FROM SAFETY NOTICES TO COHORT-CONDITIONED MONITORING: EVIDENCE-GROUNDED LLM AGENTS FOR STATISTICAL PROCESS CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Safety notices require hospitals to map natural language onto local exposure cohorts before statistical monitoring can begin. We formulate this step as a probability distribution over executable cohort predicates, linked to monitoring actions by an equal-ARL_0 expected-loss solver. An LLM uses six calls to five read-only tools to acquire cell-suppressed aggregate evidence and update its cohort distribution. Deterministic code handles calibration and action selection, while patient rows, outcomes, and monitoring losses remain outside the model's interface. We evaluate 56 synthetic notices for 14 cohorts held out from semantic-method development, using cohort memberships derived from real MIMIC-IV exposures and two prespecified process-shift templates. The Agent reduces normalized regret by 38.6% relative to the same LLM without tools and by 64.9% relative to a fixed-query non-LLM pipeline with the same six-call budget. Mean detection delay falls to 53.92 patient arrivals from 58.24 and 67.14, respectively. Fifteen of sixteen preregistered conditions were met; the largest target's 27.86% share of net gain against fixed-query evidence exceeded the 25% concentration ceiling, leaving the conjunctive confirmation criterion unmet. These results quantify how model-directed aggregate evidence use improves cohort-conditioned monitoring decisions within a fixed executable catalog.

1 INTRODUCTION

Statistical process control turns a stream of outcomes into a disciplined answer to one question: has this process changed? CUSUM and EWMA accumulate weak evidence over time, multivariate and risk-adjusted variants extend the same logic to correlated outcome vectors and to case mix, and the resulting charts have been used for decades to monitor surgical mortality, infection rates, and operational quality in hospitals (Page, 1954; Roberts, 2000; Lowry et al., 1992; Steiner et al., 2000; Woodall, 2006).

These methods share a structural premise: the population under surveillance is fixed before the chart is designed. The analyst already knows which patients belong to the process, so the open questions are which statistic to accumulate and where to place the control limit.

Safety notices break that premise. A medication recall, a device advisory, or an infection-control bulletin arrives as prose, and nothing in it identifies the affected patients in the hospital's own record. The notice may use a trade name the local formulary does not carry, describe a route or a care setting where the drug itself is never named, or name a co-exposure whose local equivalent is a conjunction of two separately recorded events. Someone must turn that text into an executable cohort before any chart can be designed. Doing so takes clinical interpretation, familiarity with the local record schema, exploratory queries against it, and monitoring expertise. Few people in any single institution hold that combination, and their time does not scale to the rate at which notices arrive.

A language model is an obvious candidate for the semantic half of that work and a poor candidate for the rest. Transformers degrade sharply once a task requires composing several exact numerical steps (Dziri et al., 2023), and apparent mathematical competence is fragile to superficial changes in how a problem is stated (Mirzadeh et al., 2025). A model asked to name a chart family or supply a control limit mixes semantic competence together with numerical optimization in a single answer

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.

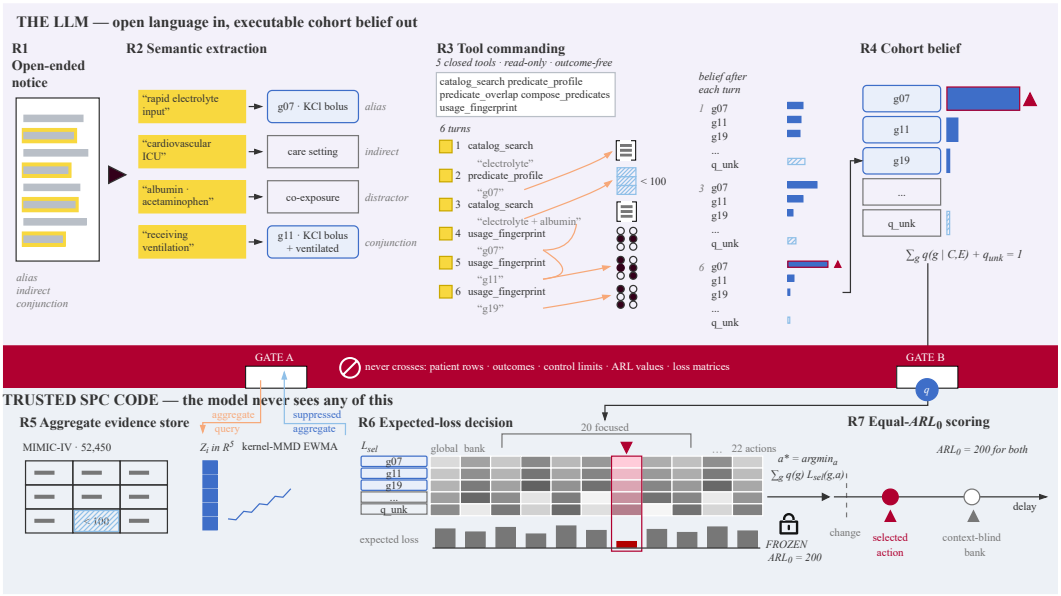


Figure 1: System overview, read top to bottom. Above the firewall the LLM works only in language: it reads the open-ended notice (R1), extracts alias, care-setting, and co-exposure cues (R2), commands six adaptive calls on five closed read-only tools (R3), and returns a distribution over opaque cohort IDs plus unknown mass (R4). It never names a chart or a threshold. Below the firewall trusted code owns every number: the aggregate store answers queries with cells under 100 patients suppressed (R5), and the expected-loss solver routes the belief to one of 22 monitors calibrated offline to a common $ARL_0 = 200$ (R6), whose detection delay defines the score (R7, a schematic of the scoring rule rather than a measured comparison). Gates A and B mark the interface the model never crosses: patient rows, outcomes, control limits, ARL values, and loss matrices stay on the trusted side. The worked example is one notice from the development split.

nobody can audit. The established remedy elsewhere is delegation: the model emits a structured intermediate object and an interpreter computes the quantities (Gao et al., 2023).

We adopt that division of labor as the design principle (Figure 1). The model reads the notice, proposes candidate cohorts from a catalog of executable predicates, and decides which aggregate evidence would most sharpen its own belief. Trusted code owns everything numerical: it exposes only outcome-free, cell-suppressed aggregates, calibrates every candidate monitor to a common in-control average run length, selects the action minimizing expected detection delay under the model's belief, and scores the outcome on an independent Monte Carlo matrix. The model returns a distribution over candidate predicates with explicit unknown mass. This represents uncertainty about the notice's target and allows the solver to account for the monitoring cost of each possible cohort interpretation. Since every action is calibrated to the same false-alarm rate, a semantic error cannot be disguised as speed; it converts directly into detection delay. Judging a cohort hypothesis by the monitoring action it induces makes this a predict-then-optimize problem (Elmachtoub & Grigas, 2022).

Three conditions have to hold for the design to be worth its complexity, and we test them in that order. Routing has to matter, so that the chosen monitor depends on which cohort a notice implicates. The interface has to carry information, so that the aggregate queries the model is permitted to make can separate one cohort from another. And the model has to recover that information under a budget, against retrieval and rule baselines that see the same catalog and issue the same number of queries.

Our contributions are the following.

- **Formulation.** We formulate notice interpretation as a probability distribution over executable exposure cohorts and connect that distribution to monitoring actions through an equal-ARL_0 expected-loss solver.

- **Benchmark.** We construct a MIMIC-grounded benchmark in which real first-24-hour EHR events determine cohort membership, notice text is generated in isolation from the evaluated model, and all patient-level truth stays behind aggregate tools.
- **Evaluation.** We score an evidence-seeking LLM once on held-out notices against keyword, retrieval, ontology, fixed-query, no-tool, exhaustive-evidence, and oracle references, under a pass/fail rule committed before the held-out set was opened.

2 Related Work

Statistical process monitoring in healthcare. CUSUM and EWMA accumulate weak evidence over time and are evaluated through in-control and out-of-control run lengths (Page, 1954; Roberts, 2000), with design, calibration, and run-length properties developed in standard treatments (Montgomery; Hawkins & Olwell, 2012). Multivariate monitoring handles correlated outcome vectors (Lowry et al., 1992), and risk-adjusted CUSUM accounts for case mix (Steiner et al., 2000). Healthcare reviews stress that the observation unit, risk model, outcome, and alarm interpretation must be fixed before deployment (Woodall, 2006; Benneyan et al., 2003), and modern frameworks make broad detector catalogs and simulation-based calibration accessible (Zago, 2025). We take the catalog and its calibration as given and study the decision that precedes them.

Natural-language cohort definition and EHR agents. Criteria2Query translates eligibility criteria into executable clinical database queries (Yuan et al., 2019), and LLM versions broaden the cohort definitions covered (Park et al., 2024). EHR agents answer questions over tabular records using code or tools (Shi et al., 2024), and recent benchmarks study multi-turn tool interaction with clinical databases (Lee et al., 2026). Our Agent never executes unrestricted code or observes patient rows, calls only five closed aggregate tools, and is graded by a hidden equal-ARL_0 loss matrix, never by query syntax or an LLM judge. With an opaque roster and small cells hidden, its only lever is which aggregate comparison to request.

Safety signals and budgeted evidence. Quantitative pharmacovigilance finds unknown associations in spontaneous-reporting databases through proportional reporting ratios (Evans et al., 2001) and empirical-Bayes shrinkage over drug-event frequency tables (DuMouchel, 1999; Bate & Evans, 2009), asking which pairs deserve attention system-wide. We begin after a notice exists and ask which executable cohort inside one hospital's records it denotes. Pricing evidence under a budget is classical: value-of-information values an observation by the downstream improvement it induces (Howard, 1966), and active learning asks which items to query when supervision is costly (Settles, 2009). Our queries return cell-suppressed aggregates in place of labels, and earn their cost only through the monitoring action they change.

Tool use and decision-aware evaluation. ReAct interleaves reasoning and environment action (Yao et al., 2023), Toolformer studies learned tool invocation (Schick et al., 2023), and retrieval-augmented generation injects external evidence into inference (Lewis et al., 2020); BM25, dense embeddings, and rank fusion remain strong non-generative baselines (Robertson & Zaragoza, 2009; Reimers & Gurevych, 2019; Cormack et al., 2009), with terminology expansion available through UMLS (Bodenreider, 2004). These are scored on answer correctness. We follow decision-focused learning instead, judging a predictor by the decision it induces (Elmachtoub & Grigas, 2022; Donti et al., 2017; Wilder et al., 2019): two distributions with similar top-1 accuracy can induce different actions, and a confident error can cost more than a diffuse one.

3 Method

3.1 Problem formulation

Patient-indexed monitoring process. Each observation is one patient's first ICU encounter. After a fixed 30-day maturation horizon, it contributes a five-dimensional adverse-outcome vector Y_i. Phase-I logistic models estimate outcome-specific risks from admission covariates. We monitor centered and scaled Pearson residuals $Z_i \in \mathbb{R}^5$ and resample complete rows to preserve their empirical dependence.

The alarm concerns deterioration in the ICU or hospital process, not deterioration of an individual patient. Run length is counted in matured patient encounters. Since MIMIC shifts dates independently across patients, we use a reproducible pseudo-order and make no claim about calendar-time drift.

Notices and executable cohort hypotheses. Let $\mathcal{G} = \{g_1, \ldots, g_K\}$ be a frozen catalog of executable exposure predicates. An atomic predicate identifies an exposure such as a medication administration or ICU procedure. A compositional predicate is a conjunction of two cross-source atomic exposures during the first 24 hours.

A notice C is generated from one hidden target $g^\star$, but it contains neither the opaque predicate identifier nor local counts. Given C, public catalog metadata, and aggregate tool results E, a semantic method outputs a distribution $q(g \mid C, E)$ over $g \in \mathcal{G}$, plus unknown mass q_{unk}, normalized so that $\sum_{g\in\mathcal{G}} q(g \mid C, E) + q_{\mathrm{unk}} = 1$.

The output schema includes explicit unknown mass, allowing the model to represent unresolved cohort identity before the solver distributes that mass uniformly over the candidate catalog.

Equal-ARL monitoring actions. Let $\mathcal{A}$ denote the frozen action set: a global monitor, a context-blind bank, and one focused monitor per candidate cohort. The global action updates on every arrival. A focused action for g_j updates only when the arriving encounter satisfies g_j; nonmembers leave its statistic unchanged, although run length is still counted in all arrivals. The blind bank runs all K focused monitors in parallel and signals on their maximum normalized statistic, with a joint multiplier for the multiple-monitor false-alarm opportunity.

Every complete action includes its monitored cohort, detector configuration, outcome subset, and threshold. Each is calibrated independently on the full mixed stream so that $\mathbb{E}_0[T_a] = A_0$ with $A_0 = 200$, where T_a is the first alarm time under the in-control process.

This target is imposed separately on the global action, each focused action, and the entire blind bank, so no method can appear faster merely by raising more false alarms. At equal ARL_0, a smaller ARL_1 means faster detection, and 200 is a benchmark design choice, not a clinical operating point. A monitor routed to g_j updates on all arrivals in g_j. When the affected cohort is g_k, only arrivals in the intersection $g_j \cap g_k$ carry the imposed shift.

For target g, action a, and future-shift template d, let $\mathrm{ARL}_1(g, a, d)$ be the mean delay after the process changes. The selection loss averages the two frozen templates:

$$L_{\mathrm{sel}}(g, a) = \frac{1}{2} \sum_{d\in\{\mathrm{location},\mathrm{scale}\}} \mathrm{ARL}_1(g, a, d).$$

Before action selection, the solver forms the effective distribution $\tilde{q}$ over every candidate g and then selects the action

$$\tilde{q}(g) = q(g) + \frac{q_{\mathrm{unk}}}{K}, \qquad a^\star(q) = \arg\min_{a\in\mathcal{A}} \sum_{g\in\mathcal{G}} \tilde{q}(g) L_{\mathrm{sel}}(g, a).$$

A disjoint Monte Carlo matrix L_{ver} scores the selected action. This separates policy choice from evaluation and prevents the LLM from observing any ARL value.

Decision regret. For true target g, let $a_o(g)$ be the best verification action and a_b the context-blind bank. We define normalized selection regret

$$r(g, a) = \frac{L_{\mathrm{ver}}(g, a) - L_{\mathrm{ver}}(g, a_o(g))}{L_{\mathrm{ver}}(g, a_b) - L_{\mathrm{ver}}(g, a_o(g))}.$$

Thus $r = 0$ matches the target-specific oracle and $r = 1$ matches the blind-bank reference. Regret can exceed one when a wrong focused cohort performs worse than the bank.

3.2 EVIDENCE-GROUNDED COHORT AGENT

Budgeted semantic decision. Let $H_0 = (C, \mathcal{G}, R)$ contain the notice, the sanitized candidate catalog, and a frozen shortlist R of the six highest-scoring catalog entries under the BM25 configuration fixed on development. The shortlist orders the catalog for the model; the distribution it returns may place mass on any of the K candidates. Write π for the query policy, $\mathcal{T}$ for the tool interface, and f for the readout. At turn t, the Agent selects one valid query $x_t = \pi(H_{t-1})$ from the tool language, receives cell-suppressed evidence $e_t = \mathcal{T}(x_t)$, and appends it to the ordered transcript $H_t = (H_{t-1}, x_t, e_t)$. After exactly $B = 6$ calls, it returns $q_B = f(H_B)$, and the deterministic solver returns $a^\star(q_B)$. The evaluated composition is

$$H_0 \xrightarrow{\pi,\mathcal{T}} H_B \xrightarrow{f} q_B \longrightarrow a^\star(q_B) \longrightarrow L_{\mathrm{ver}}(g, a^\star(q_B)).$$

This is an evaluation objective, not a training signal: neither π nor f observes target truth, patient rows, or either loss matrix. Numerical optimization begins only after the transcript and probability vector are frozen.

Responsibility boundary. The Agent never receives patient rows, membership vectors, target truth, outcome labels, shift parameters, control limits, or the selection and verification matrices, and it returns only a probability vector over opaque candidate IDs plus unknown mass.

The numerical layer owns all statistical operations. It validates the probability schema, applies the same expected-loss solver to every method, substitutes a predeclared blind action for invalid output, and records the selected action. No natural-language rationale enters the score.

Aggregate evidence tools. The closed tool set is read-only and supports five operations:

- `catalog_search`: search sanitized concepts, categories, and routes;
- `predicate_profile`: return predicate kind, sources, route families, and a coarse prevalence band;
- `predicate_overlap`: return cell-suppressed margins, intersections, and Jaccard overlap;
- `compose_predicates`: test whether two atomic concepts form a candidate conjunction; and
- `usage_fingerprint`: return coarse first-24-hour timing, care-setting enrichment, and up to five sanitized co-exposure clues.

Cells below 100 patients are suppressed. Unknown identifiers, invalid compositions, and malformed arguments fail closed. Tool calls cannot access outcomes or monitoring loss.

Active evidence acquisition. The Agent can adapt each query to preceding results. An example development trace first searches for a notice concept, profiles plausible atomic candidates, tests a compositional hypothesis, and uses timing or co-exposure evidence to resolve remaining ambiguity. The fixed-query baseline instead uses the notice to retrieve six candidates and then requests the same fingerprint for each, so its query identities depend on text retrieval while its plan stays fixed regardless of what the evidence returns.

The same-model no-tool ablation shares the candidate roster, retrieval shortlist, base model, output schema, and solver, and differs only in receiving no aggregate evidence.

Deterministic action solve. The candidate policy catalog contains 105 configurations from ten families spanning Shewhart, Hotelling T^2, univariate and multivariate CUSUM, EWMA and MEWMA, a multiscale GLR scan, and kernel-MMD EWMA. Numerical selection produced 22 fixed actions: one global action, one blind bank, and 20 focused actions. All selected actions use kernel-MMD EWMA; 21 use all five outcomes and one uses mortality plus prolonged ICU stay. The evaluated semantic choice is principally which cohort to monitor, followed by cohort-specific threshold calibration.

3.3 MIMIC-GROUNDED BENCHMARK

Patient cohort and outcomes. We use MIMIC-IV (Johnson et al., 2023). The source cohort contains 65,366 patients, each represented by the first ICU encounter. Benchmark construction and evaluation use the 52,450-patient Phase-I partition. A separate 12,916-patient partition remains reserved for subsequent studies.

Five binary outcomes mature within 30 days: 30-day mortality, ICU stay of at least 7 days, same-admission ICU readmission, invasive ventilation, and renal replacement therapy (Table 3 in Appendix A).

Five-fold Phase-I logistic models use age, gender, first care unit, admission type, insurance, and race. The SPC input is the joint vector of centered and scaled out-of-fold Pearson residuals.

Real exposure predicates. Candidate truth comes from first-24-hour exposure events in four MIMIC sources: prescription starts, administered eMAR records, ICU inputs, and ICU procedures. The Phase-I extraction contains 4.41 million qualifying event records. After minimum membership, semantic renderability, and duplicate controls, it retains 139 atomic predicates and 717 cross-source conjunctions.

Performance-blind rules select 20 semantically renderable targets: eight atomic cohorts and twelve conjunctions. Selection uses source type, prevalence anchors, generic eligibility rules, and opaque-ID tie-breaking, never outcomes or ARL. A salted target-level split assigns six targets to development and 14 to the hidden set.

Controlled notice generation. Each notice verbalizes selected facts from the target cohort's sanitized aggregate profile in one of four language styles. This construction links the language task to executable EHR predicates while controlling the evidence available through the tools. An isolated generator receives a sanitized description of one target and a subset of its true aggregate fingerprint facts. It does not receive target IDs, counts, prevalence, outcomes, DGPs, ARLs, or policy answers. It produces four styles: direct concept language; alias or lexical mismatch; indirect operational clues; and a compositional notice with a distractor.

The resulting benchmark contains 80 notices: 24 development notices and 56 held-out notices. Deterministic validators reject local identifiers, exact aggregate values, answer keys, and monitoring terminology. Development notices receive a manual semantic audit; the other 56 notices and their 14 target predicates cannot be used to modify the evaluated model, prompts, baselines, or tools. All 20 predicate descriptions and precomputed actions nevertheless remain available as catalog candidates, so the split tests semantic generalization across target notices and excludes open-world discovery. Notice generation and evaluation use distinct endpoint-reported model identifiers, `claude-haiku-4-5-20251001` and `gpt-5.5`, respectively. Appendix I documents the relay interface and identifier checks.

Future shifts and action calibration. Each hidden target faces the same two equally weighted process changes: a persistent location shift of 0.75 on mortality and prolonged ICU stay, and a persistent scale multiplier of 1.8 on ventilation and renal replacement therapy. Only patients satisfying the hidden exposure predicate are perturbed. This isolates cohort identification: a semantic method can neither invent its own DGP nor receive a target-specific shift advantage.

4 EXPERIMENTS

4.1 SETUP

Baselines and references. We compare no-context uniform probabilities, keyword and dictionary matching, BM25, MiniLM sentence-embedding retrieval, retrieval fusion, ontology expansion, a fixed six-query evidence pipeline, the evaluated LLM without tools, and the same LLM with six adaptive aggregate queries. Among the tested non-LLM baselines the fixed-query pipeline is strongest on the hidden set; exhaustive-evidence access and true-target routing are reference points, excluded from the equal-budget comparison. Development hyperparameters are frozen before the 14 hidden targets are scored.

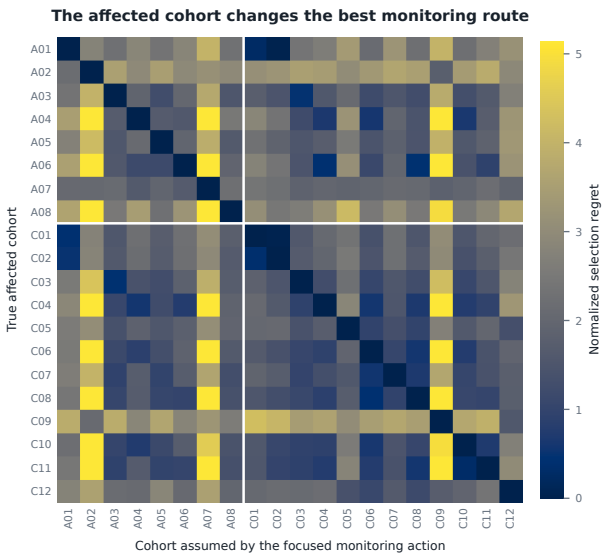


Figure 2: Cross-target verification loss for 20 frozen cohorts. Wrong-cohort monitoring can increase delay when every action targets the same system ARL_0.

Metrics and protocol. The primary endpoint is target-cluster mean normalized regret: all four notice styles for one target remain together. Target-cluster bootstrap stability intervals use 20,000 repetitions and are conditional on the frozen benchmark and Monte Carlo verification matrix. Secondary endpoints include two-template mean ARL_1, top-1 and top-3 cohort accuracy, probability assigned to the true target, Brier score, log score, schema validity, tool errors, exact call count, tokens, and latency. We report atomic and conjunction targets and all notice styles separately.

4.2 Routing depends on which cohort is implicated

The first condition is that the action must depend on cohort identity. Across all twenty catalog targets, target-specific routing achieves mean ARL_1 of 47.66, compared with 77.41 for the strongest context-blind action, a 38.4% reduction (95% target-cluster bootstrap interval: 35.7% to 40.8%). All 20 targets select their own focused action. For a typical target, choosing the median wrong focused cohort incurs 116.7% relative regret (Figure 2).

Independent audits place the achieved ARL_0 of all 22 actions between 193.56 and 207.68, within 3.84% of the target of 200; maximum censoring is 0.06%. The held-out comparison uses these frozen calibrated actions and an independently simulated verification-loss matrix.

4.3 The aggregate interface carries recoverable information

The second condition is that the permitted queries can separate cohorts. On the six development targets, the strongest text-only retrieval configuration has mean normalized regret 1.2167. Giving the same retrieval family exhaustive aggregate fingerprints reduces regret to 0.4097, a target-cluster bootstrap reduction of 66.3% (35.9% to 95.1%). Five of six targets and all four notice styles improve. This measures the headroom available to one frozen non-LLM scoring rule, and is neither an information-theoretic ceiling nor an Agent result.

4.4 One-shot held-out comparison

The third condition is that a budgeted model recovers that information in competition. Frozen methods are evaluated once on 56 notices for 14 cohorts held out from notice and semantic-method development. The held-out split is defined over target-specific notice evaluation and semantic-method development: six targets support development, and fourteen support the final comparison. All twenty executable predicates and their precomputed monitoring actions remain in the fixed candidate catalog. Table 1 reports the principal methods. The Agent is its evidence-grounded LLM row.

The Agent reduces normalized regret by 38.6% (95% target-cluster bootstrap interval: 15.1% to 59.8%) relative to the no-tool LLM and by 64.9% (41.3% to 81.6%) relative to fixed-query evidence.

Table 1: One-shot results for 14 cohorts held out from semantic-method development. All deployable methods emit the same probability schema and use the same deterministic action solver. Regret is the target-cluster mean; $q_\star$ is the probability placed on the true cohort. The exhaustive-evidence reference applies the frozen non-LLM evidence-scoring rule to fingerprints from every catalog entry; it and the oracle are references, not equal-budget competitors.

Method	Top-1	Top-3	$q_\star$	Brier	Log	Regret	Mean ARL_1	Calls
True-target oracle	100.0%	100.0%	1.000	0.000	0.000	0.000	48.57	-
Exhaustive-evidence reference	73.2%	91.1%	0.449	0.401	1.010	0.228	54.21	all
Evidence-grounded LLM	**83.9%**	85.7%	**0.780**	**0.247**	1.357	**0.206**	**53.92**	6
LLM without tools	80.4%	80.4%	0.701	0.260	**0.834**	0.336	58.24	0
Fixed-query evidence	44.6%	64.3%	0.307	0.742	2.130	0.588	67.14	6
Ontology expansion	41.1%	58.9%	0.276	0.750	2.282	0.826	76.10	0
MiniLM embedding	44.6%	51.8%	0.257	0.748	2.376	0.855	76.46	0
Keyword dictionary	35.7%	50.0%	0.199	0.776	2.341	0.846	73.72	0
Retrieval fusion	33.9%	55.4%	0.246	0.828	2.510	0.869	76.56	0
BM25	26.8%	46.4%	0.201	0.907	2.804	0.932	78.54	0
No-context uniform	7.1%	14.3%	0.050	0.950	2.996	1.000	78.33	0

Figure 3 shows mean normalized regret and target-cluster bootstrap intervals for each method. Mean detection delay is 53.92 patient arrivals, compared with 58.24 and 67.14 for the two baselines, corresponding to reductions of 4.32 and 13.22 arrivals. All 112 final probability outputs satisfy the schema. One Agent output required the preregistered format-only repair of an opaque candidate ID; its six-query evidence-acquisition trajectory was retained unchanged.

Aggregate evidence lowers unknown mass from 0.144 to 0.020, raises mean probability on the true cohort from 0.701 to 0.780, and lowers the Brier score from 0.260 to 0.247. The log score increases from 0.834 to 1.357, indicating larger penalties for predictions that assign very low probability to the true target. The six-call procedure averages 24,645 input tokens, 1,734 output tokens, and 55.0 seconds per notice; the no-tool procedure averages 1,493 input tokens, 920 output tokens, and 22.3 seconds.

4.5 Outcome of the preregistered rule

Before the held-out notices were opened, we wrote down the rule by which we would judge the comparison and committed it to the repository. The rule was all-or-nothing: sixteen conditions, every one of which had to hold for us to report the comparison as confirmed. They fall into four groups.

Two effect-size floors required the Agent to reduce regret by at least 15% against the best eligible non-LLM method and by at least 10% against the same model without tools, and two further conditions required both bootstrap lower bounds to sit above zero. Three breadth conditions tested how widely the gain was spread: it had to be positive for atomic and for conjunction targets separately, and no single target was allowed to supply more than 25% of the total net gain. One condition required 98% schema validity. The remaining conditions fixed the scope and the mechanics of the run: the exact hidden set and its 112 model states, the ARL_0 tolerance for every selected action, and five integrity conditions requiring six tool calls per trace with no backend error, model mismatch, repeated query, or unresolved transport failure.

Fifteen conditions were met. The largest target contributed 1.4895 of the total net paired gain of 5.3465, or 27.86%, exceeding the registered 25% ceiling; one target had a negative gain. The conjunctive confirmation criterion was therefore unmet. The benchmark, target set, and thresholds remained fixed throughout the held-out evaluation. Appendix D plots the target-level gains.

4.6 Where the gain comes from

The comparison separates two levels of the workflow. On regret, the no-tool LLM outperforms the tested lexical, dense-retrieval, and ontology baselines, and the Agent outperforms the no-tool model. The no-tool comparison evaluates the addition of aggregate evidence acquisition and interpretation

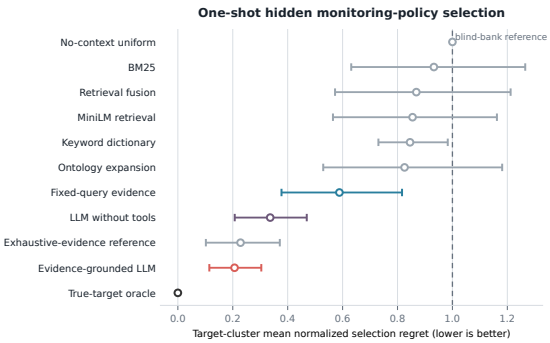


Figure 3: Mean normalized selection regret on the 14-target held-out benchmark. Horizontal bars show 95% target-cluster bootstrap intervals for each method's mean regret.

Table 2: Target-cluster mean regret by notice style and target kind. Agent regret is near the oracle floor on three controlled styles; indirect operational language carries the remaining loss.

Stratum	Agent	No tools	Fixed query
Direct concept	0.003	0.071	0.124
Alias mismatch	0.003	0.074	0.613
Compositional distractor	0.003	0.146	0.643
Indirect operational	0.818	1.054	0.974
Atomic cohorts	0.277	0.514	0.815
Conjunction cohorts	0.154	0.203	0.419

to the same LLM. The fixed-query comparison evaluates the complete model-directed procedure against a retrieval-and-scoring pipeline under an equal call budget.

The end-to-end endpoint behaves differently from top-1 accuracy: the Agent raises top-1 by only 3.6 percentage points over the no-tool model, yet cuts decision regret by 38.6%. Because the solver minimizes expected loss over the whole distribution, probability allocation and cohort-confusion costs drive that difference.

Table 2 shows where the effect lives. Gains hold for both target kinds. Agent regret is 0.003 for direct, alias-mismatch, and compositional-distractor notices, and 0.818 for indirect operational notices. The largest style-level improvements over fixed-query evidence occur for alias mismatch and compositional distractors; indirect operational language accounts for nearly all remaining Agent regret.

5 Limitations

The benchmark combines real first-24-hour MIMIC-IV exposures with synthetic notices generated from selected sanitized aggregate facts. Monitoring is evaluated on a reproducible patient-indexed pseudo-stream under two persistent shift templates. The semantic task uses a fixed 20-predicate catalog containing each target, with six targets for development and fourteen for held-out evaluation. The reported tool comparisons evaluate complete evidence-acquisition procedures. All 22 selected monitoring actions use kernel-MMD EWMA, with action differences determined by cohort, threshold, and outcome subset. The reported conclusions concern cohort-conditioned action selection within this benchmark design.

6 Conclusion

We connect natural-language notice interpretation to statistical monitoring through a distribution over executable cohorts and an equal-ARL_0 expected-loss solver. On the 14-target held-out benchmark, six aggregate queries reduce normalized regret by 38.6% against the no-tool LLM and by 64.9% against the fixed-query pipeline, with corresponding reductions in mean detection delay. Fifteen preregistered conditions were met, while the largest target's gain share exceeded the concentration ceiling and left the conjunctive criterion unmet. Style-specific results locate the largest gains in alias and compositional interpretation, with indirect operational language carrying nearly all residual regret. The study provides an executable interface through which model-directed evidence acquisition informs monitoring decisions under a fixed statistical calibration protocol.

AI USE STATEMENT

Two language models are components of the study itself and are described in the main text: Claude Haiku 4.5 generated the synthetic notice texts (Section 3.3 and Appendix C), and GPT-5.5 is the evaluated Agent. Separately, we used LLM coding assistants for analysis code and for polishing the paper. Every reported number traces to a committed result file. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Deployment would be offline by construction: a hospital defines its predicate catalog and outcome process, calibrates candidate actions, exposes approved aggregate tools, and freezes a monitoring action before prospective use. No online prompt editing, retraining, or threshold adjustment is evaluated. Any resulting alarm is system-level and is not a diagnosis or treatment recommendation for an individual patient.

MIMIC is a credentialed, deidentified research dataset (Johnson et al., 2023). Patient-level rows, local identifiers, and membership vectors never enter model prompts or committed artifacts. Our cell suppression is deterministic and outcome-free and carries no formal guarantee; differential privacy would bound disclosure explicitly, at the cost of noise our rule does not add (Dwork & Roth, 2014). The system is a research prototype and must not be used for clinical decisions without local governance, validation, and human review.

REPRODUCIBILITY STATEMENT

Code and derived benchmark data are included in the supplementary material, with instructions in its README. Patient-level MIMIC-IV data are not redistributed.

REFERENCES

Andrew Bate and Stephen J. W. Evans. Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiology and Drug Safety*, 18(6):427–436, 2009.

James C. Benneyan, Robert C. Lloyd, and Paul E. Plsek. Statistical process control as a tool for research and healthcare improvement. *Quality & Safety in Health Care*, 12(6):458–464, 2003. doi: 10.1136/qhc.12.6.458.

Olivier Bodenreider. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270, 2004. doi: 10.1093/nar/gkh061.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 758–759, 2009. doi: 10.1145/1571941.1572114.

Priya L. Donti, Brandon Amos, and J. Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5484–5494, 2017.

William DuMouchel. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *The American Statistician*, 53(3):177–190, 1999.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems*, volume 36, pp. 70293–70332, 2023.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68 (1):9–26, 2022. doi: 10.1287/mnsc.2020.3922.

Stephen JW Evans, Patrick C Waller, and S Davis. Use of proportional reporting ratios (prrs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and drug safety*, 10(6):483–486, 2001.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: Program-aided language models. In *International Conference on Machine Learning*, pp. 10764–10799, 2023.

Douglas M Hawkins and David H Olwell. *Cumulative sum charts and charting for quality improvement*. Springer Science & Business Media, 2012.

Ronald A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966.

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.

Gyubok Lee, Woosog Chay, Heeyoung Kwak, Yeong 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. In *International Conference on Learning Representations*, volume 2026, pp. 123138–123198, 2026.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Cynthia A Lowry, William H Woodall, Charles W Champ, and Steven E Rigdon. A multivariate exponentially weighted moving average control chart. *Technometrics*, 34(1):46–53, 1992.

Iman Mirzadeh, Keivan Alizadeh-Vahid, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 94743–94765, 2025.

Douglas C Montgomery. *Introduction to statistical quality control*, volume 5.

Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

Jimyung Park, Yilu Fang, Casey Ta, Gongbo Zhang, Betina Idnay, Fangyi Chen, David Feng, Rebecca Shyu, Emily R Gordon, Matthew Spotnitz, et al. Criteria2query 3.0: Leveraging generative large language models for clinical trial eligibility query generation. *Journal of biomedical informatics*, 154:104649, 2024.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

Stuart W Roberts. Control chart tests based on geometric moving averages. *Technometrics*, 42(1): 97–101, 2000.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and trends® in information retrieval*, 4(1-2):1–174, 2009.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68539–68551, 2023.

Burr Settles. Active learning literature survey. 2009.

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. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.1245.

Stefan H Steiner, Richard J Cook, Vern T Farewell, and Tom Treasure. Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics*, 1(4):441–452, 2000.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1658–1665, 2019.

William H Woodall. The use of control charts in health-care and public-health surveillance. *Journal of Quality Technology*, 38(2):89–104, 2006.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Chi Yuan, Patrick B Ryan, Casey Ta, Yixuan Guo, Ziran Li, Jill Hardin, Rupa Makadia, Peng Jin, Ning Shang, Tian Kang, et al. Criteria2query: a natural language interface to clinical databases for cohort definition. *Journal of the American Medical Informatics Association*, 26(4):294–305, 2019.

Daniele Zago. StatisticalProcessMonitoring.jl: A general framework for statistical process monitoring in Julia. *Journal of Statistical Software*, 113(7):1–45, 2025. doi: 10.18637/jss.v113.i07.

A Outcome and Evidence Contracts

Table 3: Five 30-day outcomes in the Phase-I cohort.

Outcome	Phase-I events	Prevalence
30-day mortality	7,195	13.72%
ICU stay at least 7 days	5,948	11.34%
Same-admission ICU readmission	4,423	8.43%
Invasive ventilation	23,071	43.99%
Renal replacement therapy	1,872	3.57%

All outcomes are closed at 30 days after the index ICU admission, even when they are observable earlier. Exposure predicates use only the first 24 hours. The four event sources contribute the following Phase-I records:

Table 4: Phase-I exposure records and retained atomic predicates.

Source	Qualifying event records	Patient-concept memberships	Retained atomic predicates
Prescription start	1,845,050	1,359,922	72
eMAR administered	431,135	264,950	17
ICU input	1,858,734	520,431	37
ICU procedure	277,466	237,344	13

The risk models are deliberately simple and are not proposed as clinical predictors. Their five-fold Phase-I out-of-fold diagnostics are:

Table 5: Phase-I out-of-fold risk-model diagnostics.

Outcome	AUROC	Brier	Calibration intercept	Calibration slope	E/O
30-day mortality	0.751	0.107	-0.016	0.990	1.000
ICU stay at least 7 days	0.646	0.097	-0.028	0.986	1.001
Same-admission ICU readmission	0.587	0.077	-0.287	0.877	1.000
Invasive ventilation	0.761	0.191	-0.002	0.994	1.000
Renal replacement therapy	0.698	0.034	-0.094	0.969	1.001

The maximum absolute off-diagonal covariance among the standardized outcome scores is 0.323. Complete-row resampling preserves this empirical dependence, and simulation-based calibration absorbs the Phase-I score distribution without assuming independent standard normals.

The aggregate evidence store reconstructs all 20 target memberships from constituent events. Every target has at least three unsuppressed care settings and five sanitized co-exposure clues. Two unchanged evidence builds match catalog, fingerprint, and membership-content hashes.

B Numerical Calibration and Detector Catalog

The candidate catalog contains 105 parameterized policies from ten families. For each target, ordinary calibration uses 3,000 streams, policy selection uses 1,200 streams, cross-action verification uses 4,000 streams, and the final action audit uses 10,000 in-control streams. Selection, verification, and audit seeds are disjoint.

The complete 20-by-22 cross-action verification matrix contains global, blind-bank, and 20 target-focused actions. The cross-target matrix in Figure 2 normalizes each row by its target-specific oracle and blind-bank losses. The median wrong-focused regret of 116.7% confirms that the matrix is not a cosmetic relabeling of one equivalent action.

C Notice Generation and Model Isolation

The notice generator is pinned to `claude-haiku-4-5-20251001` and has all tools disabled. It receives a target's sanitized concept description and selected aggregate facts, but no opaque ID, prevalence, count, outcome, perturbation, ARL, or policy answer. Four styles are requested in one structured generation call.

Generation produced 80 unique notices in 32 API calls. Six targets required a deterministic retry, and three development targets used one preregistered same-prompt regeneration after manual semantic audit. Hidden notices were never used to edit prompts or methods.

The evaluated model is a non-Claude `gpt-5.5` endpoint. The Agent and no-tool ablation share the base model, candidate roster, retrieval shortlist, structured probability schema, and deterministic solver. The Agent also performs six sequential tool-using turns. One held-out Agent output mistyped an opaque candidate ID after completing its six calls and used the preregistered format-only repair. That repair response corrected the invalid ID, but no tool query or semantic evidence-acquisition trajectory was repeated.

D Strata, Query Budget, and Concentration

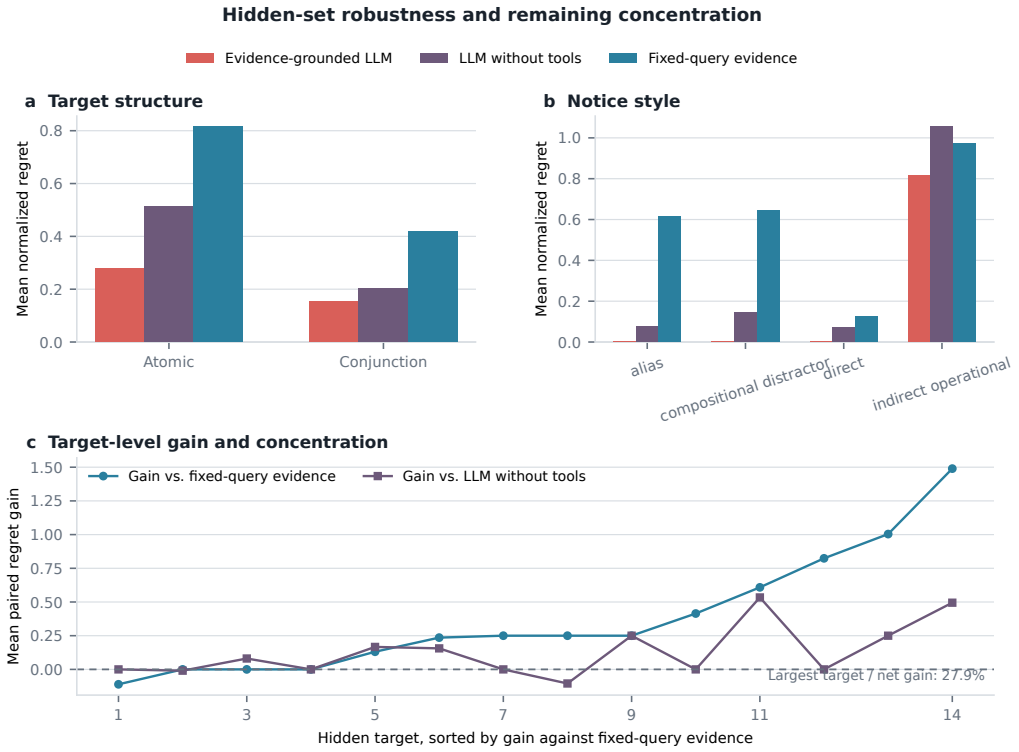


Figure 4: Hidden performance by target structure and notice style, followed by target-level paired gains.

The four hidden-style mean regrets are:

The development-only query-budget curve used 0, 2, 4, 6, and 10 calls. Mean regret was 0.4255, 0.3342, 0.2607, 0.2754, and 0.2735, respectively. The six-call budget was frozen before hidden evaluation and was retained even though four calls were numerically lowest on development.

As reported in the main Results, the preregistered robustness condition divided the largest positive target-level paired gain by the total net target-level paired gain and required that share to be at most 25%. The observed share was 27.86% (1.4895 of 5.3465), and one of the 14 targets has a negative gain. The conjunctive gate was not passed, despite all effect-size, bootstrap, integrity, target-kind,

Table 6: Hidden normalized regret by notice style.

Notice style	Tool Agent	No-tool LLM	Fixed query
Alias	0.003	0.074	0.613
Compositional distractor	0.003	0.146	0.643
Direct	0.003	0.071	0.124
Indirect operational	0.818	1.054	0.974

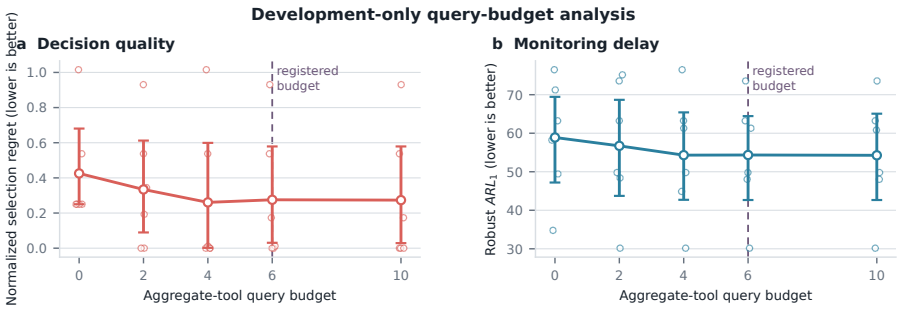


Figure 5: Development-only query-budget analysis. Six calls were frozen before hidden evaluation.

and ARL_0 conditions passing. We describe the hidden results as one-shot controlled-benchmark evidence rather than a completed independent confirmation. No target was removed, no hidden item was rerun, and the threshold was not changed after observation.

The Agent's mean log score was 1.357, compared with 0.834 for the no-tool LLM. This contrasts with its slightly better Brier score and much lower decision regret. A small number of overconfident probability errors accounts for the gap. Better uncertainty calibration is a separate improvement target. The log-score gap also shows what the distributional output buys: it makes overconfidence visible in a way a single identifier could not.

E Illustrative Development Trace

This qualitative example comes from the development split and is not part of the held-out statistical result. One indirect notice described an early cardiovascular-ICU pattern involving acetaminophen, ranitidine, enteric-coated aspirin, albumin, and rapid electrolyte input, without naming the target exposure. The no-tool LLM assigned 0.243 probability to the correct potassium-chloride-bolus cohort, left 0.522 unknown mass, and therefore selected the blind bank. The Agent searched the catalog, profiled the leading concept, searched for the electrolyte and albumin pattern, and compared three usage fingerprints. It assigned 0.820 to the correct cohort, 0.120 to a related potassium-chloride-plus-ventilation conjunction, 0.030 to intravenous acetaminophen, and 0.030 unknown mass. The common solver then selected the correct focused action, which reduced development regret from 1 to 0.

F Released Artifacts

The supplementary material contains the code and the derived benchmark data; its README describes their contents and use. The released notices are synthetic texts generated from sanitized aggregate facts and screened for local identifiers, exact aggregate values, answer keys, and monitoring terminology.

Patient rows, per-patient predicate membership vectors, and residual streams are derived from MIMIC-IV and are not ours to redistribute under the PhysioNet data use agreement. The extraction and construction code is released instead, so a credentialed reader can regenerate them from their own copy of MIMIC-IV. Credentials are read from the environment and no key is included.

G Licensing and Availability

Code is released under the MIT license. The derived benchmark data are released under CC BY 4.0. Both will be publicly available upon publication. Patient-level material remains subject to the PhysioNet agreement and is not redistributed.

H Data Access and Source Choice

We use MIMIC-IV 3.1, which is publicly listed on PhysioNet at no cost. Access requires completion of CITI human-subjects research training and acceptance of the credentialed-data use agreement; no institutional affiliation and no fee is required.

The benchmark needs four properties at once: real first-24-hour exposure events from independent prescription, administration, ICU-input, and ICU-procedure sources, so that cohort membership is a matter of record rather than of authorship; a matured multi-outcome vector per encounter; enough patients per cohort that a 100-patient suppression floor still leaves usable aggregates; and permission to publish derived aggregate statistics. A fully synthetic EHR corpus fails the first property by construction, because cohort membership would then be defined by the same process that writes the notices, which is precisely the confound the design exists to avoid. Billing and claims extracts lack administration-level and ICU-procedure granularity. MIMIC-IV satisfies all four.

I Computing Environment

Experiments ran on one workstation: Ubuntu 24.04.4 LTS under WSL2 (kernel 6.18.33.2-microsoft-standard-WSL2), Intel Core Ultra 9 285 with 24 cores, 23 GiB of RAM, and an NVIDIA RTX 2000 Ada with 16 GB. Monte Carlo calibration and verification are CPU-bound and run with 16 worker processes. The sentence-embedding baseline leaves device selection to the library default, which uses the GPU when one is present.

Python is 3.12.3; the dependency set is pinned in the lock file included with the code.

Both hosted models were reached through an OpenAI-compatible relay endpoint using the `openai` client rather than through either vendor's first-party API. Model identifiers were pinned and the resolved identifier was recorded on every call, and the evaluated Agent was required to resolve to `gpt-5.5` while the notice generator was pinned to `claude-haiku-4-5-20251001` with a hard check that the evaluated family was not Claude. One consequence should be stated plainly: because access was mediated, we can attest the identifier the endpoint returned but not independently attest the upstream weights, and the endpoint is private infrastructure, so exact re-execution requires an equivalent endpoint. Hosted models are not reproducible artifacts in the sense that the numerical layer is.

J Randomization and Seeds

Every stochastic stage draws from a named integer base, and the stages occupy disjoint seed namespaces so that no action is ever scored on the streams that selected it.

Table 7 lists them. The two hidden-set bootstrap contrasts use bases 2026072214 (against the strongest non-LLM method) and 2026072215 (against the same model without tools). The calibration tolerance is 5% of the ARL_0 target, the run-length cap is 1,600 arrivals for calibration streams and 1,000 for scenario streams, and a per-stream seed is derived deterministically from its stage base together with the target, action, and replicate indices. The Phase-I and locked patient partition follows a deterministic subject hash, and the six/fourteen development and hidden target split follows a salted target-level hash.

K Complete Hyperparameter Record

Candidate detector grid. The catalog holds 105 parameterized policies across ten families: 32 CUSUM projection, 16 EWMA projection, 8 each of Shewhart max, Hotelling T^2, multivari-

Table 7: Seed bases and Monte Carlo stream counts.

Stage	Seed base	Streams
Global calibration	2026072601	3,000
Focused calibration	2026072602	3,000
Bank reference calibration	2026072603	1,500
Bank calibration	2026072604	3,000
Candidate selection	2026072605	1,200
Cross-action selection	2026072606	1,200
Cross-action verification	2026072607	4,000
Selected-action ARL_0 audit	2026072608	10,000
Target bootstrap	2026072609	20,000

ate CUSUM, EWMA max, MEWMA, multiscale GLR, and kernel-MMD EWMA, and 1 parallel CUSUM. Every configuration is calibrated to the same ARL_0 and scored by two-template mean ARL_1, and the lowest-loss configuration per target is retained. All 22 retained actions are kernel-MMD EWMA with smoothing $\lambda = 0.1$; 21 monitor all five outcomes and one monitors mortality together with prolonged ICU stay.

Monitoring constants. $ARL_0 = 200$ with 5% tolerance; a 30-day outcome horizon; a 24-hour exposure window; five outcomes; 20 candidate predicates; 22 fixed actions; a suppression floor of 100 patients; a persistent location shift of 0.75 on mortality and prolonged ICU stay; and a persistent scale multiplier of 1.8 on ventilation and renal replacement therapy.

Retrieval and non-LLM baselines. Word n-grams over $[1, 2]$ and character n-grams over $[3, 5]$; reciprocal-rank fusion with $k = 60$; a softmax temperature grid of $\{0.25, 0.5, 1, 2, 4\}$; an unknown-mass grid of $\{0, 0.1, 0.25, 0.5\}$; and for the fixed-query pipeline a text-evidence weight grid of $\{0.25, 0.5, 0.75\}$ over the `usage_fingerprint` tool. One global configuration per method is selected on the development split by mean normalized verification regret, with lexical configuration id as the tie-break. The embedding baseline is `sentence-transformers/all-MiniLM-L6-v2` at revision `1110a243fdf4706b3f48f1d95db1a4f5529b4d41` with 384 dimensions.

Agent and ablation. Query budget $B = 6$ interpreted as an exact number of sequential calls, with the development budget curve over $\{0, 2, 4, 6, 10\}$; retrieval shortlist size 6 drawn from the frozen strongest text retrieval; reasoning effort high; temperature left at the provider default rather than set; a cap of 1,800 output tokens per tool turn and 2,400 on the final structured answer; tool choice required until the budget is exhausted; a typed error returned to the model and counted against the budget when a call is invalid; and a final structured answer without tools once the budget is spent.

L Statistical Inference

Uncertainty is reported as a target-clustered bootstrap with 20,000 resamples, and the preregistered rule required the lower bound of each contrast to sit above zero. Clustering at the target is what makes the interval honest here: the four notice styles written for one target share that target's cohort, its membership vector, and its row of the verification matrix, so they are not independent observations, and an interval over the 56 notices would understate the true variance. The estimand is also a ratio of sums of paired target-level gains rather than a location parameter, which is naturally handled by resampling and awkward for a standard rank test.

Target-cluster bootstrap intervals quantify variation across cohorts conditional on the frozen verification-loss matrix. The independent ARL_0 audit evaluates in-control calibration, while the verification losses are Monte Carlo estimates from the 4,000-stream verification stage. A paired rank test over the fourteen target-level gains would be a reasonable descriptive addition, but it was not part of the registered analysis and we do not use one to support the registered claim.