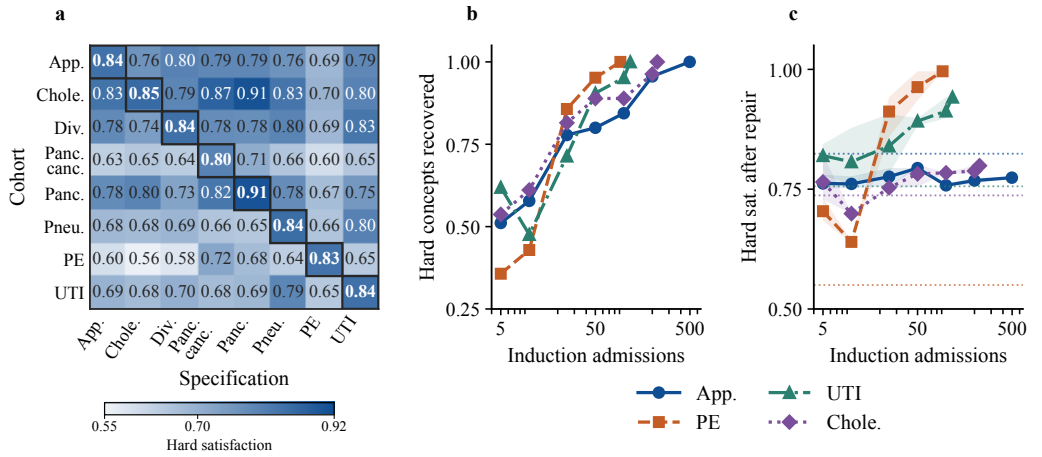


Figure 4: **Cross-cohort specificity and induction sample size.** (**a**) Cross-cohort hard satisfaction on reconstructed documented-care trajectories; rows are trajectory cohorts, columns are induced specifications, and the own-cohort diagonal is outlined. (**b**,**c**) Specifications re-induced from subsampled admissions (3 draws per reduced size; the full split once): fraction of full-split hard concepts recovered, and hard satisfaction after replayed repair. Shaded bands are observed min–max ranges, not confidence intervals; faint lines are per-cohort plain baselines. Recovery and replay scores need not be monotone in sample size.

for all four cohorts; Appendix K gives the full table. Five of six one-at-a-time threshold sweeps left reported satisfaction unchanged; stricter hard promotion reduced it. Some invariances are structural, such as support floors below the support already required by the Wilson bound (Appendix J).

**Precedence failures and cost.** All 12 appendicitis losses against the critic coincided with lower precedence satisfaction. Adding omitted diagnostics after executed treatment can activate an unsatisfied ordering constraint. Appendicitis precedence satisfaction fell by 0.072 versus plain (FDR $q = 0.034$); urinary tract infection also showed a smaller regression. The lesson is to intervene before the first therapeutic call: more coverage does not always raise the score. Offline repair required 4.08–6.78 edits per case. Live repair reduced pulmonary embolism tool calls from 29.1 to 16.6 but added 5.5–9.5 calls in the other cohorts (Appendices M and N).

**Measurement validity.** Re-evaluating ten unchanged conversations three times reproduced every string-matched score. LLM-judged medication-name and medication-class recall had mean spreads of 0.116 and 0.149, respectively (Appendix O). In contrast, the deterministic appendicitis medication-bundle score declined by 0.075 against plain and 0.079 against the critic (both Holm $p = 0.0088$). Judge variability cannot explain those declines, which temper the coverage gains reported above.

String matching can also penalise equivalent orders even when its scores are reproducible. Optional ALIGN_ITEMS unions induced modal items into already-satisfying actions and raises appendicitis urine recall to 0.442. Its live arm also changes satisfaction from 0.917 to 0.928. We report alignment separately because it edits actions that already satisfy their constraints instead of fixing violations.

## 6 Discussion

Data-induced specifications offer an explicit, auditable interface that complements language-model planning with deterministic verification and targeted repair. Across four cohorts, the shared repair mechanism improved issued-call compliance without disease-specific repair code, while blood recall gains provided evidence beyond the optimised process score. Induction also supplied routine requirements and executable order payloads that the evaluated one-shot specification-authoring prompt often omitted. Donor-specification and payload controls indicate that these gains depend on specification content, rather than merely adding another processing stage. Together, these findings

highlight the value of combining flexible model-generated plans with data-grounded requirements and traceable edits.

A further methodological advantage is that explicit specifications make residual failures inspectable, separating missing requirements from ordering violations that frozen histories cannot resolve. This transparency helps identify where earlier intervention, patient-specific contraindication checks and unnecessary-order audits could strengthen the model–verification pipeline. The observed precedence regressions and deterministic medication-bundle declines also identify trade-offs that aggregate compliance gains alone would conceal.

## 7 CONCLUSION

Data-induced requirements, precedences and executable payloads give clinical-agent plans a deterministic standard for checking and repair. In archived cohort-routed experiments, repair improved issued-call compliance over plain planning and a no-specification critic across four cohorts, and over reflection and self-consistency in the appendicitis comparison. Equal-information superiority and successful execution remain to be validated under the revised protocol. Selected independent recalls improved, but medication-bundle coverage and some precedence scores declined. The method makes conformity to an explicit specification verifiable; whether that conformity is clinically appropriate.

### AI USE STATEMENT

The generative AI tool Cursor was used to assist with manuscript writing (including language polishing, logical flow organization and draft revision) and code programming (including code drafting, syntax checking and preliminary debugging). The authors take full responsibility for all scientific claims, data interpretations and the final submitted content of this work.

### ETHICS STATEMENT

This study evaluates simulated interactions with retrospective MIMIC-IV records. Cohort-derived requirements and observed medication payloads are not patient-specific recommendations. The manuscript and accompanying aggregate tables contain no patient-level narratives or admission identifiers.

## REFERENCES

Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. Prompting is programming: A query language for large language models. *Proceedings of the ACM on Programming Languages*, 7 (PLDI):1946–1969, 2023.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Dyke Ferber, Lars Hilgers, Christiane Höper, Benedict Kinny-Köster, Jan-Niklas Eckardt, Katharina Egger-Heidrich, Marius Bill, Martin MK Schneider, Jan Clusmann, Lejla Kadric, et al. Towards autonomous medical artificial intelligence agents. *Nature*, pp. 1–10, 2026.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. Grammar-constrained decoding for structured nlp tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10932–10952, 2023.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Nan Duan, Weizhu Chen, et al. Critic: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, volume 2024, pp. 57734–57811, 2024.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622, 2024.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International conference on learning representations*, volume 2024, pp. 32808–32824, 2024.

Yixing Jiang, Kameron C Black, Gloria Geng, Danny Park, James Zou, Andrew Y Ng, and Jonathan H Chen. Medagentbench: a virtual ehr environment to benchmark medical llm agents. *Nejm Ai*, 2(9): AIdbp2500144, 2025.

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

Adharsh Kamath, Sishen Zhang, Calvin Xu, Shubham Ugare, Gagandeep Singh, and Sasa Misailovic. Enforcing temporal constraints for LLM agents. *arXiv preprint arXiv:2512.23738*, 2025. URL https://arxiv.org/abs/2512.23738.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37: 79410–79452, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

F. M. Maggi, R. P. J. C. Bose, and W. M. P. van der Aalst. Efficient discovery of understandable declarative process models from event logs. In *Lecture Notes in Computer Science*, volume 7328, pp. 270–285. Springer, 2012. doi: 10.1007/978-3-642-31095-9_18.

Gabriel Poesia, Oleksandr Polozov, Vu Le, Ashish Tiwari, Gustavo Soares, Christopher Meek, and Sumit Gulwani. Synchromesh: Reliable code generation from pre-trained language models. *arXiv preprint arXiv:2201.11227*, 2022.

Eric Rojas, Jorge Munoz-Gama, Marcos Sepúlveda, and Daniel Capurro. Process mining in healthcare: A literature review. *Journal of biomedical informatics*, 61:224–236, 2016.

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

Torsten Scholak, Nathan Schucher, and Dzmitry Bahdanau. Picard: Parsing incrementally for constrained auto-regressive decoding from language models. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 9895–9901, 2021.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce C Ho, Carl Yang, and May Dongmei Wang. Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22315–22339, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

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

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 599–621, 2024.

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

Haoyu Wang, Christopher M. Poskitt, and Jun Sun. AgentSpec: Customizable runtime enforcement for safe and reliable LLM agents. In *Proceedings of the 48th IEEE/ACM International Conference on Software Engineering*, pp. 2938–2950, 2026. doi: 10.1145/3744916.3764546.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Cyril Zakka, Joseph Cho, Gracia Fahed, Rohan Shad, Michael Moor, Robyn Fong, Dhamanpreet Kaur, Vishnu Ravi, Oliver Aalami, Roxana Daneshjou, et al. Almanac copilot: towards autonomous electronic health record navigation. *arXiv preprint arXiv:2405.07896*, 2024.

## A EXPERIMENTAL ARM CONFIGURATIONS

Table 3 lists the shared planning configurations; Table 4 reports the six-arm appendicitis comparison.

Table 3: Experimental arm configurations. Arms are selected through a shared planning code path. The LLM critic receives the same case and planning context as induced repair, but not the induced specification. Align denotes the optional item-alignment operator.

| Arm | Plan samples | Reflection | Authored rules | Intervention |
|---|---|---|---|---|
| BASELINE (ReAct-style) | 1 | no | n/a | n/a |
| VERIFIABLE PLAIN | 1 | no | off | none |
| REFLEXION | 1 | **yes** | off | none |
| SELF CONSISTENCY | **3** | no | off | none |
| LLM CRITIC | 1 | no | off | **LLM critic** |
| CONSTRAINT (ours) | 1 | no | off | **induced spec** |
| CONSTRAINT ALIGNED | 1 | no | off | induced spec + align |
| VERIFIABLE (frozen v1) | 1 | no | **on** | hand-written bundle |

Table 4: Appendicitis ladder. Satisfaction and violation counts are deterministic and computed on 50 trajectories; recall columns are marginal means on metric-specific evaluation-complete subsets. Paired comparisons use the shared applicable cases and can differ from differences between these marginal means. Deltas are paired bootstrap, * marks intervals excluding zero. Diagnosis match is 1.000 for every arm and is omitted. The separate CONSTRAINT ALIGNED arm targets string agreement but can also change process scores (§5.4).

| | induced metric | | | ground-truth recall | | | |
|---|---|---|---|---|---|---|---|
| arm | HSAT | h.viol | prec. | blood | urine | proc. | micro |
| VERIFIABLE PLAIN | 0.824 | 5.46 | **0.964** | 0.382 | 0.403 | 0.817 | 0.148 |
| REFLEXION | 0.809 | 5.90 | 0.929 | 0.400 | 0.286 | 0.801 | 0.161 |
| SELF CONSISTENCY | 0.821 | 5.18 | **0.964** | 0.228 | 0.117 | 0.813 | 0.098 |
| LLM CRITIC | 0.848 | 5.04 | 0.924 | 0.431 | 0.287 | 0.831 | 0.239 |
| CONSTRAINT (ours) | 0.917 | 3.10 | 0.884 | **0.537** | 0.341 | **0.901** | **0.552** |
| CONSTRAINT ALIGNED | **0.928** | **2.74** | 0.888 | 0.535 | **0.442** | 0.852 | 0.536 |

Against plain planning, reflexion, self-consistency and the critic yield paired deltas of $-0.015$, $-0.003$ and $+0.024$, respectively.

## B SPECIFICATION INDUCTION

Table 5 expands §3.3. Requirement promotion and precedence consistency use Wilson lower bounds ((4)); support thresholds use admission counts, and avoidance filtering uses empirical prevalence. Evaluation identifiers are excluded from induction.

Table 5: Specification induction for one cohort. $D' = D \setminus X$ is the effective induction split and $N = |D'|$. $\mathrm{anc}(c)$ contains the strict ancestors of $c$. $K$ denotes the retained data-supported concepts, stored in specification order. Severity is assigned using (5).

| | step | note |
|---|---|---|
| *Input:* candidate induction identifiers $D$, excluded evaluation identifiers $X$, thresholds $\theta$, guideline corpus $G$ | | |
| 1 | Set $D' \leftarrow D \setminus X$ and $N \leftarrow \lvert D' \rvert$; flatten every $h \in D'$ into events $(h, T, c, t)$: day-1 medications, labs, urine and microbiology in the first 24 hours, radiology, procedure ICD titles | every raw string mapped through $\varphi$ |
| 2 | discard events whose concept is not decision-relevant | the flag is data, not control flow |
| 3 | $A^+, A^- \leftarrow$ concepts asserted / under a negation cue in the bullets of $G$ matching this cohort | both $\emptyset$ in the data-only ablation |
| 4 | $n_c, n_T \leftarrow$ number of admissions in which concept $c$ / tool $T$ occurs | |
| 5 | **for** each tool $T$ with $n_T \geq n_{\min}$: | |
| 6 | $\quad s \leftarrow \mathrm{sev}(T)$ at thresholds $(0.60, 0.25)$ | |
| 7 | $\quad$ **if** $s$ is not discarded: emit `require_action`$(T)$ at severity $s$, recording the most prevalent concept of $T$ | record the concept used for tool-level injection |
| 8 | $K \leftarrow \emptyset$ | kept concepts, in specification order |
| 9 | **for** each concept $c \notin A^-$ with $n_c \geq n_{\min}$ and $\mathrm{sev}(c)$ not discarded: | |
| 10 | $\quad$ emit `require_concept`$(c)$ at severity $\mathrm{sev}(c)$, provenance DATA plus GUIDELINE when $c \in A^+$; add $c$ to $K$ in specification order | |
| 11 | **for** each $c \in A^+ \setminus (K \cup A^-)$: emit SOFT `require_concept`$(c)$ | flagged guideline-asserted, unconfirmed by prevalence |
| 12 | **for** each $c \in A^-$ with $\hat{p}_c \leq \tau_{\mathrm{avoid}}$: emit SOFT `avoid_concept`$(c)$ | a negation needs low observed prevalence too |
| 13 | $B \leftarrow \{c \in K : c$ diagnostic, required HARD$\}$; $A \leftarrow \{c \in K : c$ therapeutic, $\mathrm{anc}(c) \cap K = \emptyset\}$ | orderings are stated at the most general kept level |
| 14 | **for** each $(b, a) \in B \times A$ satisfying (6): | |
| 15 | $\quad$ among the candidates sharing $(\mathrm{host}(b), a)$, emit only the one with the largest $\big(\underline{p}(k_{ba}, N_{ba}),\ N_{ba}\big)$ | one ordering per diagnostic tool for each therapeutic concept |
| 16 | Retain up to three most frequent raw payloads for $c$, including the modal exemplar $e(c)$ in (7), for every $c \in K$ and every modal concept recorded above | |
| 17 | **return** the specification with $\theta$, $N$, $\mathrm{hash}(D)$, $\mathrm{hash}(D \setminus X)$, $\mathrm{hash}(X)$, the matched guideline identifiers, the vocabulary statistics and a content fingerprint | every reported number is tied to an exact constraint set |

## C REPAIR PROCEDURE AND GUARANTEES

Table 6 specifies the repair loop; the guarantees below separate presence preservation from precedence behaviour and budgeted termination from a fixpoint.

**Proposition 1** (Preservation of presence requirements)**.** *Adding order items, inserting actions and moving unexecuted actions preserve satisfied tool- and concept-presence requirements. Removing* `Finish` *also preserves these requirements provided that* `Finish` *is not a required tool and carries no required clinical concept.*

*Proof.* Let $C(\tau) = \{c : \sigma_\tau(c) < \infty\}$. The operators ADD_ITEM and INSERT_ACTION do not remove existing calls or argument items, so neither $C(\tau)$ nor $\mathcal{T}_\tau$ decreases. Moving an unexecuted action changes its position but not its tool or concepts. By (2), a satisfied presence requirement therefore remains satisfied. Under the stated condition on `Finish`, deferring it removes no tool or concept needed by a presence requirement. □

**Scope of monotonicity.** Presence preservation does not imply a monotone aggregate score: injection can activate a violated precedence, and movement can affect other orderings. A satisfied avoidance

Table 6: Deterministic repair at one planning step. The procedure modifies the remaining plan and leaves the executed prefix unchanged. The default severity set is $\Sigma = \{\text{HARD}\}$.

| | Procedure |
|---|---|
| | *Input:* remaining plan $P$, executed prefix $\tau_{\text{exec}}$, specification $\mathcal{S}_D$, severity set $\Sigma$, and loop iteration budget $K = 24$. |
| 1 | Initialise $P' \leftarrow P$ and the injected-concept set $J \leftarrow \emptyset$. |
| 2 | Identify required umbrella concepts $U$ subsumed by more specific required concepts. |
| 3 | For at most $K$ iterations, recompute $\text{obs}(\tau_{\text{exec}} \cdot P')$ and collect applicable violations with severity in $\Sigma$. |
| 4 | Stop if no selected violation remains. Otherwise, scan violations in specification order. |
| 5 | For a required concept $c$, or a required tool with modal concept $c$, skip injection if $c \in J \cup U$. Do not use a tool requirement alone to inject a therapeutic concept. |
| 6 | If $c$ is therapeutic and a mined predecessor is absent from the combined trajectory and plan, record the requirement as blocked and continue the scan. |
| 7 | Otherwise, attempt INJECT$(P', c)$. Add $c$ to $J$ only if an edit is emitted. |
| 8 | For a violated applicable precedence, move the planned predecessor action before the successor when this is admissible. If the inversion is fixed by the executed prefix, record it as irreducible. |
| 9 | After an edit, restart from the observation step. Stop if a full scan emits no edit. |
| 10 | If the plan ends with `Finish` while an unblocked requirement remains open, remove `Finish` using DEFER_FINISH. |
| 11 | Return the repaired plan, edit log and residual violations annotated by cause. |
| | INJECT$(P', c)$ |
| 12 | If no exemplar is available, record the missing exemplar and return without editing. |
| 13 | Set $T \leftarrow \text{host}(c)$ and obtain the order payload from the retained exemplars for $c$. |
| 14 | If $T$ is list-valued and a compatible action exists in $P'$, append the payload using ADD_ITEM; cost $0.5$. |
| 15 | Otherwise, insert a new action using INSERT_ACTION; cost $1.0$. Apply the diagnostic and therapeutic placement rules and recheck precedences on the next iteration. |

for $c$ is preserved only if the injected payload contains no $c' \preceq c$; disjoint requirement and avoidance identifiers are insufficient.

**Proposition 2** (Termination of budgeted repair)**.** *For a finite plan and specification, the procedure in Table 6 terminates after at most $K$ loop iterations. It emits at most $K$ edits within the loop and at most one additional* DEFER_FINISH *edit.*

*Proof.* Each iteration scans a finite set of violations and emits at most one edit. The loop has an explicit bound of $K$ iterations, after which the final `Finish` check is executed once. Thus both the number of iterations and emitted edits are finite. □

**Fixpoints and stronger bounds.** A complete scan without an edit establishes a fixpoint of the implemented guards; budget exhaustion does not. Neither guarantees full satisfaction. Residuals distinguish immutable-prefix violations, missing exemplars, temporary blocking and budget exhaustion.

The stronger bound $|\mathcal{S}_D| + I(P)$ requires each injection to discharge a requirement without creating precedence inversions, and each movement to reduce the inversion count without creating another. Injections are then bounded by the requirement count and movements by the initial inversion count; final `Finish` removal is counted separately. Presence preservation alone establishes neither these assumptions nor global minimum edit cost.

## D INDUCED SPECIFICATIONS AND HELD-OUT FIT

On disjoint documented-care admissions, macro hard satisfaction is $0.845$ (Table 7). This empirical reference is not a score ceiling: specification-optimised agents may exceed it without establishing greater clinical quality. Specification sizes reflect both induction support and observed care patterns.

Table 7: Specification size and held-out fit on reconstructed documented-care trajectories. Induction and evaluation splits are disjoint. All sat. includes both hard and soft constraints under the same applicability rules. The final row sums admission counts and reports unweighted cohort means for satisfaction.

| Cohort | Induction $n$ | Held-out $n$ | Total | Hard | HSAT | All sat. |
|---|---|---|---|---|---|---|
| appendicitis | 497 | 265 | 54 | 41 | 0.839 | 0.692 |
| cholecystitis | 226 | 97 | 60 | 38 | 0.851 | 0.756 |
| diverticulitis | 64 | 28 | 50 | 27 | 0.841 | 0.775 |
| pancreatic cancer | 15 | 7 | 24 | 13 | 0.802 | 0.667 |
| pancreatitis | 67 | 29 | 40 | 18 | 0.905 | 0.736 |
| pneumonia | 25 | 11 | 32 | 17 | 0.845 | 0.728 |
| pulmonary embolism | 92 | 40 | 33 | 21 | 0.834 | 0.695 |
| urinary tract inf. | 117 | 50 | 37 | 24 | 0.841 | 0.717 |
| total / macro | 1103 | 527 | — | — | 0.845 | 0.721 |

## E CROSS-COHORT SPECIFICITY

Table 8 and Figure 4a compare all eight cohorts and specifications. Row-wise comparisons change specification stringency; column-wise comparisons fix the specification but retain differences in case mix and applicability. Every specification scores its own cohort above the other-cohort mean, ranking it first in six columns, third for pancreatic cancer and second for pancreatitis.

Table 8: Cross-cohort hard satisfaction. Rows are trajectory cohorts and columns are specifications; diagonal entries are bold. The final rows report each diagonal entry minus the unweighted mean of the seven off-diagonal entries in its column, and its rank within that column. App: appendicitis; Chol: cholecystitis; Div: diverticulitis; PaCa: pancreatic cancer; Panc: pancreatitis; Pneu: pneumonia; PE: pulmonary embolism; UTI: urinary tract infection.

| Trajectory cohort | App | Chol | Div | PaCa | Panc | Pneu | PE | UTI |
|---|---|---|---|---|---|---|---|---|
| appendicitis | **0.839** | 0.764 | 0.802 | 0.789 | 0.788 | 0.764 | 0.688 | 0.788 |
| cholecystitis | 0.835 | **0.851** | 0.793 | 0.872 | 0.906 | 0.831 | 0.696 | 0.802 |
| diverticulitis | 0.784 | 0.740 | **0.841** | 0.777 | 0.775 | 0.798 | 0.685 | 0.828 |
| pancreatic cancer | 0.633 | 0.651 | 0.640 | **0.802** | 0.710 | 0.655 | 0.602 | 0.652 |
| pancreatitis | 0.781 | 0.798 | 0.730 | 0.825 | **0.905** | 0.777 | 0.673 | 0.754 |
| pneumonia | 0.684 | 0.677 | 0.690 | 0.657 | 0.650 | **0.845** | 0.658 | 0.804 |
| pulmonary embolism | 0.599 | 0.561 | 0.585 | 0.723 | 0.675 | 0.635 | **0.834** | 0.645 |
| urinary tract inf. | 0.695 | 0.677 | 0.702 | 0.675 | 0.687 | 0.793 | 0.649 | **0.841** |
| own − mean(other) | +0.124 | +0.156 | +0.135 | +0.042 | +0.163 | +0.094 | +0.170 | +0.088 |
| own rank in column | 1/8 | 1/8 | 1/8 | 3/8 | 2/8 | 1/8 | 1/8 | 1/8 |

**Exploratory cohort identification.** Assigning each trajectory to its highest $z$-normalised specification score gives accuracy $0.391$ on $527$ trajectories and macro accuracy approximately $0.572$. Overall accuracy exceeds uniform guessing ($0.125$) but falls below the majority-class baseline ($265/527 \approx 0.503$). Standardisation adjusts score location and scale, not all differences in specification difficulty.

Per-cohort accuracy and most frequent incorrect prediction are listed below, showing uneven discrimination and overlapping process signatures.

| Cohort | Accuracy | Most frequent error |
|---|---|---|
| Pulmonary embolism | 0.925 | Pancreatic cancer |
| Pneumonia | 0.818 | Urinary tract infection |
| Pancreatic cancer | 0.714 | Pancreatitis |
| Pancreatitis | 0.621 | Pancreatic cancer |
| Urinary tract infection | 0.600 | Pneumonia |
| Diverticulitis | 0.429 | Urinary tract infection |
| Appendicitis | 0.294 | Diverticulitis |
| Cholecystitis | 0.175 | Pancreatic cancer |

## F PULMONARY EMBOLISM: PER-CONSTRAINT ANALYSIS

Table 9 compares plain planning, offline replay of the same trajectories and live CONSTRAINT repair.

Table 9: Per-constraint hard satisfaction for pulmonary embolism on $n = 40$ paired cases. The specification, fingerprint `980075734d5b`, was induced from $n = 92$ admissions and contains 21 hard constraints out of 33 total. Seven constraints satisfied in every case in all three arms are grouped into one row. Aggregate HSAT is computed at the trajectory level using applicable hard constraints.

| Constraint | Plain | Offline replay | Live CONSTRAINT |
|---|---|---|---|
| `require_concept:med:anticoag_heparin_iv` | 0.000 | 1.000 | 1.000 |
| `require_concept:med:anticoag_oral` | 0.000 | 1.000 | 1.000 |
| `require_concept:med:fluid_dextrose` | 0.000 | 1.000 | 1.000 |
| `require_concept:lab:differential` | 0.025 | 1.000 | 1.000 |
| `require_concept:med:analgesia_nonopioid` | 0.150 | 1.000 | 1.000 |
| `require_concept:lab:electrolyte_extended` | 0.200 | 1.000 | 1.000 |
| `require_concept:med:anticoagulation` | 0.225 | 1.000 | 1.000 |
| `require_concept:lab:coagulation` | 0.325 | 1.000 | 1.000 |
| `require_concept:med:analgesia` | 0.600 | 1.000 | 1.000 |
| `require_concept:med:iv_fluid` | 0.600 | 1.000 | 1.000 |
| `require_concept:rad:radiograph` | 0.650 | 1.000 | 1.000 |
| `require_concept:rad:ct` | 0.875 | 1.000 | 1.000 |
| `require_action:MedicationRequestList` | 0.950 | 1.000 | 1.000 |
| `precede:rad:ct→med:iv_fluid` | 0.955 | 0.925 | 1.000 |
| *seven further constraints at 1.000 in all three columns* | | | |
| HSAT (aggregate) | 0.550 | 0.996 | 1.000 |

Both repair settings satisfy all listed presence requirements. The residual CT-before-fluid violations explain the replay–live gap: replay preserves frozen histories, whereas live repair can act before execution. Injection can also activate this precedence in additional cases, changing its applicable-case denominator.

## G REPAIR-COMPONENT ABLATIONS

We ablate repair on 25 saved plain-arm pulmonary embolism trajectories, always scoring against the original specification. Table 2 reports all six variants, case-bootstrap 95% confidence intervals and mean edit counts.

Without exemplars, satisfaction remains at $0.537$ despite $1.00$ mean logged edit, showing that this implementation needs executable payloads. LLM-authored exemplars recover $0.280$ of the full $0.461$ gain; full repair scores $0.181$ higher with fewer edits ($6.72$ versus $7.72$). These results do not establish that payloads must be data-derived.

Removing precedences has no measured replay effect. This does not establish their redundancy online: frozen histories restrict movement, while precedences also affect injection guards and placement.

## H MULTIPLE-COMPARISON CORRECTION

Two-sided paired sign-flip tests enumerate sign configurations for at most 15 non-zero differences; otherwise they use 20,000 Monte Carlo draws with a fixed seed. Of 98 tests, 45 use exact enumeration and 7 have all-zero differences. Holm and Benjamini–Hochberg (BH) corrections are applied separately within the families in §4. Sign-flip validity requires the corresponding null exchangeability or symmetry condition.

Table 10: Significant tests before and after correction within each archived family. Nominal: unadjusted $p < 0.05$; Holm: family-wise adjusted $p < 0.05$; BH: Benjamini–Hochberg adjusted $q < 0.05$. Family sizes sum to 98.

| family | tests | nominal | Holm / BH | content |
|---|---|---|---|---|
| PRIMARY | 8 | 8 | **8 / 8** | HSAT, both contrasts × 4 cohorts |
| INDEPENDENT | 48 | 18 | 10 / 17 | string-matched recalls, diagnosis match |
| COUPLED | 24 | 19 | 16 / 19 | all-sat, violation counts, precedence |
| JUDGED | 18 | 9 | 2 / 7 | medication recall and deterministic bundle coverage |

**Primary process outcome.** All eight contrasts survive Holm correction (unadjusted $p < 10^{-4}$; adjusted $p = 0.0004$), with effects of $+0.069$ to $+0.450$ over 40–50 paired admissions.

**Independent outcomes.** Blood recall improves against both baselines in all four cohorts (eight tests, $p < 10^{-4}$; Holm-adjusted $p = 0.0024$). The other Holm-significant gains are pulmonary embolism diagnosis match against plain ($+0.417$, adjusted $p = 0.0024$) and urinary tract infection microbiology recall ($+0.144$, adjusted $p = 0.019$). Appendicitis microbiology ($+0.327$, $q = 0.034$), cholecystitis microbiology ($+0.283$, $q = 0.016$) and cholecystitis procedure recall ($+0.180$, $q = 0.013$) pass BH only. Six tests have all-zero paired differences, including appendicitis diagnosis and radiology against both baselines; this is observed equality, not population equivalence.

**Secondary medication outcomes.** Only two tests in the archived JUDGED family survive Holm correction: appendicitis medication-bundle declines against plain ($-0.075$) and the critic ($-0.079$), both adjusted $p = 0.0088$. The bundle uses deterministic keywords for antibiotics, analgesics, antiemetics and intravenous fluids, so judge variability cannot explain these declines. We retain the archived family and its correction.

LLM-judged medication-name recall against the critic rises for pulmonary embolism ($+0.373$) and falls for cholecystitis ($-0.227$); both pass BH only. Judge instability is established by repeated evaluation (Appendix O), not by these opposite signs.

## I LLM-AUTHORED SPECIFICATIONS

The same model backbone receives 71 decision-relevant concept identifiers and tool bindings and, at temperature 0.2, returns JSON requirements, severities, tools, precedences and avoidances. Outputs are preserved except for filtering identifiers outside the vocabulary: one invalid tool, `RadiologyRequestList`, is removed from the cholecystitis specification; no invalid concepts are observed.

Table 11: LLM-authored specifications in offline replay of plain trajectories. LLM SPEC replaces constraints while retaining induced modal payloads; LLM SPEC FULL replaces both constraints and payloads. Scoring always uses the mined specification. Concept recall and precision measure agreement with its hard required concepts; precedence recall measures recovery of its ordering constraints. For appendicitis, fewer injections can avoid activating violated precedences (§5.3), so a higher replay score need not imply broader requirement coverage.

| | | HSAT with repair driven by | | | authored vs mined | | |
|---|---|---|---|---|---|---|---|
| cohort | $n$ | mined | LLM SPEC | LLM SPEC FULL | concept rec. | concept prec. | prec. rec. |
| appendicitis | 50 | 0.774 | 0.829 | **0.831** | 0.467 | 0.778 | 0.100 |
| pulmonary embolism | 40 | **0.996** | 0.689 | 0.555 | 0.357 | 0.833 | 0.333 |
| urinary tract inf. | 40 | **0.942** | 0.781 | 0.780 | 0.429 | 1.000 | 0.111 |
| cholecystitis | 40 | **0.799** | 0.746 | 0.746 | 0.444 | 1.000 | 0.040 |

**Coverage and granularity.** Authored specifications contain 6–9 hard concepts and 1–4 precedences, versus 14–18 and 6–25 after induction. All omit white-cell differential, intravenous fluids and non-opioid analgesia; three omit lactate and extended electrolytes, despite prevalence above $0.5$ in the relevant cohorts. The comparison therefore reflects coverage as well as specificity.

The model labels cholecystectomy as soft and requires parent anticoagulation rather than intravenous heparin or oral anticoagulation for pulmonary embolism. Parent-concept actionability depends on exemplars and tool bindings, not abstraction alone.

**Avoidance constraints.** The authored pulmonary embolism specification adds 30 avoidances, including all antibacterial classes and microbiology concepts; induction produces none. Some, including cultures, occur in the cohort. This mismatch with documented practice does not determine either order or prohibition appropriateness.

## J SENSITIVITY TO INDUCTION THRESHOLDS

We vary one threshold at a time on the same induction admissions and plain-arm trajectories, scoring all repaired outputs against the fixed production specification. Cohort abbreviations follow Table 8.

Five sweeps leave hard satisfaction and hard-concept counts unchanged, without establishing identical full specifications or actions. At 92–497 induction admissions, a Wilson lower bound of $0.50$ requires support above 20, making the tested support floors non-binding for hard promotion. Soft-promotion changes and precedence-consistency thresholds below $0.90$ can affect soft constraints without directly changing hard requirements; indirect effects depend on which constraints repair consumes.

Stricter hard promotion ($0.60$–$0.70$) removes requirements and lowers production-specification satisfaction, especially for pulmonary embolism and urinary tract infection. Lowering it to $0.40$ changes only the cholecystitis score ($0.799$ to $0.804$), supporting local stability around the default.

Table 12: One-at-a-time threshold sensitivity under offline replay. Each cell reports HSAT against the fixed production specification, followed by the number of hard required concepts in the re-induced specification. † marks the default. Identical scores and hard-concept counts do not imply identical full specifications.

| Threshold | Value | App | PE | UTI | Chol |
|---|---|---|---|---|---|
| hard promotion | 0.4 | 0.774 (21) | 0.996 (15) | 0.942 (17) | 0.804 (22) |
| | 0.5† | 0.774 (15) | 0.996 (14) | 0.942 (14) | 0.799 (18) |
| | 0.6 | 0.762 (12) | 0.642 (8) | 0.849 (11) | 0.734 (14) |
| | 0.7 | 0.748 (10) | 0.642 (6) | 0.769 (8) | 0.734 (13) |
| soft promotion | 0.1 | 0.774 (15) | 0.996 (14) | 0.942 (14) | 0.799 (18) |
| | 0.2† | 0.774 (15) | 0.996 (14) | 0.942 (14) | 0.799 (18) |
| | 0.3 | 0.774 (15) | 0.996 (14) | 0.942 (14) | 0.799 (18) |
| min. support | 3 | 0.774 (15) | 0.996 (14) | 0.942 (14) | 0.799 (18) |
| | 5† | 0.774 (15) | 0.996 (14) | 0.942 (14) | 0.799 (18) |
| | 10 | 0.774 (15) | 0.996 (14) | 0.942 (14) | 0.799 (18) |
| | 20 | 0.774 (15) | 0.996 (14) | 0.942 (14) | 0.799 (18) |
| precedence consistency | 0.7 | 0.774 (15) | 0.996 (14) | 0.942 (14) | 0.799 (18) |
| | 0.8† | 0.774 (15) | 0.996 (14) | 0.942 (14) | 0.799 (18) |
| | 0.9 | 0.774 (15) | 0.996 (14) | 0.942 (14) | 0.799 (18) |
| min. pair support | 5 | 0.774 (15) | 0.996 (14) | 0.942 (14) | 0.799 (18) |
| | 10† | 0.774 (15) | 0.996 (14) | 0.942 (14) | 0.799 (18) |
| | 20 | 0.774 (15) | 0.996 (14) | 0.942 (14) | 0.799 (18) |
| action promotion | 0.5 | 0.774 (15) | 0.996 (14) | 0.942 (14) | 0.799 (18) |
| | 0.6† | 0.774 (15) | 0.996 (14) | 0.942 (14) | 0.799 (18) |
| | 0.7 | 0.774 (15) | 0.996 (14) | 0.942 (14) | 0.799 (18) |

## K SENSITIVITY TO INDUCTION SAMPLE SIZE

Table 13 evaluates specification recovery and replay against the full-split reference, using three draws per reduced induction size. Minima and maxima describe draw ranges, not confidence intervals.

Larger samples generally improve recovery, but neither recovery nor replay satisfaction is monotone. Tens of admissions recover substantial portions of these specifications; the results do not establish a universal sample-size requirement.

For appendicitis, greater concept recovery does not lift replay satisfaction above plain planning: additional requirements can expose ordering violations in frozen histories (Appendix M). Precedence sample needs depend on joint endpoint prevalence and temporal consistency, not a general quadratic scaling law.

Table 13: Induction sample-size analysis with three draws per reduced size and one run using the full split. Satisfaction is evaluated against the full-split specification; min and max give the range across draws. Rec. and prec. describe agreement with the full-split hard concept set; p.rec. describes precedence recovery. These are specification-recovery measures, not independent clinical order recalls. Edits are mean replay edit counts.

| cohort | ind. $n$ | HSAT | min | max | rec. | prec. | p.rec. | edits |
|---|---|---|---|---|---|---|---|---|
| appendicitis | 5 | 0.762 | 0.752 | 0.780 | 0.511 | 1.000 | 0.000 | 1.33 |
| (plain 0.824) | 10 | 0.761 | 0.752 | 0.777 | 0.578 | 1.000 | 0.000 | 2.11 |
| | 25 | 0.776 | 0.756 | 0.794 | 0.778 | 1.000 | 0.117 | 2.60 |
| | 50 | 0.794 | 0.794 | 0.794 | 0.800 | 1.000 | 0.533 | 2.48 |
| | 100 | 0.758 | 0.756 | 0.762 | 0.844 | 1.000 | 0.833 | 3.35 |
| | 200 | 0.768 | 0.761 | 0.774 | 0.956 | 1.000 | 0.867 | 3.83 |
| | 497 | 0.774 | — | — | 1.000 | 1.000 | 1.000 | 4.08 |
| pulmonary embolism | 5 | 0.704 | 0.684 | 0.733 | 0.357 | 1.000 | 0.000 | 2.05 |
| (plain 0.550) | 10 | 0.640 | 0.637 | 0.642 | 0.429 | 1.000 | 0.000 | 2.55 |
| | 25 | 0.912 | 0.859 | 0.942 | 0.857 | 1.000 | 0.667 | 5.70 |
| | 50 | 0.963 | 0.935 | 0.996 | 0.952 | 1.000 | 0.889 | 6.36 |
| | 92 | 0.996 | — | — | 1.000 | 1.000 | 1.000 | 6.78 |
| urinary tract inf. | 5 | 0.820 | 0.779 | 0.844 | 0.619 | 1.000 | 0.000 | 2.07 |
| (plain 0.756) | 10 | 0.807 | 0.759 | 0.878 | 0.476 | 1.000 | 0.000 | 1.71 |
| | 25 | 0.840 | 0.782 | 0.907 | 0.714 | 1.000 | 0.481 | 2.51 |
| | 50 | 0.892 | 0.882 | 0.898 | 0.905 | 1.000 | 0.667 | 3.42 |
| | 100 | 0.913 | 0.898 | 0.942 | 0.952 | 1.000 | 0.852 | 3.84 |
| | 117 | 0.942 | — | — | 1.000 | 1.000 | 1.000 | 4.42 |
| cholecystitis | 5 | 0.765 | 0.729 | 0.822 | 0.537 | 0.974 | 0.000 | 1.88 |
| (plain 0.737) | 10 | 0.699 | 0.698 | 0.701 | 0.611 | 1.000 | 0.000 | 2.03 |
| | 25 | 0.753 | 0.752 | 0.755 | 0.815 | 0.979 | 0.173 | 3.79 |
| | 50 | 0.782 | 0.755 | 0.801 | 0.889 | 0.941 | 0.400 | 5.60 |
| | 100 | 0.784 | 0.783 | 0.785 | 0.889 | 0.980 | 0.667 | 5.42 |
| | 200 | 0.788 | 0.782 | 0.799 | 0.963 | 0.964 | 0.973 | 6.19 |
| | 226 | 0.799 | — | — | 1.000 | 1.000 | 1.000 | 5.97 |

## L GUIDELINE-PROVENANCE ABLATION

Disabling guidelines leaves hard concept requirements and replay scores unchanged. For appendicitis, pulmonary embolism, urinary tract infection and cholecystitis, scores are $0.774$, $0.996$, $0.942$ and $0.799$ with either setting.

Guidelines annotate 13, 2, 1 and 6 constraints and add $4$, $1$, $0$ and $0$ soft requirements, respectively: pregnancy testing, two cephalosporin classes and ultrasound for appendicitis, and oxygen or airway support for pulmonary embolism. No avoidance is produced across all eight cohorts.

This ablation does not establish invariance of soft, all-constraint or live-loop outcomes.

## M APPENDICITIS FAILURE ANALYSIS

Induced repair wins against the critic on $35/50$ appendicitis admissions, loses on 12 and ties on 3. All losses coincide with lower precedence satisfaction; Table 14 lists their additional hard violations.

In the loss cases, concepts present in CONSTRAINT but absent from LLM CRITIC include: non-opioid analgesia, Gram stain and blood culture ($8$ cases each), and white-cell differential ($7$). Adding diagnostics after executed treatment can activate previously inapplicable, now violated precedences. Thus greater coverage can lower the aggregate score; per-case losses range from $-0.164$ to $-0.005$.

Table 14: Additional open hard violations in CONSTRAINT relative to LLM CRITIC across the 12 appendicitis loss cases. Each grouped precedence row represents four constraints; each gives the case count for every listed endpoint. Counts can overlap across cases.

| constraint | cases |
|---|---|
| `precede:micro:blood_culture` → `med:{analgesia, antibacterial, antiemetic, iv_fluid}` | 8 each |
| `precede:lab:differential` → `med:{analgesia, antibacterial, antiemetic, iv_fluid}` | 7 each |
| `precede:lab:differential`→`proc:appendectomy` | 7 |
| `precede:micro:blood_culture`→`proc:appendectomy` | 5 |
| `require_concept:lab:differential` | 1 |
| `require_concept:med:analgesia_nonopioid` | 1 |
| `require_action:ProcedureRequestFHIR` | 1 |
| `require_concept:proc:appendectomy` | 1 |

## N RUNTIME AND EXECUTION COST

Table 15 separates live runtime and tool calls from offline replay edits and costs.

Runtime means include all numeric simulation times, including failures and timeouts, but exclude subsequent evaluation and historical retries. In contrast, arm-summary CSV `elapsed s` means use successful, evaluation-complete records; their `n` counts scored trajectories, not necessarily timed records. These estimates have different denominators.

Table 15: Runtime and execution cost. Wall-clock means use archived simulation records with numeric elapsed times, including failed or timed-out attempts. Tool-call counts come from live trajectories. Replay edit counts and weighted edit costs use saved plain-arm trajectories and describe a different execution setting.

| | | wall clock (s) | | | replay edits | | tool calls |
|---|---|---|---|---|---|---|---|
| cohort | $n$ | plain | critic | constraint | Count | Cost | Plain → constraint |
| appendicitis | 50 | 186 | 291 | 246 | 4.08 | 3.47 | 14.7 → 20.2 |
| pulmonary embolism | 40 | 596 | 525 | 353 | 6.78 | 4.62 | 29.1 → 16.6 |
| urinary tract inf. | 40 | 317 | 396 | 524 | 4.42 | 3.51 | 16.9 → 25.1 |
| cholecystitis | 40 | 202 | 354 | 419 | 5.97 | 4.71 | 14.0 → 23.5 |

Repair reduces time and tool calls for pulmonary embolism but increases both relative to plain planning in the other cohorts. These measurements include changes in subsequent agent execution and do not isolate the repair routine's computational cost.

## O REPEATED-EVALUATION AUDIT OF THE HOST BENCHMARK

We re-evaluate 10 unchanged conversations three times with the same judge (Table 16).

String-matched metrics remain stable; both LLM-judged medication metrics vary in scores and reference-set sizes.

This audit demonstrates non-reproducibility on fixed inputs but does not quantify its contribution to between-arm contrasts. Judge-dependent outcomes remain secondary and are labelled separately even when confidence intervals exclude zero.

Table 16: Repeated-evaluation spread across $k = 3$ evaluations of $n = 10$ unchanged conversations using `deepseek-v4-flash`. Cases unstable reports the number with varying scores over the applicable-case denominator. GT unstable counts cases with a changing reference-set size.

| Metric | Evaluator | Mean spread | Max spread | Cases unstable | GT unstable |
|---|---|---|---|---|---|
| diagnosis match | string | 0.000 | 0.000 | 0/10 | 0 |
| blood recall | string | 0.000 | 0.000 | 0/10 | 0 |
| urine recall | string | 0.000 | 0.000 | 0/9 | 0 |
| radiology recall | string | 0.000 | 0.000 | 0/10 | 0 |
| procedures recall | string | 0.000 | 0.000 | 0/10 | 0 |
| microbiology recall | string | 0.000 | 0.000 | 0/9 | 0 |
| clinical med bundle | string | 0.000 | 0.000 | 0/10 | 0 |
| medication names | **LLM judge** | **0.116** | 0.714 | 4/10 | 2 |
| medication classes | **LLM judge** | **0.149** | 0.833 | 7/10 | 7 |

## P REPRODUCIBILITY

Content-fingerprinted specification JSON records thresholds, prevalence and Wilson bounds where applicable, support counts, provenance and split-identifier hashes. Evaluation lists use seed 42 and are versioned with the configuration.

**Scope of the accompanying materials.** The manuscript package provides aggregate and paired statistical summaries, source data for all four figures, and the Figure 2 script; other figures are supplied as exports only. The core replication archive contains the revised checker, repair controls, vocabulary, full specification JSON, synthetic tests and version manifest. It excludes patient trajectories, patient-identifier split lists and the host MIRA runtime. Historical statistics require original run versions and authorised record access: revised code is not byte-identical, and aggregates cannot reproduce admission-level resampling or sign-flip tests.

Fixed trajectories, specifications, vocabulary mappings and verifier implementations yield deterministic scores without an LLM judge; new trajectories and judge-dependent outcomes are separate reproducibility targets. Shared-code configurations expose arm differences, and the frozen hand-authored arm supports regression checks, but shared code alone does not establish identical experimental conditions.