Seven Months of Progress on Simulated Healthcare Administration Tasks

A comparison of two generations of computer-use agents

Zach McCormick

Novoflow

November 2026

Abstract. In April 2026, Bedi et al. introduced HealthAdminBench, a benchmark of 135 simulated healthcare administration tasks. Using screenshots and general portal guidance, GPT-5.4 completed eight tasks and Claude Opus 4.6 completed 20. Seven months later, we evaluated their successors in the same interaction setting. GPT-6.2 Sol completed 54 tasks (40.0%) and Claude Opus 5.5 completed 66 (48.9%). Each model family gained 46 completed tasks, or 34.1 percentage points, under strict scoring. Both newer models also exceeded their predecessors' stronger results with vendor-native computer-use tools. Even their smaller siblings completed far more tasks than the April models: Claude Haiku 5.5 completed 53 (39.3%) and GPT-6 Luna 38 (28.1%), against 20 and eight. They cost 95% and 90% less to run than Opus 5.5 and Sol, respectively. The gains persisted across all three workflows. A separate analysis of deterministic checks found gains even with language-model judgments set aside. Changes to the harness, portals and judge limit how precisely these results isolate model capability. We also reviewed every failed Sol and Opus run: 18% contained a clear agent error, and 50% failed only on scoring behavior or unstated requirements.

Task completion in April and November

Obtaining a prior authorization can require reading an order in an electronic health record (EHR), entering the request in a payer portal and copying the confirmation back to the patient's chart. Bedi et al. turned work of this kind into HealthAdminBench, a set of 135 tasks spanning prior authorization, denials and medical equipment orders [1]. A task passes only if the agent satisfies every check. Table 1 compares the April paper's main screenshot results with our November runs, both under the benchmark's original scoring.

Table 1. Strict task completion, out of 135. Both evaluations supplied the task description and portal guidance. April included two ways of operating the interface. November used the benchmark's standard screenshot agent.

Evaluation and agent configuration	OpenAI	Anthropic
April, standard agent	8 (5.9%)	20 (14.8%)
April, vendor-native computer-use loop	36 (26.7%)	49 (36.3%)
November, standard agent	**54 (40.0%)**	**66 (48.9%)**
Gain over April standard	+46 (+34.1 pp)	+46 (+34.1 pp)
Gain over April native	+18 (+13.3 pp)	+17 (+12.6 pp)

The standard-agent rows give the closest comparison for each model family. The April paper's headline of 36.3% came from Opus 4.6 using Anthropic's computer-use tool; both November models exceeded it using ordinary screenshot-based interaction. All comparisons here concern the screenshot and portal-guidance setting. The original study also evaluated other inputs and prompts [1].

Where the gains appeared

Both model families completed more tasks in every workflow. Prior authorization accounted for 26 of the 46 additional completions for OpenAI and 19 for Anthropic. The largest percentage-point gain for Opus was in durable medical equipment, where its earlier standard agent failed all 15 tasks and the newer agent completed 12. Appeals and denials had the lowest completion rate for both newer models. Sol completed 13 of 60 and Opus completed 25.

Table 2. Strict passes by workflow, using the standard screenshot agent at both dates. Each column reports complete tasks. Historical counts come from the released April run records [2] and reproduce the original paper's Table 3.

		OpenAI		Anthropic	
Workflow	Tasks	GPT-5.4	6.2 Sol	Opus 4.6	Opus 5.5
Prior authorization	60	4	30	10	29
Appeals and denials	60	2	13	10	25
Durable medical equipment	15	2	11	0	12

In `emr-hard-3`, whose definition was the same in April and November, the agent had to read a biologic dosing regimen, calculate the billing units for six months, submit the request to the payer, record the authorization number in the EHR and clear the referral. The rubric expected 26 units. GPT-5.4 passed three of the 24 checks and Opus 4.6 passed six. Both newer models passed all 24, including the calculation and the writes in both systems [2, 3].

Deterministic checks only

To rule out a more lenient judge, we compared only the deterministic checks on the 96 tasks whose definitions were identical at both dates. These checks inspect recorded actions and portal state. The April export gives total checks passed per task across both check types. If an agent's total passes fell below the number of deterministic checks, it must have failed at least one deterministic check. Treating *every other April task* as a deterministic pass gives a deliberately generous upper bound. GPT-5.4 could have passed every deterministic check on at most 30 tasks, versus 63 observed for Sol. Opus 4.6 could have done so on at most 39 tasks, versus 53 for Opus 5.5. This comparison bypasses both text judges. The portal and harness versions still differ.

Review of failed checks

The newer agents passed 91.3% (Sol) and 92.3% (Opus) of individual v3 checks, pooled across tasks. Yet both completed fewer than half of the tasks under strict scoring. To interpret this gap, we examined all 278 failed checks across the 150 failed Sol and Opus runs. A clear agent error appeared in 27 runs (18%). In 75 (50%), we attributed every failed check to scoring behavior or an unnecessary, unstated requirement. In the other 48 (32%), the available evidence supported more than one judgment. These figures describe failed checks in the November Sol and Opus runs only.

Clear agent error. In `denial-easy-6`, both agents reported finding no authorization. A link on the patient's name opened a record showing that an authorization had expired ten days before service. Neither agent opened it. Staff would be told that no authorization exists, when the real question is whether the payer will honor an expired one.

Unstated requirement. In `denial-medium-2`, Sol appealed the two denied claim lines, distinguished them from the two paid lines and documented the confirmation. The note failed three checks because it omitted the denied amount, a diagnosis code and the appeal deadline. The task specified what the note should contain and did not request those three details. If a workflow needs those details, the task should ask for them.

Debatable decision. In `denial-hard-8`, the payer's claim and member identifiers matched the EHR, but the date of birth, provider and facility differed. Sol held the upload and documented the conflict. Opus filed and flagged it. The written goal called for an appeal and gave no rule for resolving this identity discrepancy. Four Sol checks failed as a consequence of one decision.

Table 3. Sensitivity of November completion to review judgments. Each row keeps the same 135 recorded trajectories per model. The historical comparisons in Tables 1 and 2 use only the first rule.

Scoring rule	GPT-6.2 Sol	Opus 5.5
Upstream strict	54 (40.0%)	66 (48.9%)
Excuse hidden interface behavior, key and judge errors	66 (48.9%)	81 (60.0%)
Also excuse unnecessary, unstated requirements	100 (74.1%)	95 (70.4%)
Also accept every debatable decision	124 (91.9%)	119 (88.1%)

Whether unstated requirements count decides which model ranks first: Opus leads under strict scoring, and Sol leads (100 to 95) once they are excused. Excusing checks can also hide real failures. In `emr-hard-16`, Sol saved the authorization number but interpreted a disabled button as confirmation that the referral had cleared. The queue still contained it. Requiring actual clearance where that check failed reduces the third row to 92 Sol passes and 88 Opus passes. The comparison across generations uses the original strict scores throughout.

Smaller models

Each vendor also sells a much cheaper model in the same generation. We ran Claude Haiku 5.5 and GPT-6 Luna once on each v3 task with the same harness, settings and judge as the larger models. Their list prices are about one fortieth of Opus 5.5's and one twentieth of Sol's per token. Table 4 places them beside the November results above.

Table 4. The four November agents on the same 135 v3 tasks, one run per task. Agent cost uses list API prices and excludes the judge. A task passes every deterministic check when it satisfies all checks evaluated from recorded actions and portal state, whatever the text rubrics found.

	OpenAI		Anthropic	
	6.2 Sol	6 Luna	Opus 5.5	Haiku 5.5
Strict completions	54 (40.0%)	38 (28.1%)	66 (48.9%)	53 (39.3%)
Prior authorization (60)	30	24	29	27
Appeals and denials (60)	13	9	25	15
Durable medical equipment (15)	11	5	12	11
Passed every deterministic check	94	75	85	81
Individual checks passed	91.3%	79.9%	92.3%	83.1%
Text rubrics passed	82.2%	64.6%	87.6%	74.9%
Agent cost, all 135 tasks	$87.17	$8.51	$149.18	$6.91
Cost per task	$0.65	$0.06	$1.11	$0.05

Haiku 5.5 completed one task fewer than Sol and four more than the April paper's headline result, which came from Opus 4.6 with Anthropic's computer-use tool. It matched that configuration in appeals and in equipment orders and exceeded it in prior authorization, 27 to 23. Opus 5.5 completed 13 more tasks than Haiku. Matched by task, 32 were passed only by Opus and 19 only by Haiku (exact McNemar $p = 0.09$). Most of the difference lay in appeals and denials. Haiku cost 5 cents per task in agent calls, against $1.11 for Opus.

GPT-6 Luna trailed the other three. It completed 16 fewer tasks than Sol (29 passed only by Sol, 13 only by Luna, $p = 0.02$) and 15 fewer than Haiku ($p = 0.004$). It still exceeded GPT-5.4 under both April configurations, 38 to 8 with the standard agent and 38 to 36 with OpenAI's computer-use loop. Luna also took the most steps of any model, 6,711, and ten of its runs ended at the step limit, nine of them on easy tasks.

A judge grading its own model. The November judge is GPT-6 Luna, so Luna's notes and appeal text were graded by the same model that wrote them. Language-model judges can favor output resembling their own. The data show no sign that this raised Luna's score. Luna passed 64.6% of its text rubrics, the lowest rate of the four models. On deterministic checks alone, which involve no judge, it also ranked last. A low pass rate cannot exclude a bias that offset weaker writing. Measuring one would require regrading the same outputs with a different judge. The failure review in the previous section covered only Sol and Opus, so the smaller models' failures have not been classified.

Methods and scope of the comparison

Historical data. We used Bedi et al.'s April 10, 2026 paper and the benchmark maintainers' per-task v2 export at revision `b88844c8c3e4` [1, 2]. The export includes a separate analysis that removes selected process checks. We used the original columns, with all checks retained, selected screenshot-only observations and general portal guidance, and matched runs by task ID. The historical standard agents used GPT-5.4 and Claude Opus 4.6. The native agents used the corresponding vendor computer-use loops. The two April configurations show how much the interface for taking actions could affect the same model's score.

New runs. On November 19, 2026, we ran GPT-6.2 Sol and Claude Opus 5.5 once on each v3 task using the released portals and standard agent classes. Adapters sent the requests to the vendors' APIs. Agents received screenshots, the goal and general portal guidance and acted using screen coordinates. They did not receive task-specific walkthroughs or scoring criteria. The v3 tasks contained 1,177 deterministic checks and 517 text rubrics. On November 21, we ran GPT-6 Luna and Claude Haiku 5.5 the same way. The November judge, `openai/gpt-6-luna`, voted three times per rubric. Because GPT-6 Luna was also an agent, it graded its own text in those runs. The April benchmark used GPT-5.4 as judge and contained 1,698 checks in total, compared with 1,694 in v3 [1, 3].

Retrospective review. Claude and Codex reviewed failed checks against the task goal, supplied guidance, actions, saved portal state, grader evidence and relevant source code. Each failure received a written rationale. A consequence such as an absent confirmation after a withheld appeal inherited the judgment on the underlying decision. Any independent clear error placed a run in the agent-error group. Reviewers saw the model identities and scores. No independent billing specialist adjudicated the labels, and inter-rater agreement was not measured. Passing runs were outside this review.

Interpretation. Two newer models completed substantially more simulated administrative tasks than their predecessors under the closest published interaction setting. The gains persisted on checks that require no language-model judgment. The benchmark also moved from v2 to v3 between the evaluations. We expect little effect: on the 96 tasks with identical definitions, GPT rose from 6 to 44 completions and Opus from 13 to 43. Because model, harness and portal releases changed together, the data do not identify the contribution of the model update alone. The tasks were public during the interval, and prior model exposure is unknown. Each task was attempted once, so run-to-run variance is unmeasured. The smaller models' results share every one of these limits, and Luna's text scores also depend on a judge from its own model. A model-only comparison would rerun both generations in the same frozen environment with a common judge. Measuring value in a clinic would also require repeatability, appropriate escalation and confirmation that the requested changes actually reached the patient record.

Reproduction. The source archive includes the per-task comparison extract, scripts for reproducing the tables, and the 540 November task records, with check-level review judgments for the 270 Sol and Opus runs. The comparison script can rebuild its extract from the pinned upstream checkout and the November per-task statistics.

References

[1] S. Bedi, R. Welch, E. Steinberg, et al. *HealthAdminBench: Evaluating Computer-Use Agents on Healthcare Administration Tasks*. April 2026. arXiv:2604.09937. Figure 3 and Table 3.

[2] Shah Lab. Per-run v2 scores, original and process-check variants. Revision `b88844c8c3e4`.

[3] Shah Lab. HealthAdminBench source, task definitions and harness. Same revision.