The Same Answer Every Time: Auditing Small Medical LLM Agents

Abhineet Rajendra Kulkarni
Independent author

Pranav Madhav Kuber
Independent author

Anurag Bihani
Independent author

Abstract

We investigate how small open-weight language models make decisions in medical tasks and whether their failures also occur in other settings. We evaluated nine models on wearable fatigue and intensive-care tasks, where they chose among six recommendations, including continued monitoring. We then extended testing to a non-medical tool router and subsets of the Berkeley Function-Calling Leaderboard and JSONSchemaBench. We introduce the Minimum Viable Audit (MVA) to examine which actions models choose, whether their responses contain usable decisions, and whether they copy prompt content. Comparisons across tasks help establish where these behaviors recur. Under the baseline intensive-care prompt, MedGemma chose continued monitoring in all 450 attempts, achieving 33.3% accuracy. On the router, it used all six tools and answered 144 of 150 requests correctly. Meditron-7B produced unusable responses and copied prompt content across intensive-care monitoring and the non-medical evaluations at both four-bit and eight-bit quantization. MedGemma's results motivate checking the medical instructions. Meditron's recurring failures warrant checking the model configuration and how outputs are generated. Differences between tasks and mismatches between medical prompts and scoring rules limit conclusions about the causes of these failures. We recommend reporting selected actions, invalid and copied responses, and results for each prompt condition alongside accuracy.

1 Introduction

Large language models assist with clinical documentation and information handling by drafting notes, summarizing patient records and retrieving medical information. A pilot study of an LLM-powered scribe found that adoption was associated with a median reduction of 6.89 minutes in daily documentation time (Ma et al., 2025). Research is also connecting language models to electronic health records so they can retrieve measurements and request further operations, extending their role from generating text to selecting actions based on patient data (Jiang et al., 2025).

Using patient records in such workflows requires control over where information is processed, who can access it and how long it is retained. The applicable requirements depend on jurisdiction: U.S. HIPAA guidance specifies safeguards and business associate arrangements for relevant cloud processing, and European data-protection rules impose additional conditions on transfers outside the European Economic Area (U.S. Department of Health and Human Services, 2022; European Data Protection Board, n.d.). Local deployment keeps inference inputs within institutional infrastructure, where teams can manage model versions, access, logging and approval requirements. Open weights also allow teams to adapt and evaluate models in their deployment environment (Google, 2026a).

Small, medically adapted models offer a way to run specialized tasks locally with limited hardware. MedGemma illustrates the size tradeoff within one medical model family: Google lists four-bit weight footprints of approximately 3.4 GB for its four-billion-parameter model and 21 GB for its 27-billion-parameter model, excluding runtime memory (Google, 2026a). Local execution also supports settings with intermittent connectivity, such as mobile clinics (Google, 2026a). An agentic workflow can use separate small models for extracting laboratory values and summarizing records (Belcak et al., 2025). We study models with at most eight billion parameters to examine this compact deployment setting.

An agent asked whether a patient needs another test may first need to retrieve the previous result and its date. The model can obtain that information through a *tool*, a software operation with specified inputs and outputs. MedAgentBench evalu-

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3915–3933
November 6-10, 2026 ©2026 Association for Computational Linguistics

ates such interactions in a virtual electronic health record environment, including a task that asks an agent to retrieve the latest glycated hemoglobin measurement and order another test if the result is more than a year old. Across its 300 tasks, the best evaluated model achieved 69.67% task success (Jiang et al., 2025). Completing the workflow requires an explicit request identifying the operation and its inputs; an accompanying explanation can give the clinician a rationale for reviewing it.

We examine the selection step by asking models to choose among a fixed set of tools or recommendations. Each response must identify an allowed choice and select the appropriate one under the task's rules, bringing together the requirements of output validity and decision correctness examined in function-calling and structured-generation benchmarks (Patil et al., 2025; Geng et al., 2025). Our experiments score the selections without executing tools or clinical interventions.

Interpreting accuracy requires examining the choices and errors behind it: repeated defaults and unusable responses call for different investigations. The *Minimum Viable Audit* (MVA) builds on behavioral testing in CheckList and multidimensional evaluation in HELM (Ribeiro et al., 2020; Liang et al., 2023). It combines validity checks, *action distributions* (the proportions of valid selections assigned to each action), inspection for copied prompt content, and a comparison task evaluated with the same checkpoint. When a model repeatedly chooses one action on one task but performs well elsewhere, we inspect the task's instructions and rules. Task differences in labels, instructions and difficulty make this a diagnostic comparison of where failures occur.

In the intensive-care task, each model selects one of six recommended next steps: continue monitoring, recheck vital signs, adjust positioning or oxygen, adjust fluids or vasopressors, escalate to the attending clinician, or transfer to higher-acuity care; Section 3.1 describes the task and the rules used to score these choices.

Two medically adapted models illustrate the distinction (Figure 1). MedGemma 1.5 4B is a four-billion-parameter model trained for medical text and image tasks (Google, 2026b). Under the baseline prompt across three seeds, it selects `continue_monitoring` for every intensive-care scenario, giving it 33.3% accuracy, the frequency of that reference action. On the non-medical tool router, the same checkpoint uses all six tools and makes 144 correct selections in 150 attempts, or 96.0%.

Meditron-7B was developed through continued medical pretraining of the seven-billion-parameter Llama 2 model; we evaluate a four-bit quantized version, Q4 (Chen et al., 2023). It scores 4.9% on the intensive-care task. Of its 450 responses, 384 contain no valid action, including 282 with placeholder actions, and the remaining 66 all select `continue_monitoring`.

We apply MVA to five medical and four general-purpose or comparator models on a non-medical six-tool router, wearable fatigue decisions and intensive-care monitoring. Adapted subsets of the Berkeley Function-Calling Leaderboard (BFCL) and JSONSchemaBench extend the analysis to supplied-function naming and schema-valid generation. Additional experiments examine whether the observed patterns persist across quantization levels, prompt variants, output formats, databases and related checkpoints.

Our contributions are:

- A reusable audit procedure that combines action-distribution measurements, output inspection and cross-task comparison, with explicit scoring definitions, reference mappings and documentation of mismatches between deployed prompts and scoring rules.
- An empirical analysis of nine small open-weight models, demonstrating how the audit distinguishes a medical-task default accompanied by strong routing performance from frequent unusable responses and concentrated valid choices.

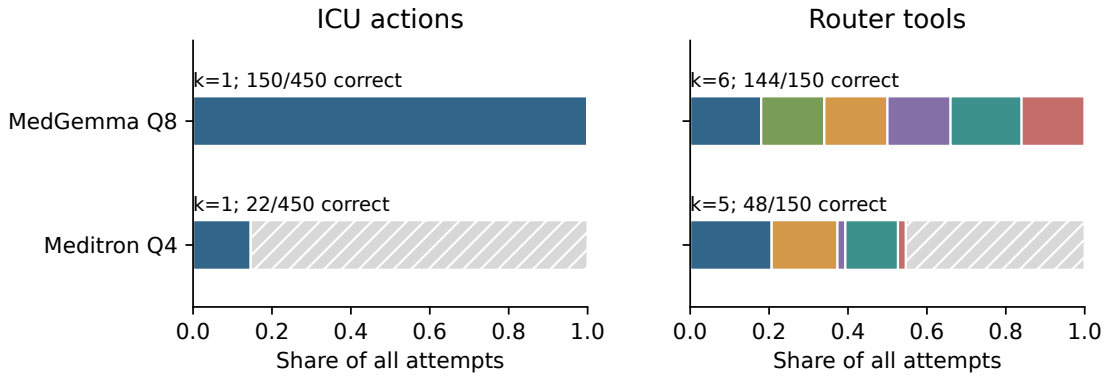
We first define the audit and its measurements, then describe the experimental setup and results.

2 The Minimum Viable Audit

2.1 Procedure

The Minimum Viable Audit combines measurements of a model's selected actions with inspection of the responses that produced them. The procedure has four steps.

1. **Specify the decision interface.** Record the inputs, allowed actions, reference rules, prompt and output parser. Check whether the prompt communicates the rules used to score the responses, since a disagreement between the



Valid selection. Asked to find a saved spreadsheet, MedGemma returns `file_search`.

Copied placeholder. Meditron's response to router case `tr_008` retains `<one of the 6 tool names exactly>`.

Figure 1: ICU and tool-router outputs across three seeds. Colored segments show valid actions; hatched gray shows invalid outputs. Colors identify actions separately within each panel. MedGemma selects one ICU action and all six router tools. Meditron's valid router responses cover five tools, alongside frequent invalid responses. Each seed uses the same scenarios.

prompt and the reference can lower accuracy even when the model follows its instructions.

2. **Measure the outputs.** Report accuracy, validity and the distribution measurements defined in Section 2.2. Preserve the raw responses so that the summaries can be checked against what the model emitted.

3. **Inspect failures.** Look for missing or malformed actions and content copied from the prompt. Where a copying pattern can be defined precisely, quantify it across the evaluated cases. Copying and invalidity require separate annotations: a copied response may pass the parser, while an invalid response need not contain copied content. Appendix B defines the parsers and copying detectors.

4. **Evaluate a comparison task.** Test the same checkpoint on an unrelated structured-output task and check whether the original pattern recurs, recording differences in instructions, action menus, schemas and scoring.

Additional seeds, prompt variants and decoding controls can test the stability of the observed behavior.

2.2 Measurements

Let N be the number of attempted decisions and V the number for which the declared parser recovers an allowed action. Validity is V/N; tables report the invalid rate, $1 - V/N$, as a percentage. The definition of a valid output depends on the task's parser and output requirements; Appendix B specifies the criteria used for each task. Accuracy is the proportion of all attempts whose selected action matches the reference. Invalid outputs count as errors in accuracy and recall.

For each allowed action a, let n_a be its count among valid responses and $p_a = n_a/V$. We describe the action distribution using:

$$k = |\{a : n_a > 0\}|, \tag{1}$$

$$H = -\sum_{a:n_a>0} p_a \log_2 p_a, \tag{2}$$

$$d = \max_a p_a, \tag{3}$$

where k counts the distinct actions used, H is Shannon entropy in bits, and d is the share assigned to the most frequent action. A model that always emits one valid action has $k = 1$, $H = 0$ and $d = 1$.

We also report recall for each reference class: the fraction of cases with that reference action that the model selects correctly. Reference classes with no cases in the evaluation have undefined recall and are excluded from the zero-recall criterion below.

Our descriptive concentration flag counts four conditions: $k \leq 2$, $H \leq 1$ bit, $d \geq 0.80$, and zero recall for at least one reference class present in the evaluation. We denote the number of satisfied conditions by M and flag $M \geq 2$. When no response is valid, H and d are undefined; we report output failure rather than computing M.

The concentration flag prioritizes cases for inspection. We interpret it using reference-class frequencies, the confusion matrix, invalidity and copying, since a skewed task may legitimately favor one action. Appendix F.2 examines sensitivity to the thresholds.

2.3 Interpreting a cross-task comparison

We use *context-specific defaulting* to describe a concentrated action profile on the original task accompanied by varied, successful decisions on a comparison task. The checkpoint's successful selections elsewhere direct attention to the original task and its instructions.

Cross-surface output failure refers to related malformed or copied responses across distinct tasks or benchmark formats, which we call evaluation surfaces. Recurrence motivates examining the checkpoint, output template and runtime alongside the original task instructions. The two patterns can coexist: a model may produce invalid responses on some inputs while concentrating its valid selections on one action.

Tasks differ in difficulty, reference construction, wording and output requirements. We interpret the comparison as evidence about where a behavior recurs and use it to guide follow-up checks of task instructions, output handling and the checkpoint.

3 Experimental setup

3.1 Tasks and reference actions

We apply the audit to three tasks that require selecting one of six permitted actions. All evaluations are offline: no tools are invoked and no recommended interventions are delivered to participants or patients. Each task requests structured JSON output, with its own action menu, schema, parser and reference rules. Medical responses use an `action` field; router responses use a `tool` field.

Non-medical tool routing. The router provides the non-medical comparison task and a simple example of the shared decision interface: selecting an operation for an everyday request. Its menu contains a calculator, web search, calendar, email, file search and code execution. The 50 unique authored requests, evaluated at three seeds for 150 attempts, comprise 30 with an unambiguous tool choice and 20 whose interpretation depends on priority rules supplied in the prompt. For example, those rules distinguish arithmetic over supplied numbers from a request to execute a program. Each response includes the selected tool, a rationale and a confidence label.

Wearable fatigue decisions. This task replays observations from a 14-participant study of trunk flexion, or bending forward at the waist (Kuber et al., 2024). At each cycle, the model receives a fatigue classification, confidence and session context. It can continue the session, suggest a short break, recommend a posture change, modify the task, end the session, or recommend an offline exoskeleton refit. Reference actions follow an author-written mapping of participants' perceived-exertion ratings and their changes over time. The wearable task replays recorded cycles in order: selected interventions update the history shown to the model, task modifications are logged as simulator annotations without changing later fatigue observations, and selecting the end-session action stops the replay.

The ten scenarios provide up to 174 cycles per seed. A model can end a session before its final cycle, so the number of observed decisions depends on its actions: Gemma 3n E2B produces 222 decisions across the main runs, compared with 522 for each other model. We score only the cycles each model encounters. Some reference rules use the following cycle to confirm an exertion spike, giving the retrospective labels information unavailable to an online decision agent. Appendix A details the mapping and the differences in exposure and information.

Intensive-care monitoring. The intensive-care-unit (ICU) task uses 150 one-hour scenarios extracted from eICU-CRD (Pollard et al., 2018, 2019, 2026). Inputs summarize vital signs, recent treatments and an early-warning signal. The model selects among continuing monitoring, rechecking vital signs, adjusting positioning or oxygen, adjusting fluids or vasopressors, escalating to the attending clinician, and transferring to higher-acuity care. Vasopressors are drugs used to support blood pressure.

Reference actions come from an author-written priority hierarchy informed by sepsis guidance and adapted to the available retrospective measurements (Evans et al., 2021; Singer et al., 2016). Independent clinical adjudication of the labels is undocumented. The National Early Warning Score, NEWS2, supplies the deterioration signal (Royal College of Physicians, 2017). An additional input called "confidence" increases with the distance of NEWS2 from the deterioration threshold. The

prompt describes this value as calibrated, although its relationship to the probability of a correct label has not been measured.

On the supplied single-window inputs, the prompt's gate blocks intervention in 87 of the 100 scenarios with non-monitoring reference actions. ICU accuracy therefore measures agreement with authored rules that partly conflict with the supplied instructions. Appendix A gives the rule ordering, input definitions and prompt–reference differences. Additional evaluations use 17 held-out eICU scenarios and 141 scenarios from MIMIC-IV (Johnson et al., 2023, 2024); their construction and use are described in Appendices A and E.

3.2 Public output tests and model cohort

Two public benchmark subsets extend the output checks beyond the authored tasks. We use 100 instances from BFCL's simple function-call setting and 100 schemas from JSONSchemaBench (Patil et al., 2025; Geng et al., 2025). Each BFCL instance supplies one function. Our adaptation scores the returned name and detects copied definitions; it omits argument and execution checks. JSONSchemaBench measures whether a generated instance validates against its target schema. These subsets cover six cohort models and four controls (Table 4); Gemma 3n E2B, Phi-4-mini and Llama-2 Chat were not evaluated on them. Scores use all attempts, including runtime errors, under the local scoring rules.

We selected locally runnable, quantized open-weight models within the study's eight-billion-parameter limit. The main cohort contains five medical models: MedGemma 1.5 4B, BioMistral 7B, Meditron 7B, OpenBioLLM 8B and Aloe-Qwen 7B (Sellergren et al., 2026; Labrak et al., 2024; Chen et al., 2023; Pal and Sankarasubbu, 2024; Garcia-Gasulla et al., 2026). Gemma 3n E2B and Phi-4-mini are the general-purpose models; Mistral Instruct v0.2 and Llama-2 Chat serve as comparators (Google DeepMind, 2025; Abouelenin et al., 2025; Jiang et al., 2023; Touvron et al., 2023). All evaluations use text inputs, including for models that support other modalities.

Additional comparisons use MedGemma Q4, Gemma 3n E4B, Meditron Q8 and Meditron3 8B. Quantization variants test whether output patterns persist with different weight representations. The successor checkpoint, Meditron3, tests whether they recur in a later release. Gemma 3n E4B provides an approximate size comparison; MedGemma's parent checkpoint belongs to the separate Gemma 3 family (Google, 2026b). Differences in training history also limit attribution of model differences to medical adaptation. Appendix H lists the exact model tags and run groups.

3.3 Execution, baselines and uncertainty

The main evaluations ran locally through Ollama[1] using GGUF weights at temperature 0.2 with three seeds. Appendix E describes the temperature checks and additional seed configurations. The Q8 router-format extension used a later Ollama release than the historical experiments, so differences between those runs may reflect execution changes as well as quantization. Appendices D and H document the configurations, prompt variants and run groups.

Four baselines provide context for ICU accuracy. Always selecting `continue_monitoring` gives the corresponding reference-class frequency. A NEWS2 threshold rule escalates to the attending at scores of at least five and continues monitoring otherwise. A multilayer perceptron (MLP) and XGBoost (Chen and Guestrin, 2016) learn the reference mapping from tabular features, with performance evaluated using saved five-fold out-of-fold predictions. The supervised baselines train on reference labels; the main prompted evaluations use no labeled demonstrations. Appendix C gives the baseline features, preprocessing and hyperparameters.

We estimate descriptive accuracy intervals by resampling scenarios while keeping each scenario's repeated seed responses together, so seeds characterize variation on the same cases. The intervals do not account for all dependence between scenarios from the same patient or remove the effects of stratified sampling. Wearable comparisons remain descriptive (Section 3.1). Router comparisons are paired by request; Appendix F specifies the bootstrap procedure and the exploratory equivalence analysis.

The investigation developed from a locally dated ICU protocol into an exploratory cross-domain analysis. Appendix G documents the planning history, differences between the planning documents and subsequent additions.

[1] https://github.com/ollama/ollama

4 Results

4.1 One-action medical defaults coexist with accurate tool routing

Under the baseline prompts, MedGemma 1.5 4B selects `continue_monitoring` in all 522 wearable decisions and all 450 ICU decisions. Its accuracies, 40.2% and 33.3%, equal the frequency of that reference action in each evaluation (Table 1). Recall is zero for the other five ICU actions. On the supplied inputs, the gate directs continuation in 137/150 ICU scenarios (Section 3.1). MedGemma also continues in the 13 cases where intervention is permitted, comprising eight transfer and five escalation references.

On the non-medical router, the same Q8 checkpoint uses all six tools and selects correctly in 144/150 attempts (Figure 1). For example, it selects `file_search` for a request to locate a spreadsheet. Gemma 3n E4B, the approximate size comparator, answers 147/150 correctly. MedGemma names the supplied function correctly in 299/300 BFCL attempts and yields schema-valid instances in 200/300 attempts on JSONSchemaBench (Appendix B, Table 4). Its Q4 variant retains the single-action ICU pattern and answers 147/150 router attempts correctly. Together, its medical default and varied, accurate router selections fit the *context-specific defaulting* pattern (Section 2.3).

Concentration is common across the cohort (Table 1): eight of nine models meet the ICU flag, $M \geq 2$. Mistral Instruct v0.2 and Llama-2 Chat always select `continue_monitoring`; the general-purpose Gemma 3n E2B and Phi-4-mini assign over 90% of valid actions to it. BioMistral assigns 415/445 valid responses to rechecking vital signs, despite a gate favoring continuation. Only Aloe-Qwen falls below the threshold, with four actions, $H = 1.42$ bits and $M = 1$.

4.2 Meditron's failures combine invalid output, copying and concentrated choices

Meditron-7B-Q4 produces 66 valid actions in 450 ICU attempts, all `continue_monitoring`, with 22 correct decisions overall. Its 384 invalid responses comprise 282 placeholder actions, 62 dictionary-valued actions, 36 null actions and four non-JSON responses (Appendix D, Table 6). Figure 1 shows the overall distribution; the confusion matrices in Appendix C, Figure 2 show which reference classes receive invalid or incorrect responses.

The router exposes a different distribution alongside similar output problems. Meditron produces 82 valid selections spanning five tools, with $H = 1.90$ bits and $M = 1$, but answers only 48/150 attempts correctly. Of the 150 responses, 68 contain the literal placeholder `<one of the 6 tool names exactly>`, including the saved response to scenario `tr_008`.

Copying also occurs on the public subsets (Appendix B, Table 4). On BFCL, 220/300 responses reproduce a function definition and another 29 echo a placeholder. On JSONSchemaBench, 58/300 responses reproduce the complete input schema, ignoring whitespace and key order. A broader metadata heuristic flags 177/300 responses, although it can also match legitimate instance fields. The exact-schema count excludes partial copies.

Output validity also varies across tasks. For OpenBioLLM, 165/300 BFCL attempts fall in the *other-invalid* category: parseable JSON that is not a detected copy but lacks a recognized string-valued function-name field. It produces valid selections in 150/150 router and 449/450 ICU attempts.

4.3 Meditron-7B copying persists at Q8 and is reduced in Meditron3

The Meditron-7B Q4 and Q8 controls share prompts, scoring, seeds and non-weight model layers; the historical Q8 runtime version was not recorded (Appendix D). Table 2 compares their results across four tasks.

At Q8, valid ICU responses increase from 66/450 to 140/450 and correct decisions from 22 to 47. Every valid response still selects `continue_monitoring`, leaving $k = 1$, $H = 0$, $d = 1$ and zero recall for five reference classes. Both quantizations receive $M = 4$. Placeholder counts vary sharply across seeds: 0, 132 and 150 at Q4, compared with 0, 0 and 150 at Q8 (Appendix D, Table 6).

Router accuracy rises from 48/150 to 66/150, and invalid responses fall from 68 to 39. Responses containing the exact tool-list placeholder fall from 68 to 38. Q8 uses all six tools with nonzero recall for each, reducing M from 1 to 0.

Performance shows no consistent improvement across the two public subsets. Q8 names the supplied BFCL function correctly in nine attempts, compared with 23 at Q4, and produces 225 function-definition copies and 56 placeholder echoes. JSONSchemaBench validity rises from 76/300 to 97/300; exact schema copies fall from 58 to 40 (Appendix B, Table 4).

Model	N_A	Acc. A	Acc. B	Inv. B (%)	k_B	H_B	d_B	M_B
Gemma 3n E2B	222	0.57	0.33	3.33	3	0.29	0.96	3
Phi-4-mini	522	0.24	0.34	0.00	4	0.41	0.93	3
MedGemma 4B Q8	522	0.40	0.33	0.00	1	0.00	1.00	4
BioMistral 7B Q4	522	0.27	0.16	1.11	4	0.42	0.93	3
Meditron 7B Q4	522	0.13	0.05	85.33	1	0.00	1.00	4
Mistral Instruct v0.2	522	0.40	0.33	0.00	1	0.00	1.00	4
Llama-2 Chat	522	0.40	0.33	0.00	1	0.00	1.00	4
OpenBioLLM 8B	522	0.40	0.33	0.22	2	0.31	0.94	4
Aloe-Qwen 7B	522	0.21	0.33	0.00	4	1.42	0.62	1

Table 1: Main-cohort wearable (A) and ICU (B) results. Wearable denominators vary because early session termination changes exposure; those accuracies are not a common-case ranking. ICU uses the same 150 scenarios at three seeds ($N = 450$ per row). Accuracy includes invalid actions as errors; invalidity is action-parser failure, not full-field schema validation. H (bits) and dominance (d) condition on valid actions. M counts four concentration criteria (Section 2.2).

Weights	ICU acc.	ICU inv.%	k_I	Router acc.	Router inv.%	k_T	BFCL%	JSB%
Q4	0.05	85.33	1	0.32	45.33	5	7.67	25.33
Q8	0.10	68.89	1	0.44	26.00	6	3.00	32.33

Table 2: Meditron Q4/Q8 comparison across four tasks. ICU: $N = 450$; router: $N = 150$; each public-benchmark subset: $N = 300$. BFCL reports correct function name under the local adaptation; JSB reports schema validity, including outputs also flagged by a copying heuristic. The prompts and scoring match; the runs come from different collection periods without a verified runtime match.

Meditron3 changes the base model (Llama 3.1 8B) and training-data mixture (EPFLiGHT, 2026). It answers 135/150 router attempts correctly and names the supplied BFCL function correctly in 297/300 attempts, compared with 48/150 and 23/300 for Meditron-7B-Q4. No function-definition copies or placeholder echoes are detected on BFCL. On JSONSchemaBench, it produces 204/300 valid responses and 11 exact schema copies, compared with 76 valid responses and 58 copies for Meditron-7B-Q4 (Appendix B, Table 4).

All 450 Meditron3 ICU responses are valid, but its choices remain concentrated: 344 select `continue_monitoring` and 106 select fluid or vasopressor adjustment, giving $M = 3$. The continue-monitoring share is 76.4% overall, with counts of 62, 141 and 141 across the three seeds. Appendix I collects the additional controls.

4.4 Reversing the tool list produces a single-tool default

Meditron-7B-Q8 answers 63/150 router attempts correctly with the bullet-format change and 60/150 with the heading change. Reversing the tool order reduces this to 9/150 (Table 3). All 75 valid responses in the reversed condition select `code_execution`, giving $k = 1$, $H = 0$ and $M = 4$.

Accuracy spans 36 percentage points across the Q8 formats. The historical Q4 runs span 30.7 points and show the same single-tool pattern under reversal, subject to the runtime difference described in Section 3.3.

Across the three Q8 conditions, 155/450 responses are invalid. Of these, 148 contain the exact tool-list placeholder; the remaining seven contain other invalid outputs. Pooling the conditions would show use of all six tools and obscure the reversed condition's exclusive selection of `code_execution` among valid responses.

5 Discussion

5.1 Interpreting medical-task defaulting

The ICU prompt and scoring rules specify conflicting decisions for 87 of the 100 non-monitoring reference cases on the supplied inputs (Section 3.1). This conflict leaves the contribution of instruction following to MedGemma's concentration unresolved. Supplying the complete reference hierarchy and required inputs while holding the checkpoint, cases and output format fixed would test that contribution. Improvement concentrated in previously conflicting or underspecified branches would support instruction alignment as a contributor. The MLP and XGBoost achieve 60.0% and 76.0% agreement with the authored labels using

Format	Weights	Correct	Invalid	k	H	d	M
Bullet	Q4	47	69	5	1.88	0.42	1
Bullet	Q8	63	44	6	2.05	0.41	0
Header	Q4	51	48	5	1.56	0.63	1
Header	Q8	60	36	6	1.77	0.54	0
Reverse	Q4	5	118	1	0.00	1.00	4
Reverse	Q8	9	75	1	0.00	1.00	4

Table 3: Tool-router format perturbations, 50 cases at three seeds per row ($N = 150$). Q4 is historical; Q8 was collected on Ollama 0.20.7. Reversing the tool list concentrates every valid response on code execution at both quantizations. H is in bits; d is the share among valid actions and M counts concentration criteria.

saved out-of-fold predictions (Appendix C). Their access to labeled training cases differs from the prompted models' access to prose instructions; aligning those instructions can clarify how much of the gap remains.

Caution-related language may also favor continued monitoring when intervention conditions are complex or underspecified. Continued monitoring can itself be inappropriate, so a test of risk framing needs to vary the surrounding language while preserving the decision rules. Google's model card describes MedGemma as a starting point for task-specific adaptation and notes its potential prompt sensitivity (Google, 2026b).

The ICU hierarchy combines thresholds, trends and precedence. Paired cases differing in one relevant field can test recognition of individual triggers; cases satisfying several conditions can test their precedence. Concentration across medical and general-purpose models also calls for comparisons of matched parent and adapted checkpoints under compatible execution settings, with instruction alignment checked for each task.

5.2 Repairing the output interface can leave the decision failure intact

With the joint reasoning-field and schema intervention, Meditron produces 449 valid ICU responses out of 450, all selecting `continue_monitoring`, although its wearable choices expand to four actions (Appendix E). Under the same intervention, MedGemma uses four ICU actions, but accuracy falls from 33.3% to 28.9%, with 56 invalid responses. Repair evaluations need to track corrected cases, newly introduced errors and per-class recall.

An unfilled placeholder suggests reproduction of the output specification. Checking rendered inputs, role delimiters, generation prefixes and stopping conditions can identify configuration problems; models sharing a base can require different chat formats (Hugging Face, n.d.). Comparing placeholder templates, completed examples and formal schemas with runtime and requests fixed tests how output specification affects validity and decision accuracy. Raw generations support separate decision and format scores. Fixing the rendered template across checkpoints tests whether copying recurs across model versions. Our joint intervention leaves the reasoning field's and schema constraint's individual contributions unresolved.

The reversal result motivates separate tests of tool position, name and description. Placing each tool in several positions distinguishes choices that follow position from persistent `code_execution` selection; separate changes to names and descriptions then test those features. Wei et al. (2024) document sensitivity to both order and option tokens. Reporting each permutation separately preserves concentrated conditions that pooling can hide.

5.3 Applying the audit in research and development

For a record-retrieval agent selecting laboratory results, clinical notes, medication history or clarification, the audit set needs clear requests, missing information and competing tools. Repeated note searches on laboratory requests identify cases for inspection. Successful routing on an unrelated document task directs attention to medical instructions; placeholders across both tasks prioritize the output interface. Retained failure cases track repairs during development; fresh cases test generalization.

Logs must distinguish raw responses, parser decisions and software fallbacks. Silently converting an unparseable response to continued monitoring conceals the failure and makes it indistinguishable from a model-generated recommendation.

Report performance on representative workload cases separately from a stratified challenge set of uncommon actions. Clinical evaluations addition-

ally need adjudicated reference decisions and an assessment of missed and unnecessary interventions. For an escalation-support prototype, separate counts of missed and unnecessary escalations expose clinical consequences that similar accuracy scores can obscure. A fixed regression set, with versioned prompts, model identifiers, parsers and runtime settings, can track changes after updates to quantization, schemas or serving software. Accounting for case-mix changes avoids misattributing shifts in action frequencies to the model.

The audit suits fixed-menu agents where repeated defaults can obscure failures to distinguish cases and a suitable comparison task is available. Candidate applications include cybersecurity incident routing, content-moderation escalation, industrial maintenance triage, legal intake and customer-support routing. These applications remain untested here.

5.4 Priorities for validating and extending the audit

Validation requires failures characterized independently of MVA. Controlled systems can supply always-default policies, malformed outputs, deliberate copying and conflicting instructions, including combinations because these behaviors can coexist. A separate evaluation of naturally occurring failures tests whether the distinctions extend beyond constructed examples.

Varying reference frequencies, menu size and sample size tests whether the concentration flag mistakes legitimate imbalance for failure. False flags and missed failures on separate tasks assess thresholds chosen during development. Comparisons with accuracy alone, invalidity alone and a dominant-action rule can establish what the full procedure adds.

A copying-detector annotation study needs complete and partial copies, legitimate schema fields and unrelated malformed responses. Exact matching and metadata heuristics need separate precision and recall estimates; validity needs separate annotation because copied content can pass a parser.

Matched medical and non-medical versions of a logical problem can test domain framing while preserving rules, action count, output requirements and reference frequencies. Comprehension checks and independent difficulty assessments can evaluate their comparability.

A prompt-aligned test measures rule execution; clinical judgments require independently adjudicated cases. Wearable comparisons need a common replay set independent of model stopping decisions, with labels restricted to information available at each decision time.

Multi-step failures include incorrect arguments, retrieval of the wrong record and misuse of returned information, even when the selected tool is appropriate. Following these transitions tests error recovery and dependencies between decisions.

5.5 Relationship to prior evaluation work

MVA links an observed output pattern to a diagnostic cross-task comparison, building on CheckList and HELM (Ribeiro et al., 2020; Liang et al., 2023). MEDIC and MedAgentBench assess broader clinical utility and workflows (Kanithi et al., 2026; Jiang et al., 2025); biomedical-specialization studies compare adaptation benefits (Dorfner et al., 2025). The MIMIC-IV/eICU kit describes a sequential setting (Özkan, 2026). Prompt-format research measures presentation sensitivity (Sclar et al., 2024); our tool-order reversal produces a single-tool default.

6 Conclusion

The Minimum Viable Audit combines response inspection and cross-task testing to characterize failures in small language-model agents. Across nine small open-weight models, MedGemma's repeated monitoring recommendation coexisted with accurate non-medical tool selection, while Meditron-7B produced copied and unusable responses across tasks at both four-bit and eight-bit quantization. Adding a reasoning field and schema constraint improved output validity or action diversity without consistently improving decisions. These findings support evaluating changes through both decision accuracy and inspection of the resulting responses. We recommend reporting action frequencies, invalid and copied responses, and results for each prompt condition alongside accuracy, with a non-medical comparison showing whether observed failures extend beyond the medical task.

Acknowledgments

Generative AI tools assisted with coding and editing the manuscript. The authors take responsibility for the final content, analyses and citations.

Limitations

MVA is demonstrated on selected small open-weight models and mostly single-step tasks; its diagnostic sensitivity and specificity remain unmeasured. Concentration thresholds and copying detectors can misclassify outputs. The clinical reference actions are authored rules without documented independent adjudication. Prompt–reference mismatches, the uncalibrated ICU confidence input and wearable labels using later observations qualify the accuracy results. Model-dependent session termination also changes wearable exposure. Small samples, stratification and correlated observations limit generalization; repeated seeds do not add independent cases.

Cross-task comparisons change instructions, action menus, parsers and difficulty, while checkpoint comparisons change training or execution conditions. They provide diagnostic contrasts without isolating causes. Evaluations of larger hosted models are excluded because of clinical-data processing restrictions. Our findings are limited to the evaluated small open-weight models.

Historical model and runtime records are incomplete, and the differing local planning documents lack independently verified registration timestamps. Clinical-source agreements constrain reuse of record-level data. Appendices G and H document provenance and availability.

Ethical considerations

These retrospective evaluations executed no clinical interventions. The reported open-weight inference ran locally. The ICU databases are deidentified resources subject to credentialing and data-use requirements (Pollard et al., 2018, 2019, 2026; Johnson et al., 2023, 2024). The original wearable study documents its human-subject procedures (Kuber et al., 2024). Clinical safety requires separate validation: action diversity and passing the audit provide no safety guarantee, and exact-label accuracy does not capture the differing consequences of missed escalation and unnecessary intervention.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi-ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, and 55 others. 2025. Phi-4-Mini technical report: Compact yet powerful multimodal language models via Mixture-of-LoRAs. *Preprint*, arXiv:2503.01743.

ARDS Definition Task Force. 2012. Acute respiratory distress syndrome: The Berlin definition. *JAMA*, 307(23):2526–2533.

Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. 2025. Small language models are the future of agentic AI. *Preprint*, arXiv:2506.02153v2. Version 2.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. Association for Computing Machinery.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. MEDITRON-70B: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Clément Christophe, Praveen K. Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. Med42-v2: A suite of clinical LLMs. *Preprint*, arXiv:2408.06142.

ContactDoctor. 2024. ContactDoctor-Bio-Medical: A high-performance biomedical language model. Model card. Accessed September 26, 2026.

Felix J. Dorfner, Amin Dada, Felix Busch, Tianyu Han, Daniel Truhn, Jens Kleesiek, Madhumita Sushil, Jacqueline Lammert, Marcus R. Makowski, Lisa C. Adams, and Keno K. Bressem. 2025. Evaluating the effectiveness of biomedical fine-tuning for large language models on clinical tasks. *Journal of the American Medical Informatics Association*, 32(6):1015–1024. Author list as corrected in doi:10.1093/jamia/ocaf189.

EPFLiGHT. 2026. Meditron3-8B model card. Accessed September 26, 2026.

European Data Protection Board. n.d. International data transfers. Data protection guide for small business. Accessed September 27, 2026.

Laura Evans, Andrew Rhodes, Waleed Alhazzani, Massimo Antonelli, Craig M. Coopersmith, Craig French, Flávia R. Machado, Lauralyn Mcintyre, Marlies Ostermann, Hallie C. Prescott, Christa Schorr, Steven Simpson, W. Joost Wiersinga, Fayez Alshamsi, Derek C. Angus, Yaseen Arabi, Luciano Azevedo, Richard Beale, Gregory Beilman, and 41 others. 2021. Surviving Sepsis Campaign: International guidelines for management of sepsis and septic shock 2021. *Intensive Care Medicine*, 47(11):1181–1247.

Dario Garcia-Gasulla, Jordi Bayarri-Planas, Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Adrian Tormos, Daniel Hinjos, Pablo Bernabeu-Perez, Anna Arias-Duart, Pablo Agustin Martin-Torres, Marta Gonzalez-Mallo, Sergio Alvarez-Napagao, Eduard Ayguadé-Parra, and Ulises Cortés. 2026. The Aloe family recipe for open and specialized healthcare LLMs. *npj Digital Medicine*, 9(1):558.

Saibo Geng, Hudson Cooper, Michał Moskal, Samuel Jenkins, Julian Berman, Nathan Ranchin, Robert West, Eric Horvitz, and Harsha Nori. 2025. JSONSchemaBench: A rigorous benchmark of structured outputs for language models. *Preprint*, arXiv:2501.10868.

Google. 2026a. Frequently asked questions: Health AI Developer Foundations. Updated July 14, 2026. Accessed September 27, 2026.

Google. 2026b. MedGemma 1.5 model card. Accessed September 26, 2026.

Google DeepMind. 2025. Gemma 3n model card. E2B and E4B family documentation. Accessed September 26, 2026.

Hugging Face. n.d. Chat templates. Transformers documentation. Accessed September 27, 2026.

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

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. 2025. MedAgentBench: A virtual EHR environment to benchmark medical LLM agents. *NEJM AI*, 2(9):AIdbp2500144.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. 2024. MIMIC-IV (version 3.1). PhysioNet.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Liwei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.

Praveenkumar Kanithi, Clément Christophe, Marco AF Pimentel, Tathagata Raha, Prateek Munjal, Nada Saadi, Hamza A Javed, Svetlana Maslenkova, Nasir Hayat, Ronnie Rajan, and Shadab Khan. 2026. MEDIC: Comprehensive evaluation of leading indicators for LLM safety and utility in clinical applications. *Transactions on Machine Learning Research*.

Pranav Madhav Kuber, Abhineet Rajendra Kulkarni, and Ehsan Rashedi. 2024. Machine learning-based fatigue level prediction for exoskeleton-assisted trunk flexion tasks using wearable sensors. *Applied Sciences*, 14(11):4563.

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

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Stephen P. Ma, April S. Liang, Shreya J. Shah, Margaret Smith, Yejin Jeong, Anna Devon-Sand, Trevor Crowell, Clarissa Delahaie, Caroline Hsia, Steven Lin, Tait Shanafelt, Michael A. Pfeffer, Christopher Sharp, and Patricia Garcia. 2025. Ambient artificial intelligence scribes: Utilization and impact on documentation time. *Journal of the American Medical Informatics Association*, 32(2):381–385.

Meta. 2024. Meta Llama 3 model card. Accessed September 26, 2026.

Göktug Özkan. 2026. LLM Clinical Inertia: Cross-task safety evaluation kit (code & data). Zenodo. Software, version 1.0.1.

Ankit Pal and Malaikannan Sankarasubbu. 2024. OpenBioLLMs: Advancing open-source large language models for healthcare and life sciences. Hugging Face model card. Accessed September 28, 2026.

Shishir G. Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. 2025. The Berkeley Function Calling Leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *PMLR*, pages 48371–48392.

Tom Pollard, Alistair Johnson, Jesse Raffa, Leo Anthony Celi, Omar Badawi, and Roger Mark. 2019. eICU Collaborative Research Database (version 2.0). PhysioNet.

Tom Pollard, Benjamin E. Moody, Li-wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. 2026. PhysioNet as a global platform for biomedical research. *Nature Health*, 1(8):792–795.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. 2018. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific Data*, 5:180178.

Qwen Team. 2024. Qwen2.5-7B-Instruct model card. Accessed September 26, 2026.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912. Association for Computational Linguistics.

Todd W. Rice, Arthur P. Wheeler, Gordon R. Bernard, Douglas L. Hayden, David A. Schoenfeld, and Lorraine B. Ware. 2007. Comparison of the SpO2/FiO2 ratio and the PaO2/FiO2 ratio in patients with acute lung injury or ARDS. *Chest*, 132(2):410–417.

Royal College of Physicians. 2017. National Early Warning Score (NEWS) 2: Standardising the assessment of acute-illness severity in the NHS. Updated report of a working party, Royal College of Physicians, London.

Donald J. Schuirmann. 1987. A comparison of the Two One-Sided Tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, Liron Yatziv, Tiffany Chen, Bram Sterling, Kenneth Philbrick, Richa Tiwari, Yun Liu, Madhuram Jajoo, Chandrashekar Sankarapu, Swapnil Vispute, and 23 others. 2026. MedGemma 1.5 technical report. *Preprint*, arXiv:2604.05081.

Mervyn Singer, Clifford S. Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, Richard S. Hotchkiss, Mitchell M. Levy, John C. Marshall, Greg S. Martin, Steven M. Opal, Gordon D. Rubenfeld, Tom van der Poll, Jean-Louis Vincent, and Derek C. Angus. 2016. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA*, 315(8):801–810.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

U.S. Department of Health and Human Services. 2022. Guidance on HIPAA & cloud computing. Content last reviewed December 23, 2022. Accessed September 27, 2026.

Sheng-Lun Wei, Cheng-Kuang Wu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. Unveiling selection biases: Exploring order and token sensitivity in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5598–5621, Bangkok, Thailand. Association for Computational Linguistics.

A Task definitions and reference mappings

A.1 Wearable reference actions

The wearable labels use back and leg perceived-exertion ratings, recent changes and a back-weighted maximum. The frozen function `gt_action_v2` in `score_agent_v2.py` applies the rules in order. An early-session guard returns continue monitoring in the first two cycles when maximum exertion is below eight; a drop of at least two points from the preceding cycle also returns continue. Maximum exertion of at least eight, or at least seven sustained into a second cycle, can end the session.

A rise of at least two points to a rating of at least five triggers a check of the next recorded cycle. A confirmed spike can end the session; an unconfirmed spike modifies the task unless its current rating reaches seven, which also triggers session termination. This retrospective check gives the reference information unavailable to an online agent. Later branches cover sustained load, back–leg divergence, breaks, posture changes and refit decisions. The complete ordering is preserved in the scoring function.

The ten scenarios contain 174 possible cycles per seed. Session-ending actions reduce exposure: Gemma 3n E2B produces 222 decisions across three seeds, compared with 522 for each other main-cohort model. Unvisited cycles are excluded from scoring, so these models are evaluated on different observed cycles.

A.2 ICU rule hierarchy

The ICU labels come from `apply_gt_icu_v1` in `extract_icu_scenarios.py`. Mean arterial pressure (MAP), oxygen saturation (SpO_2), inspired oxygen fraction (FiO_2), treatments and trends enter the following ordered rules:

1. *Transfer.* MAP below 65, vasopressin use, and a maximum norepinephrine or epinephrine dose of at least 0.25 micrograms/kg/min in the preceding hour, with at least four hours since the earliest recorded nonzero infusion of that drug in the preceding 24 hours. This duration proxy does not verify continuous high-dose exposure.

2. *Escalate.* Low, non-improving MAP after adequate recorded fluid volume; elevated lactate with vasopressor use and that fluid volume; NEWS2 at least seven; or NEWS2 at least five with selected worsening trends. Adequate fluid volume is defined as 30 mL/kg over three hours.

3. *Adjust fluids or vasopressors.* Borderline falling MAP without a recent bolus, or ongoing low-dose norepinephrine with MAP below 65.

4. *Adjust positioning or oxygen.* $SpO_2 < 92$, $SpO_2/FiO_2 \leq 315$ when $SpO_2 \leq 97$, or a saturation drop of at least four percentage points over the implemented one-hour trend. The code multiplies FiO_2 by 100 and uses the equivalent ratio threshold of 3.15.

5. *Recheck.* MAP within one unit of 65 or SpO_2 within one unit of 92.

6. *Continue.* No earlier branch applies.

The thresholds draw on sepsis guidance and NEWS2 (Evans et al., 2021; Singer et al., 2016; Royal College of Physicians, 2017); their precedence, trend windows and six-action mapping are authored adaptations. Rice et al. (2007) relate a saturation ratio of 315 to an arterial-oxygen-pressure ratio of 300. Our saturation branch omits the additional diagnostic conditions in the Berlin definition (ARDS Definition Task Force, 2012). These research labels have no complete independent clinician-adjudication record.

The 150 cases comprise 50 continue, 25 recheck, 25 oxygen, 20 fluid, 20 escalation and 10 transfer references. This stratification differs from natural action prevalence. Three seeds yield 450 attempts on the same 150 scenarios. Missing fluid documentation, unit conversions and trend construction affect the implemented branches.

A.3 Prompt alignment and sampling

The prompt gates intervention on the deterioration signal and gives a narrower oxygen rule: low saturation together with circulatory conditions. The reference can select oxygen adjustment from its ratio branch alone. The confidence input is also inconsistent with its description: the generator starts at 0.5, increases with distance from the NEWS2 threshold of five, scales that distance by 14 and caps the value at one. It is an uncalibrated magnitude proxy, although the prompt describes 0.80 as approximately 80% correctness.

Applying the four prompt-gate conditions in Appendix J to the 150 archived scenarios permits intervention in 13 and blocks it in 137. Each evaluation supplies one window and no preceding-window history, so the consecutive-window condition cannot be established. Of the 100 non-monitoring references, 87 occur behind the closed gate: all 25 recheck, 25 oxygen and 20 fluid cases, plus 15 escalation and two transfer cases. The other 50 blocked cases have continue references. This is a check of the supplied instructions against the saved inputs, not a clinical adjudication or a rescore of model outputs.

The held-out evaluation contains 17 cases at seed 0. Later few-shot prompts draw examples from that source, so it is no longer untouched for the few-shot condition. The MIMIC replication has 141 cases per seed, including just one transfer reference.

B Parsers and copying measures

B.1 Action validity

The ICU parser accepts a JSON object whose `action` is an allowed string after whitespace trimming. Auxiliary fields are outside this validity check. Dictionary-wrapped values and near-miss strings count as errors. The router parser accepts `tool`, `action` or `name`, can recover nested strings, and normalizes case, spaces, hyphens and nearby enum strings.

The format grid uses a separate router parser. Applying the general parser recovers six additional Mistral outputs, all for request `tr_033` in bullet/header conditions. We retain the grid parser for that comparison. The verification records identify these discrepancies and independently check the stored medical actions and schema-validity categories.

B.2 Copied content

ICU placeholder detection recognizes angle-bracket content in the action field. The Q8 format extension also counts an exact schema-placeholder marker. BFCL scoring accepts alternative function-name fields and a namespace-style fuzzy match; correct arguments are outside its score. Function-definition copies and placeholders are separate categories.

The JSONSchemaBench metadata marker requires at least two schema-related keys, including names such as `type`, `properties` and `required`. Legitimate instances can contain these keys. Exact schema copying instead requires equality of the complete parsed output and input-schema objects, ignoring whitespace and key order; it misses partial copies. Table 4 reports both measures and runtime-error counts. Runtime-error attempts count as unsuccessful in both public-subset scores; they are separate from returned responses that fail parsing or validation. Thus the reported percentages measure success per attempt, rather than validity conditional on receiving a response.

C Recall and baseline construction

The majority rule always continues. The NEWS2 rule escalates to the attending at scores of at least five and continues otherwise, scoring 68/150 (45.3%). The supervised models use saved five-fold out-of-fold predictions.

The MLP inputs are MAP, SpO_2, heart rate, respiratory rate, temperature, lactate, FiO_2, fluid bolus in the preceding three hours, active vasopressor count, NEWS2 and its six-hour change. Median imputation and standardization are fitted within training folds. Hidden layers contain 32 and 16 units; the model uses 2,000 maximum iterations, no early stopping, and random seed 20260517 for fitting and shuffled stratified folds. It scores 90/150 (60.0%). Removing NEWS2 and its change leaves nine features and gives 80/150 (53.3%); the remaining vitals still carry information related to NEWS2.

XGBoost uses the same folds, native missing-value handling, 200 trees, depth six, learning rate 0.1, row/column subsampling of 0.9, a multiclass probability objective, histogram construction and four worker threads. It scores 114/150 (76.0%).

D Quantization and router-format controls

The historical Meditron Q4/Q8 control uses the baseline prompt-9 ICU runner and identical scoring. The archived log records matching template, system and parameter layer digests. Table 6 separates the seed-specific output categories. Incorrect or malformed responses were not retried. The later Q8 format extension freezes three variants, three seeds and fifty requests per cell: 450 attempts, with zero transport errors. Settings are temperature 0.2, JSON output mode and the recorded context setting of 2048. The unchanged format-grid parser scores both quantizations.

For bullet, heading and reversal, exact placeholder counts are 42, 36 and 70 out of 150; invalid

	BFCL				JSONSchemaBench			
Model	Correct (%)	Err.	Def.	Placeh.	Valid (%)	Err.	Marker	Exact
MedGemma 4B Q8	99.67	1	0	0	66.67	35	17	0
MedGemma 4B Q4	100.00	0	0	0	63.00	31	19	0
Gemma 3n E4B	98.00	0	0	0	64.33	39	21	0
BioMistral 7B Q4	86.00	3	0	1	56.33	20	39	0
Meditron 7B Q4	7.67	8	220	29	25.33	20	177	58
Meditron 7B Q8	3.00	7	225	56	32.33	17	147	40
Meditron3 8B	99.00	0	0	0	68.00	11	51	11
Mistral Instruct v0.2	98.67	1	0	0	68.67	5	26	0
OpenBioLLM 8B	40.67	0	2	0	57.00	3	34	5
Aloe-Qwen 7B	100.00	0	0	0	77.33	3	25	0

Table 4: Public-subset audit, 100 instances at three seeds per surface. BFCL: correct function name (%), function-definition category (count), placeholder category (count). JSB: valid (%), broad schema-metadata marker (count), and exact full-schema copying (count). Err.: runtime-error attempts (count), retained as unsuccessful in each percentage. Metadata markers are fallible and overlap valid outputs. Exact matches compare parsed JSON objects, ignoring whitespace and key order.

Model	C	R	O	F	E	T
Gemma 3n E2B	0.96	0.00	0.00	0.00	0.10	0.00
Phi-4-mini	1.00	0.00	0.03	0.00	0.00	0.00
MedGemma 4B Q8	1.00	0.00	0.00	0.00	0.00	0.00
BioMistral 7B Q4	0.01	0.95	0.00	0.00	0.00	0.00
Meditron 7B Q4	0.15	0.00	0.00	0.00	0.00	0.00
Mistral Instruct v0.2	1.00	0.00	0.00	0.00	0.00	0.00
Llama-2 Chat	1.00	0.00	0.00	0.00	0.00	0.00
OpenBioLLM 8B	0.97	0.00	0.00	0.07	0.00	0.00
Aloe-Qwen 7B	0.83	0.17	0.00	0.18	0.00	0.00
Meditron 7B Q8	0.31	0.00	0.00	0.00	0.00	0.00
Meditron3 8B	0.87	0.00	0.00	0.22	0.00	0.00

Table 5: Strict ICU per-class recall; invalid actions remain errors. C: continue; R: recheck; O: oxygen/positioning; F: fluid/pressor; E: escalate; T: transfer. Support per seed is 50, 25, 25, 20, 20 and 10, respectively.

counts are 44, 36 and 75. Correct counts by seed are (21, 23, 19), (20, 21, 19) and (1, 6, 2). Reversal has 14, 43 and 18 valid outputs, respectively, all selecting `code_execution`.

The Q8 format extension used Ollama 0.20.7; historical records describe a 0.6.x stack. The earlier Q4 runs were not repeated under the later runtime. The July four-surface Q8 control has no recorded runtime version. These records therefore leave runtime effects unresolved in historical quantization comparisons. Q8 is still quantized; no FP16 control was run.

E Interventions and robustness

E.1 Reasoning field and schema constraint

Table 7 reports the joint addition of a reasoning field and a schema constraint. Because both change together, their individual effects cannot be separated. Meditron's 449 valid ICU responses all continue monitoring. Its wearable outputs instead comprise 305 continue, 93 posture, 80 break and 44 task-modification choices. MedGemma uses four ICU actions, yet its 130/450 correct decisions fall below the baseline 150/450.

E.2 Prompt and context changes

Three alternative ICU framings retain MedGemma's and Llama-2's all-continue profile. BioMistral stays recheck-dominant among valid outputs, but invalid counts rise from 5/450 to 241, 101 and 247. Meditron produces identical saved responses in all three conditions: 14/450 correct and 377 invalid. The cause of that equality is unresolved.

Few-shot ICU prompts add six examples, one per class. BioMistral has 445/450 invalid responses; the other six evaluated models score between 5.6% and 33.3%. Increasing Meditron's context setting to 4096 leaves its responses unchanged at 25/450 correct and 389 invalid. On the router, MedGemma falls from 144 to 138/150, Gemma E4B stays at 147/150, and Meditron falls from 48 to 25/150 with 95 invalid outputs.

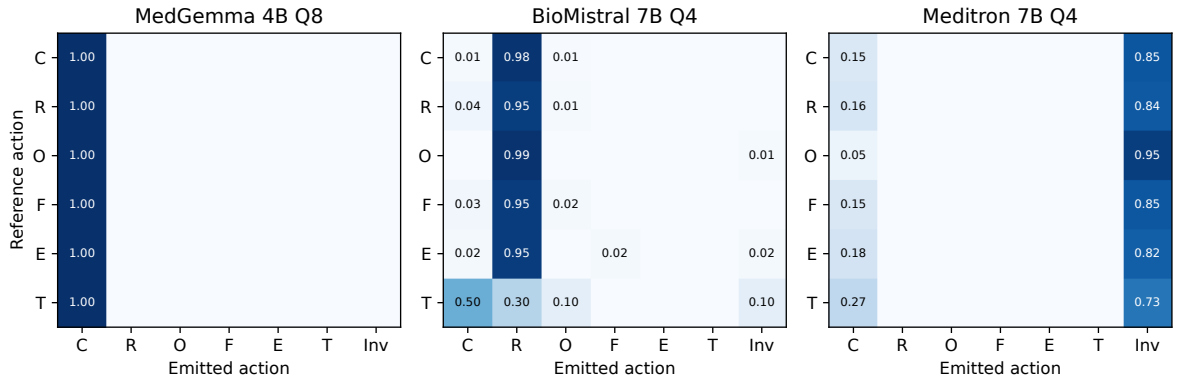


Figure 2: ICU confusion matrices. Rows are reference classes and sum to one; columns include invalid responses. The diagonal gives the recalls in Table 5. The invalid column shows which reference classes contribute to output failures.

Weights	Seed	Valid	Placeholder	Null	Dict-wrap	Non-JSON
Q4	0	66	0	22	62	0
Q4	1	0	132	14	0	4
Q4	2	0	150	0	0	0
Q8	0	140	0	6	3	1
Q8	1	0	0	110	0	40
Q8	2	0	150	0	0	0

Table 6: ICU response categories in the historical Meditron quantization control. Every row contains 150 attempts. All valid actions are continue monitoring. The seed numbers are archived seed indices; ICU requests use the recorded per-window seed derivation.

Removing decoder JSON mode, or both decoder mode and the prompt's JSON instruction, produces 450/450 invalid outputs in each reported cell under the original strict parser. A separate semantic extractor finds continue-monitoring words throughout MedGemma's decoder-only condition. That free-text endpoint does not satisfy the original output contract.

E.3 Transfer and additional perturbations

MIMIC contributes 423 attempts per model. MedGemma and the two historical comparators remain all-continue; BioMistral remains recheck-dominant. MedGemma scores 150/423 (35.5%), matching reference prevalence; BioMistral scores 75/423 (17.7%); Meditron scores 30/423 (7.1%) with 341 invalid outputs. The extraction reuses the authored hierarchy and its limitations concerning missing fluids, units and trends (Appendix A).

The historical router-format grid covers seven main-cohort models plus E4B, totaling 3,600 attempts; OpenBioLLM and Aloe are absent. The additional Q8 grid covers Meditron only. Temperature checks at 0.0, 0.5 and 0.7 yield 95.3–96.0% router accuracy for MedGemma and 98.0% for E4B. Extra router/BFCL seeds remain separate from the original three-seed summaries.

After function-name renaming, both tested models score 300/300 on BFCL. Property-renamed JSONSchemaBench gives 196/300 valid instances for MedGemma and 194/300 for E4B, compared with 200 and 193 before renaming. This perturbation leaves task structure and examples intact, so prior exposure remains untested.

F Uncertainty and threshold sensitivity

F.1 Resampling and paired comparisons

Accuracy intervals use 10,000 case-cluster bootstrap samples with random seed 20260926. Each sampled case retains all its seed responses. Residual dependence between cases from one patient and the effects of stratified sampling remain outside these intervals. Baselines use matched cases; wearable exposure differences are described in Appendix A.

For the baseline prompts, the descriptive 95% case-cluster bootstrap intervals for ICU accuracy

Task	Model	N	C_b	C_j	I_b	I_j	k_b	k_j
ICU	MedGemma	450	150	130	0	56	1	4
ICU	BioMistral	450	73	100	5	0	4	4
ICU	Meditron	450	22	150	384	1	1	1
Wearable	MedGemma	522	210	124	0	4	1	1
Wearable	BioMistral	522	142	106	0	0	3	2
Wearable	Meditron	522	70	159	340	0	1	4

Table 7: Baseline (b) and joint reasoning-field/schema condition (j). C is the correct count, I the invalid count and k the number of valid actions used. Both components change together; all attempts remain in the denominator.

Comparator pair	Δ (pp)	90% CI (pp)	TOST p
MedGemma / Gemma E4B	-2.00	[-5.35, 1.35]	0.0021
BioMistral / Mistral	-10.00	[-18.93, -1.07]	0.6456
OpenBioLLM / Llama-3	0.00	[-6.39, 6.39]	0.0205
Aloe / Qwen2.5	-0.67	[-5.59, 4.25]	0.0079
Med42-v2 / Llama-3	1.33	[-2.22, 4.89]	0.0014

Table 8: Exploratory router comparisons. Difference is the mean within-case difference in accuracy after averaging three seeds; $n = 50$ cases. Intervals are paired t intervals (90%). TOST p uses a post-hoc ± 8 percentage-point margin, without multiplicity adjustment. These are comparator contrasts, not isolated fine-tuning effects.

are 26.0–40.7% for MedGemma and 3.1–6.9% for Meditron Q4; their router intervals are 90.0–100.0% and 20.0–44.7%, respectively. These intervals use 150 unique ICU cases and 50 unique router requests, keeping each case's three seed responses together. The paired comparisons below use a separate 90% t interval for the difference between models.

For each router pair, we average seed correctness within each case and then compute the fifty paired differences. Table 8 gives 90% paired t intervals and TOST results (Schuirmann, 1987) using a post-hoc ± 8-point margin, with no multiplicity adjustment. The margin has no independently established operational meaning. MedGemma's interval fits eight points but not five; BioMistral trails Mistral by ten points. Different checkpoint histories further limit causal attribution.

F.2 Concentration thresholds

The sensitivity analysis varies the required criterion count from one through four and the entropy threshold from 0.5 through 1.5 bits. MedGemma's ICU/router contrast persists; Meditron's baseline/reversed router profiles differ, and Aloe's ICU flag can change with the threshold. Dominance conditions on valid actions. With none, entropy and dominance are undefined and the cell is an output failure. Diagnostic sensitivity and specificity remain unmeasured.

G Protocol history

Two local planning documents contain different hypothesis sets and lack independently verified registration timestamps. We use the initial May 17, 2026 document to describe the recorded history. Its comparator-isolation hypothesis fails because both historical comparators concentrate on continue. Twenty held-out cases were planned and seventeen realized; the protocol specified 512 output tokens and the ICU runner requested 400. Exact prompts were frozen after a smoke test. Cohort expansion, lenient parsing, cross-domain tests and most robustness checks were exploratory additions.

H Implementation and availability

The principal runs use Ollama/GGUF, temperature 0.2 and seed indices 0, 1 and 2. ICU seeds are derived per window, so these indices do not imply identical request seeds across tasks. Held-out runs use seed 0; sweeps and extra seeds are separate groups. Appendix D describes runtime records. Table 9 preserves the principal model tags; historical immutable digests are incomplete. Additional router pairs use Llama-3 Instruct, Qwen2.5 Instruct, Med42-v2 and Bio-Medical-Llama-3 (Meta, 2024; Qwen Team, 2024; Christophe et al., 2024; ContactDoctor, 2024). The run inventory records condition, model, observed counts and source identifiers.

Code, prompts, rule mappings, the authored router task and aggregate results are available at `github.com/abhi789/mva-audit`.

Model	Archived Ollama tag
MedGemma Q8	`medgemma1.5:4b-it-q8_0`
MedGemma Q4	`medgemma1.5:4b-it-q4_K_M`
BioMistral	`adrienbrault/biomistral-7b:Q4_K_M`
Meditron Q4 / Q8	`meditron:7b-q4_K_M` / `meditron:7b-q8_0`
Mistral comparator	`mistral:7b-instruct-v0.2-q4_K_M`
Llama-2 comparator	`llama2:7b-chat-q4_K_M`
Gemma comparators	`gemma3n:e2b` / `gemma3n:e4b`
Phi-4-mini	`phi4-mini:latest`
OpenBioLLM	`koesn/llama3-openbiollm-8b:Q4_K_M`
Aloe-Qwen	`hf.co/mradermacher/Qwen2.5-Aloe-Beta-7B-i1-GGUF:Q4_K_M`
Meditron3	`hf.co/mradermacher/Meditron3-8B-GGUF:Q4_K_M`

Table 9: Archived identifiers for the principal cohort and controls. Tags preserve the distinction between a model family, instruction tuning and quantization.

Task	Model	N	Correct	Invalid	k	H	d	M
Router	MedGemma Q8	150	144	0	6	2.58	0.180	0
Router	MedGemma Q4	150	147	0	6	2.58	0.180	0
Router	Gemma E4B	150	147	0	6	2.58	0.180	0
Router	Meditron Q4	150	48	68	5	1.90	0.378	1
Router	Meditron Q8	150	66	39	6	2.11	0.369	0
Router	Meditron3	150	135	0	6	2.55	0.240	0
ICU	MedGemma Q4	450	150	0	1	0.00	1.000	4
ICU	Meditron3	450	144	0	2	0.79	0.764	3

Table 10: Additional controls cited in the main Results. Entropy H (bits) and dominance d use valid actions; M counts concentration criteria.

Record-level outputs derived from eICU-CRD and MIMIC-IV are available on request to credentialed PhysioNet users, consistent with the source data use agreements.

I Additional controls

Table 10 collects additional controls used in the main Results. Meditron3's ICU continue counts are 62, 141 and 141 by seed; the remaining 88, 9 and 9 choices select fluid or vasopressor adjustment.

J Prompt excerpts

The following lines are excerpts from the evaluated prompts. The ICU confidence wording conflicts with its generator (Appendix A).

The ICU gate permits an action beyond continuing when the current deterioration signal has confidence at least 0.85; the current signal indicates deterioration with confidence at least 0.70 and MAP below 65 or SpO_2 below 92; deterioration persists for at least two windows with confidence at least 0.70 in each; or the trend is increasing with current deterioration confidence at least 0.70. Otherwise the prompt instructs the model to continue. These are the four conditions referred to as (a)–(d) in the excerpt.

ICU confidence: "confidence: 0.0-1.0, calibrated"; "conf=0.80" means this label is correct approximately 80% of the time.
ICU gate: "If none of (a)-(d) hold, default to continue_monitoring even if the classifier shows D."
Router priority: "For pure arithmetic on numbers the user provided, prefer 'calculator' over 'web_search' and over 'code_execution'."
Router output placeholder: `"tool": "<one of the 6 tool names exactly>"`