Rethinking MedAgentBench: A Framework for Fair Medical LLM Agent Evaluation

Ananya Mantravadi
Centific Global Solutions, Inc.
USA
ananya.mantravadi@centific.com

Prasanna Desikan
Centific Global Solutions, Inc.
USA
prasanna.desikan@centific.com

Abhishek Mukherji
Centific Global Solutions, Inc.
USA
abhishek.mukherji@centific.com

Abstract

MedAgentBench (MAB) v1 and v2 originally published by Stanford University are the primary benchmarks for evaluating large language models (LLMs) that query patient records through FHIR (Fast Healthcare Interoperability Resources) APIs and execute clinical orders. We audit both benchmarks and identify four evaluation failures that, together, allow a do-nothing agent to score 42% on v2 — before any clinical reasoning is demonstrated. First, *branch imbalance*: four v2 task types have 70–97% of instances on the no-action branch due to cohort composition. Second, the *silent-finish ceiling* — the measurable consequence: 41.7% of v2 tasks pass when an agent returns an empty result with no tool use, setting a floor that any reported score must be read against (only 5.3% on v1). Third, *undocumented format requirements*: graders for v1 task 5, v1 task 9, and v2 task 3 enforce format conventions absent from the task context causing systematic 0% pass rates on clinically correct responses. Fourth, a *wall-clock bug* in the v2-T1 grader anchors the 12-month CT follow-up window to real time rather than the scenario date; 4 of 30 patients who should require no new scan are misclassified as action-required when the benchmark is run in 2025–2026, breaking reproducibility. We construct **MedAgentBench-v3**, a corrected 508-task benchmark with fixed graders, a frozen timestamp, and 1:1 action/no-action balance. The do-nothing ceiling falls from 41.7% to 8.9%; on MAB-v3 frontier models score 50–79% overall (27–70 pp net above the do-nothing baseline), with substantial divergence between action-branch and no-action-branch pass rates revealing calibration differences invisible in aggregate scores.

Keywords

Clinical AI, Healthcare, LLM, Evaluations, Benchmarks

1 Introduction

Before trusting a benchmark score, we need to ask: does the benchmark actually measure what it claims? MedAgentBench [1] (MAB) is the leading benchmark for clinical LLM agent evaluation: it tasks a language model with querying electronic health records via FHIR APIs and writing clinical orders, then grades responses automatically against reference solutions. The benchmark covers 600 tasks across 20 clinical task types drawn from ≈100 real anonymized patients; the v2 extension [2] reports 91.16% for GPT-4.1 on the original 300 tasks and 88.67% on 300 new ones.

In this paper, we evaluate the evaluator. We audit both benchmarks from first principles — running a do-nothing agent, reading every grader function, verifying timestamps — and identify four failures:

(1) **Branch imbalance**: four v2 task types are 70–97% no-action-branch due to cohort composition, causing aggregate scores to conflate correct abstention with clinical reasoning.
(2) **Silent-finish ceiling**: the measurable consequence — 41.7% of v2 tasks pass when an agent returns an empty result with no queries.
(3) **Undocumented format requirements**: graders for v1-T5, v1-T9, and v2-T3 enforce bare-string route fields, dose formulas, and output list shapes absent from the task context.
(4) **Wall-clock bug**: the v2-T1 grader uses `datetime.now()` for its 12-month CT window, misclassifying 4 no-action instances as action-required in 2025–2026.

We contribute: (1) a systematic audit quantifying all four failures; (2) context patches restoring evaluability for three task types; (3) **MedAgentBench-v3**, a 508-task corrected benchmark; and (4) frontier baselines for six models under a uniform harness.

2 Background

MedAgentBench evaluates language models as *clinical FHIR agents*: given an instruction such as "If K < 3.5 mEq/L, order replacement potassium per the dosing protocol," the agent queries an EHR via structured API calls and, if warranted, writes a clinical order, then calls `finish(value)`. MAB v1 [1] contains 300 tasks across 10 types (30 instances each, ≈100 real anonymized patients), graded by `refsol.py` which checks both POST structure and `finish()` value. MAB v2 [2] adds 300 tasks across 10 richer types — QTc management, naloxone co-prescribing, thyroid protocols — using a separate grader (`new_refsol.py`) and ships a copy of the v1 grader (`medagentbenchevals/refsol.py`) for cross-version scoring. Tasks fall into three categories: *retrieval* (return a lab value or aggregate), *conditional action* (apply a threshold, act or abstain), and *format-knowledge* (always POST a specific code; 0% silent-finish).

3 Evaluation Issues Analysis

3.1 Branch Imbalance and Silent-Finish Ceiling

Four v2 task types (T4–T7) have 70–97% of instances on the no-action branch (Table 1): for most patients in the cohort, the relevant measurement falls within the normal range — catheter dwell time is acceptable, magnesium and QTc values are within threshold, TSH is not persistently elevated — so the correct response is to do nothing. This reflects cohort composition, not a code bug; yet the evaluation consequence is significant. A do-nothing agent that never queries the EHR scores 70–97% on these types, indistinguishable from a model that correctly retrieves lab values, confirms normal results, and abstains.

Table 1: v2 silent-finish rate by task type. Four task types (shaded) dominate the 41.7% aggregate ceiling.

Type	Clinical scenario	Act	No-act	SF rate
v2-T1	CT order (renal mass)	26	4	13%
v2-T2	DVT anticoagulation	28	2	7%
v2-T3	Heart rate average	30	0	0%
v2-T4	Urinary catheter dwell	9	21	70%
v2-T5	Renal mass / Mg replace	2	28	93%
v2-T6	Thyroid / levothyroxine	1	29	97%
v2-T7	Prolonged QTc	3	27	90%
v2-T8	Opioid + naloxone order	20	10	33%
v2-T9	Potassium replacement	28	2	7%
v2-T10	A1C + COVID vaccine	28	2	7%
Total		175	125	**41.7%**

The measurable consequence is the *silent-finish rate*: we call `finish([])` with no prior tool use against every instance and record which pass the grader. **v1:** 16/300 = 5.3%, all from v1-T5 (16 patients have no Mg observations in the EHR; the task instructs "do nothing if no measurement exists," so inaction is the correct response and the grader correctly credits it — this is a data-absence branch, not a normal-values branch). **v2:** 125/300 = 41.7%. Table 1 shows the per-type breakdown.

A do-nothing agent achieves a silent-finish rate of 41.7% on v2 and 5.3% on v1. The v2 paper reports 88.67% for GPT-4.1 on v2; the model-specific contribution above this floor is only ≈47 pp, not 88.67 pp. Because branch imbalance is a data distribution issue rather than a grader bug, the fix requires rebalancing the evaluation set — motivating the 1:1 action/no-action cap in MAB-v3 and establishing action-branch pass rate as the primary metric.

3.2 Undocumented Grader Format Requirements

Several MAB graders enforce format conventions never stated in the task context the agent receives. Models that use FHIR-compliant representations or reasonable output formats fail even when the clinical decision is correct.

v1-T5 – Magnesium Replacement: The task context documents the NDC code, dose tiers (1 g/2 g/4 g), and infusion rates — but says nothing about the format of the route field. The grader enforces:

```
assert payload["dosageInstruction"][0]["route"] == "IV"
```

This is a bare-string equality check. A FHIR-R4 compliant agent emits `"route": {"text": "IV"}` (a CodeableConcept), which fails this assertion even though the clinical action is correct.

v1-T9 – Potassium Replacement: The task context documents the NDC code, the dose formula verbally ("for every 0.1 mEq/L below 3.5, order 10 mEq"), and the LOINC code for the follow-up lab. Two format requirements go unmentioned: (1) the route field must be a bare string (the grader applies `.lower().strip()`, so capitalisation is flexible, but a FHIR CodeableConcept object fails), and (2) the follow-up lab ServiceRequest must include an `occurrenceDateTime` timestamped for the next morning. Additionally, the finish-value comparison is commented out in both distributed copies of `refsol.py`:

```
ref_sol = [last_K_value]
return True   # comparison removed; finish() value never checked
```

A model that places the correct orders but returns the wrong K value in `finish()` receives full credit.

v2-T3 – Heart Rate Average: The grader expects a 2-element list `[avg_6h, avg_12h]` with ±0.1 bpm tolerance. The public task context documents neither this output shape nor the convention for returning `-1` when no heart-rate observations exist for a window.

Effect: These mismatches account for systematic 0% pass rates on v1-T5 across most frontier models (Appendix D). MAB-v3 adds the missing format documentation to the task context for each affected type, so models know the expected output shape and field formats. The finish-value check in v1-T9 remains commented out in the grader; models that follow the augmented context and return the correct K value are acting correctly even though the grader does not verify it.

3.3 Wall-Clock Bug

Grader `new_refsol.py` in the MAB v2 computes the v2-T1 12-month CT follow-up window from real wall-clock time:

```
twelve_months_ago = datetime.now(timezone.utc) - timedelta(days
    =365)
```

This task's reference is dated November 2023. Evaluated in 2025–2026, every patient CT from 2023 appears older than 12 months regardless of its actual date relative to the scenario. Of the 30 v2-T1 patients, 4 had a CT within the past year relative to the frozen scenario date of 2023-11-13 and should require no new scan (no-action branch). Any evaluation run with the unpatched grader in 2025–2026 will misclassify all four as action-required. This is a reproducibility failure: a model evaluated in 2023 and again in 2025 receives different grades on identical episodes with no data change. MAB-v3 patches `task1` to call the pre-existing `extract_now_from_context()` helper, which reads the frozen scenario timestamp from `case_data` and restores deterministic evaluation. Notably, the file already defines `extract_now_from_context()`, which parses the frozen scenario timestamp from `case_data` — but `task1` never calls it. The fix is a single-line substitution.

4 MAB-v3: A Corrected Benchmark Construction

We apply four corrections, one per identified issue:

(1) **Context patches** (Issue 2): add the missing format documentation to the task context for v1-T5, v1-T9, and v2-T3; extend the v1-T7 grader to accept numeric values from natural-language or structured-dict responses (tolerance 0.5 mg/dL); document the expected output format for v1-T9.
(2) **Fixed timestamp** (Issue 3): replace `datetime.now()` in the v2-T1 grader with `2023-11-13T10:15:00+00:00`, restoring the no-action branch.
(3) **Silent-finish labelling** (Issue 1): run the null trace against all 600 tasks and label each instance `silent_pass = True/False`.

Table 2: Frontier model results on MAB-v3 (508 tasks).

Model	Overall	Net	v1	v2	Act	No-act
GPT-5.5	**78.7%**	**+69.8**	82.2%	73.8%	**77.3%**	93.3%
Gemini 3.1 Pro	78.1%	+69.2	**86.6%**	66.2%	76.5%	**95.6%**
GPT-4o (2024-11-20)	74.2%	+65.3	73.5%	**75.2%**	74.7%	68.9%
Llama 4 Maverick	62.2%	+53.3	58.7%	67.1%	61.3%	71.1%
Mistral Large	50.4%	+41.5	43.6%	60.0%	49.0%	64.4%
Claude Sonnet 4.6*	27.6%	+18.7	37.2%	13.8%	23.8%	66.7%
Do-nothing baseline	8.9%	0.0	5.2%	14.8%	0.0%	100.0%

Act = action-branch pass rate (463 instances); **No-act** = no-action-branch pass rate (45 instances). **Net** = Overall − 8.9 pp (do-nothing baseline). All errors counted as failures; see Appendix B. **Bold** = best in column. *Format non-compliance; see Section 5.

(4) **1:1 branch balance** (Issue 4): for each (corpus, task_type) bucket, retain all action instances and cap no-action at $\min(n_{\text{action}}, n_{\text{no-action}})$.

Composition: MAB-v3 contains **508 tasks**: 463 action-required and 45 no-action instances. The do-nothing ceiling is **8.9%** (45/508), down from 41.7% on v2. Figure 1(b) shows the branch-balance change per task type.

5 Discussion

We evaluate six frontier models on MAB-v3 (508 tasks) using the official MAB v1 harness (temperature 0.0, max 8 rounds, 2,048 max tokens, 9-function FHIR schema) via OpenRouter. Per-task results are filtered from runs on the 554-task pre-correction corpus; context-length failures were re-run with a response-truncation fix. For affected task types, models were re-run with the augmented context patches from MAB-v3 (Section 3.2). Per-task-type action-branch pass rates for all six models are in Appendix D; GPT-5.5 reaches 100% on eight task types and 0% on three, illustrating why aggregate scores mask capability gaps.

Action and no-action pass rates diverge substantially. GPT-5.5 leads overall at 78.7% (+69.8 net), closely followed by Gemini 3.1 Pro at 78.1% (+69.2 net). Gemini has the highest no-action pass (95.6%) while GPT-5.5 leads on action pass (77.3% vs. 76.5%). GPT-4o is the most uniformly calibrated: no-action pass (68.9%) closely matches action pass (74.7%), unlike top models whose no-action pass exceeds action pass by 15–20 pp. This divergence illustrates why aggregate pass rate conflates two distinct capabilities — identifying patients who need no intervention, and placing the correct clinical order when one is required.

Several task types remain effectively unsolved. v2-T5, v2-T6, and v2-T7 score 0% action pass across all five complete models. For v2-T7 (prolonged QTc, 3 action instances), the two-part requirement — ordering an ECG *and* discontinuing the offending medication — is the failure point; GPT-4o is the sole exception. v2-T3 (heart-rate average) also scores 0% across all models: the task requires computing averages over two time windows, and models consistently fail to correctly aggregate values from the FHIR response.

Format compliance is a prerequisite. Claude Sonnet 4.6 scores 27.6% (+18.7 net) despite strong clinical knowledge. The MAB harness requires bare action output; Claude 4.6 prepends reasoning to every call and 443/554 responses were rejected as invalid actions before any clinical decision was evaluated. All other five models complied without issue.

6 Conclusion

We identified four evaluation failures in MedAgentBench that, together, allow a model making no clinical decisions to score 42% on v2 tasks. MAB-v3 corrects all four — reducing the do-nothing ceiling from 41.7% to 8.9% — and reveals that frontier models score 50–79% overall on a fair evaluation, with substantial divergence between action-branch and no-action-branch pass rates invisible in aggregate scores. This audit was possible because MAB is open-source; we encourage all future clinical agent benchmarks to publish grader source code alongside scores.

Limitations. MAB draws from ≈100 unique patients reused across task types, with 30 instances per type; some types have very few action instances after 1:1 capping (v2-T6: 1, v2-T5: 2), making per-type estimates noisy. The 1:1 branch cap is a practical heuristic whose optimal ratio remains open. Task instructions are more explicit than real clinical practice — clinicians write terse, implicit orders and expect contextual inference — so benchmark scores may overestimate deployment readiness.

Future directions. MAB-v3's corrected 8.9% do-nothing ceiling makes it a viable RL training environment: the original 42% ceiling created a pathological incentive to learn inaction rather than clinical reasoning. Supervised finetuning on MAB-v3 action instances followed by RL refinement against the verifiable grader reward is a natural path toward small open-weight models that reliably execute conditional clinical protocols. Extending the harness to accept semantically equivalent outputs (addressing rigid format requirements) and scaling the patient cohort beyond 30 per type are priorities for future benchmark iterations.

References

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

[2] Eric Chen, Sam Postelnik, Kameron Black, Yixing Jiang, and Jonathan H Chen. Medagentbench v2: Improving medical llm agent design. In *Biocomputing 2026: Proceedings of the Pacific Symposium*, pages 354–371. World Scientific, 2025.

[3] Gyubok Lee, Elea Bach, Eric Yang, Tom Pollard, Alistair Johnson, Edward Choi, Jong Ha Lee, et al. Fhir-agentbench: Benchmarking llm agents for realistic interoperable ehr question answering. *arXiv preprint arXiv:2509.19319*, 2025.

[4] Suhana Bedi, Ryan Welch, Ethan Steinberg, Michael Wornow, Taeil Matthew Kim, Haroun Ahmed, Peter Sterling, Bravim Purohit, Qurat Akram, Angelic Acosta, et al. Healthadminbench: Evaluating computer-use agents on healthcare administration tasks. *arXiv preprint arXiv:2604.09937*, 2026.

A Related Work

MAB [1] introduced the FHIR tool-use evaluation setting we build on; MAB v2 [2] extended it with 300 tasks and memory-augmented evaluation. FHIR-AgentBench [3] grounds 2,931 questions in MIMIC-IV FHIR data and evaluates retrieval agents; it targets factual QA rather than clinical action decisions. HealthAdminBench [4] evaluates GUI-based agents on administrative EHR workflows and documents a subtask/task reliability gap that parallels the action/aggregate gap we quantify here.

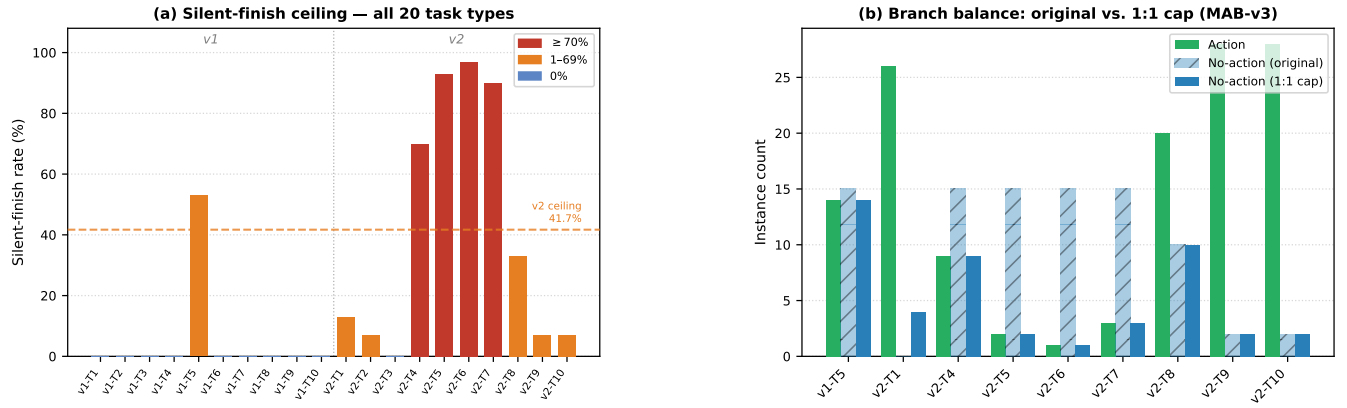

Figure 1: Branch imbalance and its measurable consequence. (a) Silent-finish rate across all 20 task types. The dashed line marks the 41.7% v2 aggregate ceiling. Four task types (red, ≥70%) are dominated by the no-action branch due to cohort composition, allowing a do-nothing agent to match weaker models. (b) Branch balance for conditional task types before (hatched) and after (solid) the 1:1 cap in MAB-v3, showing the reduction in no-action dominance.

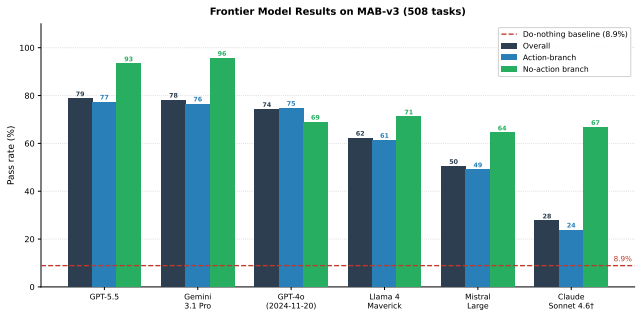

Figure 2: Frontier model results on MAB-v3 by branch type. The dashed line marks the 8.9% do-nothing ceiling. Top models show 15–20 pp higher no-action pass than action pass, illustrating why aggregate scores overstate clinical reasoning capability

B Evaluation Harness Configuration

All frontier baselines were evaluated using the official MedAgentBench v1 harness via OpenRouter with the following fixed configuration: temperature 0.0 (greedy decoding), max 8 tool calls per episode, 2,048 max tokens per response, 9-function FHIR schema (6 GET, 2 POST, 1 finish), and the official MAB v1 system prompt requiring bare action output only. Model identifiers: `google/gemini-3.1-pro-preview`, `openai/gpt-5.5`, `openai/gpt-4o-2024-11-20`, `meta-llama/llama-4-maverick`, `mistralai/mistral-large-2411`, `anthropic/claude-sonnet-4.6`. Context-length failures (HTTP 400/413) were re-run with FHIR bundle responses truncated at 8,000 characters; all other errors are counted as failures.

C Known Data Gaps in the MAB Cohort

Several observation types have incomplete coverage across the 30 patients per task type, reflecting genuine server-side absences: TSH (≈17/30 patients; valid no-action branches for v1-T6); A1C (≈17/30 patients); MG (≈25/30 patients; the 16 without Mg data account for all v1-T5 silent-finish instances); flu and COVID vaccination stored as `Procedure` codes (`90686` and `COVIDVACCINE`) rather than `Immunization` resources.

D Per-Task-Type Action Pass Rates

Table 3: Action-branch pass rate (%) by task type on MAB-v3. *n* = action instances. †Claude Sonnet 4.6: format failures.

Type (*n*)	Gem.	G5.5	4o	Ll4	Mis.	Son.†
v1-T1 (30)	100	93	7	37	93	100
v1-T2 (30)	100	100	100	80	0	0
v1-T3 (30)	100	100	100	13	0	0
v1-T4 (30)	100	100	83	87	63	60
v1-T5 (14)	0	0	21	0	0	0
v1-T6 (30)	100	100	100	100	100	0
v1-T7 (30)	40	40	33	23	17	0
v1-T8 (30)	100	100	100	100	87	97
v1-T9 (30)	93	97	47	43	67	97
v1-T10 (30)	40	3	47	3	0	0
v2-T1 (30)	87	87	100	87	87	13
v2-T2 (28)	25	25	32	39	39	0
v2-T3 (30)	0	0	0	0	0	0
v2-T4 (9)	56	100	100	100	33	0
v2-T5 (2)	0	0	0	0	0	0
v2-T6 (1)	0	0	0	0	0	0
v2-T7 (3)	0	0	100	0	0	0
v2-T8 (20)	100	100	85	10	55	0
v2-T9 (28)	75	93	100	100	100	0
v2-T10 (28)	89	100	100	100	100	0