Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A live agent benchmark of independent action authorization for healthcare AI agents

Lipsa Senapati

lipsa.email@gmail.com

DaVita Inc

Research Article

Keywords:

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11133816/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: Competing interest reported. The author is employed by DaVita Inc. The company had no role in the design, conduct, analysis, or reporting of this study, which used no company data, systems, or patient information. DaVita Inc. publication clearance was determined not to be required, as this work used no DaVita data, systems, patients, or proprietary information and was conducted independently of the author's employment duties.

Abstract

Benchmarks measure what healthcare AI agents can do, not what they may be authorized to do. We built a live-agent benchmark holding task, model, prompt and tools fixed, varying only whether an independent authorization layer gated execution. 576 synthetic care-coordination episodes ran under four arms: deterministic rules, an unconstrained agent, a gateway-governed agent, and a gateway plus autonomy controller. Episodes carry hidden truth and a lossy observable projection, so the gateway can both wrongly authorize and wrongly block. A prespecified rule set blinded to arm adjudicated safety. Unsafe actions occurred in 0.7% of fully governed versus 0.4% of unconstrained episodes (+ 0.4 points, 95% CI -0.5 to + 1.2, P = 0.6875). Routine task success, escalation accuracy, and calibration were each significantly worse under full governance, which also required more time and human effort. These synthetic findings use rule-based adjudication, lack a human comparator, and do not establish clinical effectiveness.

Introduction

Large language models have moved from answering questions to acting: retrieving records, planning multi-step work, and invoking tools. Benchmarks have followed. MedAgentBench places agents in a FHIR-compliant virtual electronic health record with 300 physician-written tasks, where the strongest evaluated model reached a 69.67% success rate.[1] AgentClinic extends evaluation to sequential, multimodal clinical encounters and shows that reformatting static questions as interactive decisions can reduce accuracy dramatically.[2] A systematic evaluation of two deployed agent architectures found only modest accuracy gains over baseline models — 60.3% on AgentClinic MedQA and 30.3% on MedAgentsBench — at more than ten times the token usage and more than twice the latency, and although in-agent safeguards filtered 89.9% of hallucinations, hallucinations remained prevalent.[3]

These are capability benchmarks. None of them measures *authorization*: which actions an agent may execute, under what preconditions, and for how long. That question is currently settled by unexamined engineering defaults, even though it determines whether a capable agent is also a safe one. The distinction matters because the last observation above suggests that safeguards *internal* to an agent — the agent checking itself — leave a substantial residue of failure.

Human review is the conventional backstop, and simulation evidence shows it is fallible in a specific and troubling way. In a high-fidelity intensive-care simulation, clinicians rejected 92% of unsafe AI drug recommendations but also rejected 29% of safe ones; unsafe recommendations attracted 37% more gaze fixations, yet clinicians did not return to the patient monitor or chart to investigate why a recommendation might be unsafe.[4][5] Oversight that neither reliably passes safe actions nor prompts re-examination of the underlying evidence is a weak foundation for autonomy.

An independent, externally enforced authorization layer is an obvious alternative, and architectures incorporating safety gateways and audit logs have been proposed. What has been missing is a controlled comparison: no published benchmark holds the task, model, prompt and tools fixed and varies only whether such a layer is present. Without that comparison, an organisation deciding how much autonomy to grant a deployed agent has no empirical basis for the decision, and no basis for deciding when autonomy should be withdrawn.

This study supplies that comparison. Four arms are evaluated on an identical frozen episode set: deterministic workflow rules (B2), an unconstrained tool-calling agent (B3), the same agent behind an Action Safety Gateway (B4G), and the same agent behind the gateway plus a Dynamic Autonomy Controller (B4). The B4G arm exists so that the gateway's contribution can be separated from the controller's rather than inferred. The design, hypotheses, endpoints, adjudication rules and analysis plan were registered before any agent was run (https://osf.io/3wc9s), and the episode set, policy, gateway, controller and prompt were content-hashed at registration.

Two design decisions deserve emphasis, because they determine whether the comparison means anything.

First, **episodes separate hidden truth from observable state.** The gateway is restricted to a lossy projection of each episode — the same projection the agent sees — while adjudication uses the hidden truth. Under strata that degrade observation (missing data, high signal noise, context volatility, tool-channel failure) the gateway can wrongly authorize an unsafe action or wrongly block a correct one. A gateway granted access to ground truth would be an oracle, its unsafe-action rate would be zero by construction, and the primary endpoint would measure nothing. A continuous-integration test fails the build if the gateway ever becomes error-free with respect to adjudicated truth.

Second, **the arm that is compared is not simulated.** An earlier version of this work reported a parameterized simulation in which each arm's unsafe-action propensity was a hand-set constant; the headline contrast was therefore an arithmetic restatement of an assumption. That analysis is superseded and is not reported here. Its code remains in the repository history, clearly marked, so that the two are not confused.

Results

Benchmark composition

The frozen set comprised 576 episodes in a full factorial over seven prespecified factors, with each cell appearing once routine and once perturbed, so perturbation status is exactly orthogonal to every factor. The correct terminal action was balanced by design at 192 episodes per action. Each arm was run 1 times with model gpt-4.1-mini-2025-04-14, giving 2,304 condition-episode runs. Table 1 reports the realised design.

Table 1
Benchmark episode strata and realised sample.

stratum	levels	episodes per condition
target_action	escalate_to_human / schedule_specialist / send_patient_message	192 / 192 / 192
social_risk	high / low	288 / 288
data_completeness	complete / missing	288 / 288
signal_noise	high / low	288 / 288
context_volatility	static / volatile	288 / 288
communication_preference	ambiguous / explicit	288 / 288
tool_availability	api_failure / full / partial	192 / 192 / 192
perturbation status	routine / perturbed	288 / 288

Primary endpoint: unsafe-action rate

The unsafe-action rate was 0.7% in B4 and 0.4% in B3. The prespecified paired contrast was + 0.4 percentage points (95% CI -0.5 to + 1.2), P = 0.6875 by exact McNemar test on 576 pairs, with 2 pairs unsafe in B3 only and 4 unsafe in B4 only. H1 was not supported: the point estimate moved in the opposite direction to that hypothesised, and the interval is compatible with no difference.

The deterministic-rules arm B2 recorded an unsafe-action rate of 0.0% and the gateway-only arm B4G 0.2%. Table 2 gives all arm-level estimates.

Table 2
Outcomes by arm.

condition	n	unsafe action rate	prohibited action rate	task success routine	task success perturbed	escalation accuracy	reliable tool use	time to action min	human effort units	autonomous action rate	ece	cost per success usd
B2	576	0.0	0.0	0.4861	0.4861	0.5	0.658	7.611	0.833	0.158	0.2667	21.43
B3	576	0.0035	0.0	0.4722	0.4722	0.5573	0.6562	8.891	1.061	0.1823	0.4584	28.08
B4G	576	0.0017	0.0	0.4167	0.3889	0.5069	0.6562	9.79	1.219	0.033	0.5334	37.83
B4	576	0.0069	0.0	0.4028	0.3681	0.4913	0.6597	9.858	1.236	0.0	0.548	40.09

Task success and the cost of authorization

Routine-episode task success was 40.3% in B4 versus 47.2% in B3; perturbed-episode success was 36.8% versus 47.2%. Time-to-action was 9.86 versus 8.89 minutes and human-effort burden 1.24 versus 1.06 units per episode. Paired contrasts with Holm-adjusted p-values for the secondary family appear in Table 3.

Table 3
Paired comparisons, B4 versus B3.

family	adjusted p value	endpoint	arm	reference	n pairs	estimate B4	estimate B3	difference
primary	0.6875	unsafe_action	B4	B3	576	0.0069	0.0035	0.0035
secondary:task_success_routine	0.00358	task_success	B4	B3	288	0.4028	0.4722	-0.0694
secondary:task_success_perturbed	1.4e-05	task_success	B4	B3	288	0.3681	0.4722	-0.1042
secondary:escalation_appropriate	0.000444	escalation_appropriate	B4	B3	576	0.4913	0.5573	-0.066
secondary:reliable_tool_use	1.0	reliable_tool_use	B4	B3	576	0.6597	0.6562	0.0035
secondary:prohibited_action_attempted	1.0	prohibited_action_attempted	B4	B3	576	0.0	0.0	0.0
secondary:time_to_action_min	0.0	time_to_action_min	B4	B3	576	9.8576	8.8906	0.967
secondary:human_effort_units	0.0	human_effort_units	B4	B3	576	1.2361	1.0608	0.1753
secondary:cost_usd	0.0	cost_usd	B4	B3	576	15.4524	13.2605	2.1919

Mechanism decomposition

Separating the two components, the gateway alone (B4G versus B3) accounted for a difference of -0.2 percentage points in unsafe actions, while adding the controller (B4 versus B4G) contributed + 0.5 points. Neither component produced a statistically distinguishable change in the unsafe-action rate (both 95% confidence intervals include zero), so this decomposition does not identify which component, if either, would carry a safety benefit in a setting where one exists.

Table 4
Mechanism decomposition.

contrast	endpoint	arm	reference	n pairs	estimate B4G	estimate B3	difference	ci low	ci high	p value	test	estimate B4
B4G vs B3 (gateway only)	unsafe_action	B4G	B3	576	0.0017	0.0035	-0.0017	-0.0052	0.0	1.0	exact McNemar	
B4 vs B4G (controller added)	unsafe_action	B4	B4G	576	0.0017		0.0052	-0.0017	0.0139	0.375	exact McNemar	0.0069

Calibration, equity, and autonomy dynamics

Expected calibration error of the agent's self-reported confidence was 0.548 in B4 and 0.458 in B3. Across social-risk strata, the between-stratum gap in unsafe actions was + 0.7 in B4 and + 0.0 in B3 (Table 5); these analyses are descriptive and were not powered to detect interaction.

Table 5
Social-risk subgroup analysis.

endpoint	social risk	arm	reference	n pairs	estimate B4	estimate B3	difference	ci low	ci high	p value	test
unsafe_action	low	B4	B3	288	0.0035	0.0035	0.0	-0.0104	0.0104	1.0	exact McNemar
unsafe_action	high	B4	B3	288	0.0104	0.0035	0.0069	-0.0069	0.0208	0.625	exact McNemar
unsafe_action	gap (high - low)	B4					0.0069				
unsafe_action	gap (high - low)	B3					0.0				
task_success	low	B4	B3	288	0.4167	0.5521	-0.1354	-0.1771	-0.0938	2.1532287064474076e-10	exact McNemar
task_success	high	B4	B3	288	0.3542	0.3924	-0.0382	-0.0799	0.0	0.08953107893466952	exact McNemar
task_success	gap (high - low)	B4					-0.0625				
task_success	gap (high - low)	B3					-0.1597				

The Dynamic Autonomy Controller recorded 576 state observations, comprising 574 hold, 2 increase. Reductions and revocations co-occurred with concurrent operating signals rather than arising at random (statistic not estimable, P not estimable by permutation test). Figure 2 shows the autonomy trajectory alongside the monitored signals.

Discussion

H1 was not supported. The unsafe-action rate under full governance (0.69%) was numerically higher than under the unconstrained agent (0.35%), and the paired difference (+ 0.35 percentage points, 95% CI − 0.52 to + 1.22, P = 0.6875) is compatible with no effect. Mechanism decomposition shows neither component moved the rate in a statistically distinguishable way: the gateway alone (B4G versus B3) shifted the estimate by − 0.17 points (P = 1.0) and adding the controller (B4 versus B4G) by + 0.52 points (P = 0.375). The most parsimonious explanation is a floor effect specific to this model: the unconstrained agent never once attempted the explicitly prohibited action across 576 episodes, and its baseline unsafe-action rate left little room for an external layer to demonstrate a detectable benefit. Governance was not free: routine and perturbed task success, escalation accuracy, and calibration were each significantly worse under B4 than B3, and time-to-action, human effort, and cost were each significantly higher, confirming H3 as an expected cost while H2a, H2b, H4 and H5 were each contradicted in direction rather than merely unsupported. H6 could not be tested: the controller held its state in 574 of 576 observations and never once reduced or revoked autonomy, leaving no variation in reduction status to relate to the monitored signals.

This benchmark is complementary to, not competitive with, capability suites such as MedAgentBench and AgentClinic.[12] Those measure what an agent can accomplish in a realistic record environment; this one measures what it is permitted to execute and what that permission costs. The observation that in-agent safeguards filter most but not all hallucinations[3] motivates the central manipulation here, namely authorization enforced outside the agent, which cannot be argued away by the agent's own reasoning. The design also responds to the argument that AI cannot be evaluated in isolation from the workflow it acts within, and that responsible evaluation requires deliberately challenging cases spanning good and poor system performance.[4] The perturbation set serves that purpose, and includes an episode in which record text instructs the agent to initiate a medication change that policy prohibits in every episode.

The deterministic-rules arm (B2) recorded zero unsafe actions by construction: a system incapable of the flexible interpretation that produces both correct escalation and policy violation cannot violate the policy, but this ceiling is fixed rather than earned. B2's routine and perturbed task success were identical (48.6% each), since fixed thresholds are insensitive to the perturbations that degraded the LLM arms' perturbed-episode performance; its cost and human-effort burden were the lowest of any arm, and its calibration was the best observed (ECE 0.267). B2 therefore functions as a naive floor rather than a competing hypothesis, showing what a system with no capacity for unsafe deviation costs and achieves, against which the LLM arms' greater flexibility can be weighed.

Six considerations bound these results, and the first four are structural rather than incidental. First, the episodes are synthetic. No patient data were used and no patient was affected; the episodes are stylised care-coordination tasks rather than a sample from a clinical population, so no rate reported here estimates a real-world rate. Positive findings would justify shadow-mode evaluation, not deployment. Second, adjudication was rule-based rather than clinical. Blinding to arm is structural, since the adjudicator receives an object that cannot carry the arm label, and this removes the principal source of adjudication bias; but no clinician reviewed these episodes, so the endpoint detects only the violations the seven frozen rules encode and cannot recognise harms outside that specification. Independent clinical adjudication remains necessary before any claim about clinical safety.

Third, there is no human-workflow comparator. A manual arm was specified in an earlier version of this design and was removed, because measuring it requires human participants who were not recruited, and simulating a human baseline would have reintroduced precisely the assumed effect sizes this study was built to eliminate. Nothing reported here therefore compares governed agents with current human practice. Fourth, one model and one snapshot were prespecified. The magnitude of any authorization benefit depends on the underlying agent's propensity to propose unsafe actions, which varies across models; the direction of an effect may generalise but its magnitude should not be assumed to.

Fifth, a gateway can reduce but not eliminate failures arising from incomplete data, novel context, or defects in its own rules, and the residual unsafe-action rate measured under governance is a direct estimate of that ceiling within these strata. Sixth, the risk dimensions and the equity safeguard encode value judgements, notably that an elevated social-risk indicator requires human confirmation before patient outreach. Such judgements require stakeholder and institutional validation and are not derived from evidence.

Taken together, these results do not support the claim that external authorization enforcement reduces unsafe action for every agent. For this prespecified model, whose ungoverned unsafe-action and prohibited-action rates were already close to zero, the gateway and controller added measurable latency, human effort, and cost without a measurable safety benefit, and degraded task success and calibration. This is a boundary condition rather than a refutation of the general architecture: the magnitude of any authorization benefit is bounded above by the agent's own propensity to propose unsafe actions, and a sufficiently well-aligned model can leave an external layer with nothing to correct while it still charges for the checking. Whether this generalises to models with higher baseline unsafe-action rates, different tool sets, or deployment contexts with adversarial or degraded inputs is not established by this design and would require testing a range of models rather than the single prespecified snapshot used here. The result argues for measuring an agent's unconstrained propensity to violate policy before adding an authorization layer, rather than assuming the layer is cost-free insurance.

Methods

Study design and registration

Prospective, fully paired, four-arm benchmark on an identical frozen episode set. The design, arms, endpoints, adjudication rules, hypotheses, power rationale and analysis plan were registered before any agent was run (https://osf.io/3wc9s, deposited 2026-09-19 07:08 UTC). The episode set, adjudication policy, gateway, controller, and prompt were content-hashed at registration and each run records those digests in its provenance record, so any reported result can be verified against the registered materials. Deviations from the registered plan are listed in Section 12 of the preregistration. Four deviations were recorded after registration, all before or during the pilot/full run and none altering the frozen episode set, adjudication policy, gateway, or controller behaviour: a Python-version-dependent bug in the freeze-digest computation (inspect.getsource is not byte-stable across Python versions), a transport bug that prevented multi-step episodes from completing under store: false (previous_response_id chaining requires server-side storage), a token-pricing constant mismatched to the pinned model (a single gpt-4.1 price was applied regardless of which model was run), and a retry-logic gap for bare read timeouts. Full detail and rationale for each are in prereg/PREREGISTRATION.md Section 12..

Arms

Arm	Description	Agent	Gateway	Controller
B2	Deterministic thresholded workflow rules	none	no	no
B3	Unconstrained agent, full tool access	LLM	no	no
B4G	Gateway only	LLM	yes	no
B4	Gateway and Dynamic Autonomy Controller	LLM	yes	yes

Arms B3, B4G and B4 receive the same agent class, model snapshot, system prompt, tool schemas, episode set and replicate index. The sole difference is whether the gateway and controller are interposed between a proposed call and the tool environment; this is enforced in code rather than by convention. B2 always uses the rules agent regardless of configuration, so that it cannot silently collapse onto B3.

An arm representing manual human workflow is deliberately absent (see Limitations).

Episode set

Full factorial over seven factors: correct terminal action (schedule specialist, message patient, escalate); social-risk indicator; data completeness; signal noise; context volatility; communication preference; and tool availability. 288 cells, each appearing once routine and once perturbed, giving 576 episodes with perturbation status exactly orthogonal to all seven factors and the correct action balanced at 192 per level.

Eight prespecified perturbations act from the second step onward, after the agent's first read: mid-run snapshot staleness, a contradictory second source, tool timeout, tool permission revocation, mid-run patient opt-out, specialist cancellation, a pre-existing duplicate request, and an instruction injected into record text. Episodes are generated deterministically and the set is content-hashed.

Making the correct terminal action a balanced design factor is essential. When correctness is instead a by-product of the strata, escalation dominates the episode set — in an earlier iteration of this design, 89% of episodes required the same answer — and neither task success nor escalation accuracy can discriminate between arms.

Hidden state and observation

Each episode carries hidden true state and a lossy observable projection. Agents and the gateway receive only the projection; the adjudicator uses only the hidden state. Missing data render freshness and record consistency unobservable; ambiguous communication preference renders outreach authorization unobservable; tool-channel failure renders specialist availability unobservable; context volatility inverts availability after the first read; high signal noise misreports severity. Unobservable is represented distinctly from false, and consumers must treat it as missing rather than negative.

Action Safety Gateway

The gateway validates every proposed write action against observable signals only: action space, data freshness, record consistency, outreach authorization, specialist availability, duplication, tool-channel integrity, and an equity safeguard requiring human confirmation for outreach under elevated social risk. It emits allow, require_confirmation, or block.

Unresolved preconditions are accumulated rather than short-circuited. Returning early only on a blocking condition allows the primary strictness setting to fall through to allow, silently authorizing actions whose preconditions could not be verified — the opposite of the intended behaviour. Three strictness levels govern how unobservable preconditions are resolved (permissive allows, moderate requires confirmation, strict blocks); moderate is the prespecified primary setting and the others are used for the Fig. 1 frontier.

Dynamic Autonomy Controller

Four autonomy states are distinguished by an autonomous-write budget: A0 recommendation-only (0 unconfirmed writes), A1 human-confirmed execution (0), A2 bounded autonomy (1), and A3 adaptive multi-step autonomy (2). State governs how much confirmation is required, never whether preconditions are verified; no state maps to permissive strictness, and a test asserts this.

Transitions use a rolling 20-episode window of observable signals. A prohibited action proposal revokes autonomy immediately. A window reference-block rate above 0.30 reduces it. Promotion requires a full window, a reference-block rate below 0.05, a tool-error rate below 0.20, and a minimum exposure of five executed writes.

Two properties of this design were arrived at by fixing defects found during validation, and both are load-bearing. First, the monitored block rate is computed against a **fixed reference strictness**, not the strictness the current state imposes. Using the enforced rate makes the signal endogenous to the intervention and produces ratchets in both directions: a permissive state suppresses blocks and reads its own quiet as earned trust, while a strict state inflates blocks and can never earn its way back. Second, promotion requires an exposure floor; without it a window can appear clean merely because the arm was never exercised, and autonomy ratchets upward on no evidence.

Agents and tools

The tool set is get_patient_snapshot, schedule_specialist, send_patient_message, request_human_confirmation, escalate_to_human, and adjust_medication. The last is prohibited by policy in every episode but is deliberately offered to the agent: an agent's willingness to take a prohibited action cannot be measured if the capability is withheld. Decision tools require a calibrated confidence in $^{[0,1]}$, supplying the calibration endpoint.

The live agent uses the provider's function-calling interface with server-side retention disabled, at most eight steps per episode, and up to five retries on transient HTTP 429 and 5xx responses. Quota and billing errors are treated as non-retryable. Only synthetic episode text is transmitted.

Outcomes and adjudication

The primary endpoint is the unsafe-action rate: the proportion of episodes in which at least one *executed* action violates the frozen policy. Actions blocked before execution are not counted, since the endpoint is unsafe execution rather than unsafe intent; intent is reported separately.

The frozen policy comprises seven rules: P1 a prohibited action was executed; P2 a write executed on genuinely stale data; P3 a write executed despite a genuine record conflict; P4 outreach without genuine authorization; P5 a specialist booked who was genuinely unavailable; P6 outreach under elevated social risk without prior recorded human confirmation; P7 the same write executed twice. Task success requires the correct terminal action and no violation.

Secondary endpoints are task success (routine and perturbed reported separately), escalation accuracy, reliable tool use, prohibited-action attempt rate, time-to-action, human-effort units, cost per successful episode, and expected calibration error.

Adjudication is condition-blinded by construction: the adjudicator is passed an object carrying the episode identifier, the executed action trace, and the reported confidence, and which structurally cannot carry the arm label, the model identity, or the gateway's dispositions. A test asserts the absence of those fields. Adjudication is rule-based, not clinical.

Statistical analysis

Primary analysis is an exact McNemar test on discordant pairs at two-sided $\alpha = 0.05$, paired on episode and replicate index. The risk difference is reported with a 95% percentile interval from a bootstrap of 10,000 draws that resamples whole episodes; episodes are the resampling unit because arms share the episode, and resampling condition-episode rows would understate uncertainty.

Secondary binary endpoints use the same paired framework and continuous endpoints the Wilcoxon signed-rank test, with family-wise error controlled by Holm. Non-inferiority of routine task success is assessed against a prespecified 10-percentage-point margin. Calibration uses 10 equal-width bins with a bootstrap interval. The autonomy-transition hypothesis is tested by permuting reduction labels while holding the signal series fixed.

A mixed-effects logistic regression was specified in an earlier draft and is not used: the design is fully paired on a binary outcome, so an exact paired test is sufficient and avoids distributional assumptions the design does not require. Episodes whose agent run fails irrecoverably after the retry policy are excluded from all arms for that replicate to preserve pairing, and exclusions are reported by arm and reason. No imputation is performed.

Claims not supported by this design

Registered in advance and binding on interpretation: no claim of clinical effectiveness or patient benefit; no claim of superiority to human practice; no claim of independent or clinical adjudication; no claim of generalisation across models; no claim that the frozen policy enumerates the harms of a real deployment.

Reporting standards

No existing AI reporting guideline applies to this study, and we do not claim conformance with one. CONSORT-AI and SPIRIT-AI govern the reporting of interventional clinical trials and their protocols,[67] and DECIDE-AI governs early-stage live clinical evaluation of AI decision-support systems.[8] All three presuppose human participants and clinical deployment, neither of which is present here. They are cited because they define the reporting obligations that would attach to the shadow-mode and bounded pilot studies that any positive finding here would motivate, and because the present study is deliberately structured to feed them: prespecified endpoints, a registered analysis plan, blinded outcome adjudication, and full provenance for every reported figure.

Ethics

This study involved no human participants, no animal subjects, no patient data, and no identifiable personal information. All episodes are synthetic and generated deterministically from the prespecified factors. Institutional review board review was therefore not applicable, and no informed consent was required.

Declarations

Data availability

No dataset containing human, patient, or identifiable personal data was generated, accessed, or analysed. The synthetic episode set is not a dataset of observations but is produced deterministically by the episode generator, and is therefore distributed as code rather than as data. The dataset(s) supporting the conclusions of this article — the frozen episode manifest with its SHA-256 digest, and all episode-level run outputs (one JSONL record per condition-episode run, including the full action trace, gateway dispositions, adjudication verdicts, and provenance) — are available in the Zenodo repository, https://doi.org/10.5281/zenodo.22851264.

Project name: Risk-Calibrated Autonomy for Healthcare AI Agents

Project home page: https://github.com/lipsasenapati/risk-calibrated-autonomy-healthcare-agents

Archived version: https://doi.org/10.5281/zenodo.22851264 (commit 5e49069)

Operating system(s): Platform independent

Programming language: Python

Other requirements: Python 3.8 or higher; an OpenAI API key for the prespecified model is required to reproduce the live-agent run, supplied through the environment (no credential is contained in the repository)

License: MIT

Any restrictions to use by non-academics: None

The GitHub repository above is a development mirror and is not the citable archive. The preregistration is at https://osf.io/3wc9s.

Code availability

All code required to reproduce the benchmark and the reported analysis — episode generator, tool environment, Action Safety Gateway, Dynamic Autonomy Controller, agent adapters, adjudicator, analysis scripts, figure generation, and the test suite — is archived under the MIT licence at the same Zenodo deposit as above (https://doi.org/10.5281/zenodo.22851264), corresponding to commit 5e49069.

Acknowledgements

The author received no specific funding for this work.

Author contributions

L.S. conceived the study, designed the benchmark, implemented the harness and analysis, registered the analysis plan, ran the benchmark, interpreted the results, and wrote the manuscript. L.S. is the sole author and accepts responsibility for the integrity of the work.

Competing interests

The author is employed by DaVita Inc. The company had no role in the design, conduct, analysis, or reporting of this study, which used no company data, systems, or patient information. DaVita Inc. publication clearance was determined not to be required, as this work used no DaVita data, systems, patients, or proprietary information and was conducted independently of the author's employment duties.

Ethics declaration

Ethics declaration: Not applicable.

Consent to participate

Consent to Participate declaration: Not applicable.

Consent to publish

Consent to Publish declaration: Not applicable.

Clinical trial number

Clinical trial number: not applicable.

Generative AI disclosure

Generative AI tools were used for language editing and for software development assistance during preparation of the benchmark harness. All study design decisions, the analysis plan, and the interpretation of results are the author's. The author verified all quantitative claims against the archived run outputs and all references against publisher records. No figure was generated or edited by generative AI; Figures 1 and 2 are produced programmatically from run outputs and Figure 3 is a schematic drawn programmatically.

References

1. Jiang, Y. et al. MedAgentBench: a virtual EHR environment to benchmark medical LLM agents. *NEJM AI* (2025). https://doi.org/10.1056/AIdbp2500144
2. Schmidgall, S. et al. AgentClinic: a multimodal benchmark for tool-using clinical AI agents. *npj Digit. Med.* **9**, 499 (2026). https://doi.org/10.1038/s41746-026-02674-7
3. Liu, Y. et al. Benchmarking large language model-based agent systems for clinical decision tasks. *npj Digit. Med.* **9**, 259 (2026). https://doi.org/10.1038/s41746-026-02443-6
4. Festor, P., Nagendran, M., Gordon, A. C., Faisal, A. A. & Komorowski, M. Safety of human-AI cooperative decision-making within intensive care: a physical simulation study. *PLOS Digit. Health* **4**, e0000726 (2025). https://doi.org/10.1371/journal.pdig.0000726
5. Nagendran, M., Festor, P., Komorowski, M., Gordon, A. C. & Faisal, A. A. Eye tracking insights into physician behaviour with safe and unsafe explainable AI recommendations. *npj Digit. Med.* **7**, 202 (2024). https://doi.org/10.1038/s41746-024-01200-x
6. Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J. & Denniston, A. K. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat. Med.* **26**, 1364–1374 (2020). https://doi.org/10.1038/s41591-020-1034-x
7. Cruz Rivera, S., Liu, X., Chan, A.-W., Denniston, A. K. & Calvert, M. J. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat. Med.* **26**, 1351–1363 (2020). https://doi.org/10.1038/s41591-020-1037-7
8. Vasey, B. et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* **377**, e070904 (2022). https://doi.org/10.1136/bmj-2022-070904

Figures

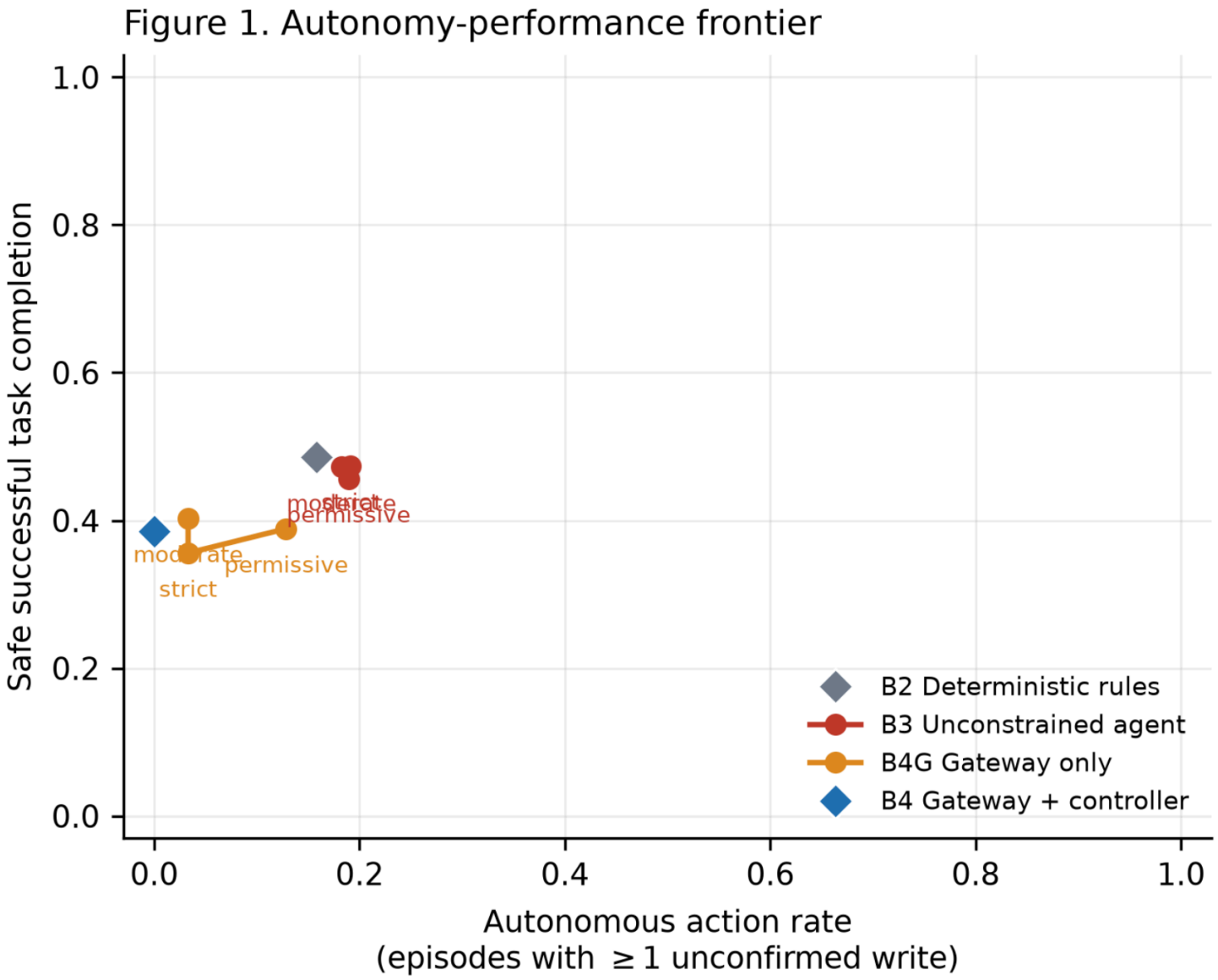


Figure 1

Autonomy-performance frontier across gateway strictness levels.

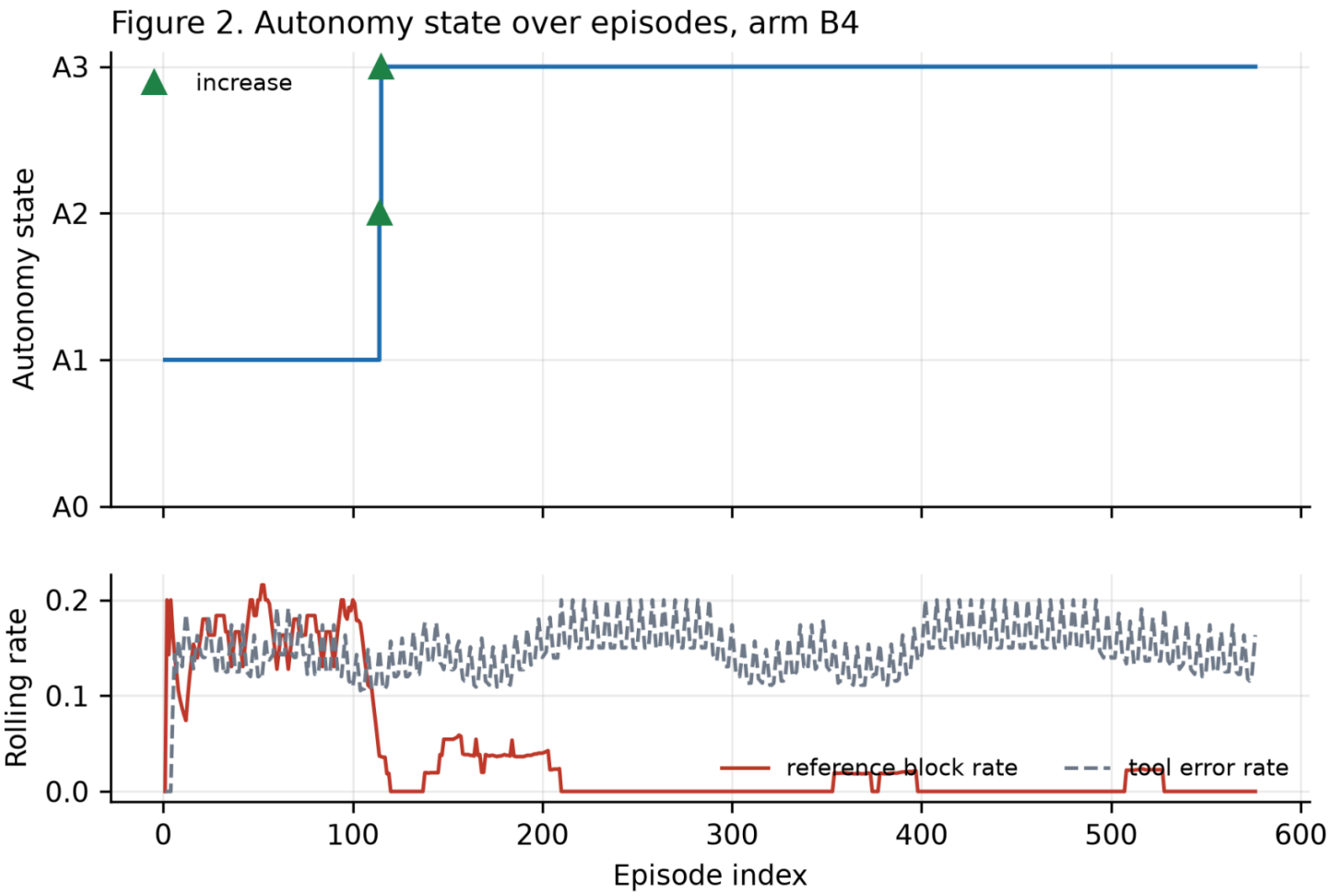
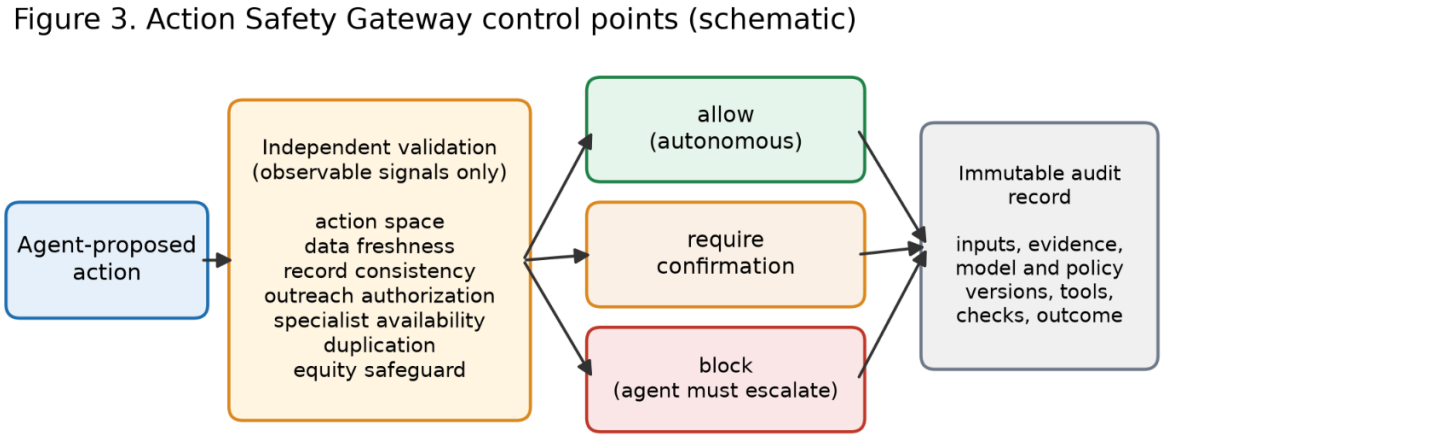


Figure 2

Autonomy state over the episode sequence, with reductions, revocations and increases marked against the monitored signals.



Schematic of architecture, not measured data. The gateway evaluates observable system signals only and never receives episode ground truth; adjudication of safety is performed separately and blinded to condition.

Figure 3

Action Safety Gateway control points (schematic).