How Agents Reach the Record Matters: A Parity-Controlled Study of Access Modality for Clinical EHR Agents

Anonymous authors
Paper under double-blind review

Abstract

How access modality affects the measured performance of agents that act on electronic health records (EHRs) remains poorly characterized. We evaluated five frontier models from OpenAI, Google and Anthropic on ten EHR tasks, from chart retrieval to consults with orders and signed notes, under three access modalities: Pixel (screenshots, mouse and keyboard), Fhir (patient-scoped FHIR search and read) and Packet+Action (the full record in the prompt, with Fhir's actions). Tasks, synthetic records, actions and grading were held constant under a protocol frozen before any model call. Physicians wrote or reviewed each task's clinical content. Grading combines deterministic checks on the final chart and recorded actions with a blinded judge for consult notes. Across 1,500 trials (10 per cell), pass@1 was 32.4% under Pixel, 54.0% under Fhir and 60.0% under Packet+Action. Exact task-level sign-flip tests place Pixel 21.6 points below Fhir ($p = 0.031$) and 27.6 below Packet+Action ($p = 0.008$); the structured modalities do not differ reliably ($p = 0.109$). Every model performed worst under Pixel, between-model differences were largest under Pixel (6.0% to 48.0% pass@1), and modality explains about as much variance in pass rates as model choice. Most of the Pixel deficit arises in reaching evidence and in finishing and persisting writes; for one model it may largely reflect the screen driver. Auditing our harness exposed parity defects, and every affected cell was rerun under a frozen plan. These results establish access modality as a critical dimension of EHR-agent evaluation, and we state the parity requirements such comparisons need.

1 Introduction

Language-model agents are now evaluated on clinical work that goes beyond question answering: reading a longitudinal chart, reconciling medications, placing orders and documenting a signed assessment. The benchmarks that do so disagree about how the agent reaches the record. MedAgentBench exposes a FHIR server through function calls (Jiang et al., 2025), FHIR-AgentBench asks questions over FHIR resources (Lee et al., 2026), EHRAgent writes code against tabular EHR data (Shi et al., 2024), and MedAlign pairs clinician instructions with records supplied as text (Fleming et al., 2024). HealthBench, which gives the model no record at all, grades single responses to fixed conversations against physician-written rubrics (Arora et al., 2025). General computer-use benchmarks work through the rendered interface instead: OSWorld gives the agent screenshots and raw mouse and keyboard control (Xie et al., 2024), while WebArena and VisualWebArena have it act on page elements it observes through an accessibility tree or annotated screenshots (Zhou et al., 2024; Koh et al., 2024). A deployed clinical agent may use any of these routes, yet a score obtained on one of them can be read as evidence about clinical capability in general.

If the interface changes what is measured, such comparisons are confounded. A graphical chart may show one value per grid cell where the record holds two. An API may omit the episode date that the screen displays, and a structured action schema may enumerate the only acceptable answer. Each difference can move a pass rate by tens of points with no change in the model's clinical reasoning. For general web agents, API access has been found to outperform browsing on the same sites (Song et al., 2025), and API- and GUI-based agents have been contrasted largely conceptually, with a small

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

prototype on office tasks (Zhang et al., 2025). In the clinical setting, FHIR-AgentBench compares retrieval strategies within FHIR, such as raw queries against specialised tools and natural-language against code-based reasoning (Lee et al., 2026). We are not aware of a controlled comparison on clinical tasks in which the task, record, grader and action semantics are held fixed while the route to the record itself changes, from pixels to an API to the full record in the prompt.

We construct one. The same ten tasks, over the same synthetic patient charts, are run under three modalities. In PIXEL the agent sees screenshots of a simulated EHR web application and acts with mouse and keyboard. In FHIR it reads the record through a FHIR R4 search and read interface scoped to the patient (Health Level Seven International, 2019). In PACKET+ACTION the complete source record is placed in the prompt. The two structured modalities take the same order, medication, note and allergy actions as the UI through one shared executor, and all three are graded by the same deterministic checks on the final chart state and recorded actions and, for signed consult notes, by the same blinded judge.

The measurement is only as good as the parity between modalities, and most of our effort went there. We state parity as concrete requirements on the decision space, the visible state and the evidence used for grading (§3). Auditing our own harness against them exposed defects both in the pilot and after the main run; those found after it were repaired and all 350 affected trials rerun under three addenda, each frozen before its runs, and Appendix B records each defect and its effect as a checklist for others who compare interfaces. A further frozen plan then extended every cell from five to ten trials. On ten tasks that range from deterministic retrieval to consults with medication changes and timed orders (§4), and under a protocol in which no valid outcome is retried (§5), we then estimate the modality effect for five frontier models with exact task-level tests and task-clustered uncertainty (§6). A failure taxonomy and exploratory process checks separate failures of access from failures of reasoning (§7).

With 10 trials per cell, PIXEL trails both structured modalities by a wide margin in exact task-level tests, and both gaps stay negative when any one task is dropped. The two structured modalities are close. PIXEL is the lowest modality for every model, and modality accounts for about as much of the variance in pass rates as the choice of model. The size of the deficit varies across models, and with ten tasks no single model's deficit is established on its own. The pooled effect survives a bound on driver artefacts, although most of one model's deficit may be such an artefact. Ten tasks cannot support claims about clinical competence, so we read the results as evidence about measurement rather than about fitness for clinical use.

2 RELATED WORK

Benchmarks for agents that operate software through its interface range from offline replay of recorded web traces in Mind2Web (Deng et al., 2023) to interactive environments: WebShop (Yao et al., 2022), WebArena (Zhou et al., 2024) and VisualWebArena (Koh et al., 2024) on the web, OSWorld (Xie et al., 2024) and Windows Agent Arena (Bonatti et al., 2025) on desktop operating systems, and AndroidWorld (Rawles et al., 2025) on mobile devices. SeeAct finds grounding a textual plan in the page to be a major source of failure for screenshot-based web agents (Zheng et al., 2024), and CogAgent and SeeClick train GUI grounding as a capability of its own (Hong et al., 2024; Cheng et al., 2024). Song et al. (2025) show that giving web agents API calls in addition to, or instead of, browsing changes success rates on WebArena tasks, and Zhang et al. (2025) compare API- and GUI-based agents largely conceptually, with a small prototype experiment on office tasks. We hold task and grader fixed and vary the modality, in a domain where a wrong write is costly.

Reasoning-and-acting prompting (Yao et al., 2023), tool-augmented models (Schick et al., 2023; Qin et al., 2024) and multi-environment agent suites (Liu et al., 2024) made tool and API calls a standard way for language agents to act, and function-calling benchmarks such as BFCL (Patil et al., 2025), which builds on Gorilla (Patil et al., 2024), evaluate that interface directly. τ-bench grades tool agents against database end states and introduced the pass^k reliability metric (Yao et al., 2025), which we report alongside pass@k (Kulal et al., 2019), using the unbiased estimator of Chen et al. (2021). Recent guidance on agentic benchmarks separates task validity from outcome validity and warns against graders that can be passed without doing the task (Zhu et al., 2025); several of the parity defects we describe are instances of the same problem seen from the interface side. We follow Miller (2024) in reporting clustered intervals, because trials within a task are not independent.

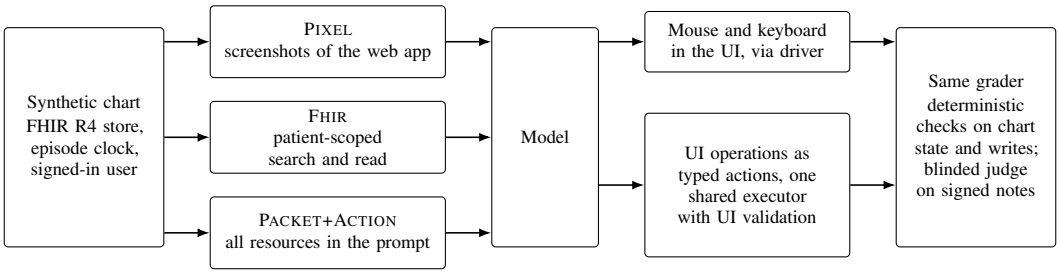

Figure 1: One chart reached three ways. The task prompt, record, episode clock and grader are shared; the modalities differ in how the model observes the record and, for PIXEL, in how its actions reach the chart.

Clinical evaluations of language models began with examination-style and consumer questions (Jin et al., 2021; Singhal et al., 2023) and have since moved to clinical decision-making with realistic information gathering (Hager et al., 2024), simulated patient encounters (Schmidgall et al., 2026), rubric-graded conversations (Arora et al., 2025), broad task taxonomies (Bedi et al., 2026) and instruction following over EHRs (Fleming et al., 2024). EHR agents have been evaluated through code over tables (Shi et al., 2024), FHIR function calls (Jiang et al., 2025) and FHIR question answering (Lee et al., 2026), and FHIR-grounded patient-facing chat has been piloted in an app (Schmiedmayer et al., 2024). Each fixes one modality. FHIR-AgentBench comes closest to our question, comparing direct FHIR queries with retrieval tools, single- with multi-turn interaction and natural-language with code-based reasoning, but all of these read the same FHIR store and none acts on the chart. We instead vary the route to the record, screen included, with actions and grader fixed.

Language-model judges are widely used for open-ended outputs (Zheng et al., 2023; Liu et al., 2023). They can be sensitive to position (Wang et al., 2024; Zheng et al., 2023) and favour their own generations (Zheng et al., 2023; Panickssery et al., 2024), and some also rate models from their own family more highly (Spiliopoulou et al., 2025). Agentic judges extend the idea to whole trajectories (Zhuge et al., 2025). Our judge is multi-turn and uses tools, but it grades the signed note rather than the trajectory, runs only after the deterministic checks pass, is blind to model and modality, and was qualified on fixture notes before use (§4). Its agreement with clinicians has not yet been measured (§8).

3 BENCHMARK AND ACCESS MODALITIES

Each trial runs against a fresh instance of a simulated EHR web application backed by a FHIR R4 store (Health Level Seven International, 2019). Its layout follows common chart-review designs, with a patient header, a laboratory grid and trend view, and pages for medications, orders, notes and allergies. Patients are synthetic, and no real patient data are used. Every task fixes an episode clock (the simulated "now"), a signed-in practitioner and the source record from which its facts were derived. The records contain hazards found in real charts: duplicate rows, a stopped duplicate of an active order, conflicting values on the same date and medication rows dated after the clock.

The modalities differ in how the agent observes the record and, for PIXEL, how it acts (Figure 1, Table 1). In PIXEL the agent receives screenshots of the application in a Chromium browser at 1920×1080 and acts with mouse and keyboard. OpenAI models use their provider's native computer-use tool; the other models use a function-calling computer tool with basic mouse and keyboard actions. In FHIR the agent reads through a FHIR search and read interface restricted to the task patient, which returns the raw resources of the source record as of the episode clock. In PACKET+ACTION the same resources are serialised into the prompt. The two structured modalities share the system prompt, the action tools and the executor that applies them, so they differ only in whether evidence has to be fetched. What we estimate is therefore the effect of each access route as built, meaning its observation channel, its action channel and the driver that connects the model to them, rather than a model's visual grounding ability in isolation.

Structured actions mirror the operations of the UI one to one. An agent can write, sign and reopen a consult note, order laboratory tests and stop or hold medications, and the potassium and allergy tasks

Table 1: Access modalities. All three share the task prompt, the synthetic record, the episode clock, the grader and the judge.

	PIXEL	FHIR	PACKET+ACTION
Observation	Screenshots (1920×1080)	Patient-scoped search and read over raw resources	All raw resources in the prompt
Actions	Mouse and keyboard in the UI	The UI's operations as typed actions, one executor	
Driver	Native computer-use tool (OpenAI); function-calling tool (others)	Provider function calling	
Episode clock	Displayed in the UI	Stated in the prompt	Stated in the record
Budget	7,200 s, no turn cap[†]	40 model turns, 1,800 s	

[†]Allergy task: 80 turns, 160 actions and 1,800 s for every PIXEL model.

add the UI's protocol and coded-allergy operations with the same fields and choices as the UI. The executor applies the UI's own pre-save validation. Grading reads only trusted evidence: captures of the chart before and after the trial, a receipt for every write, an audit event attributed to the acting user for each clinical action, and proof that the chart was restored to its baseline. A trial with missing or inconsistent evidence cannot pass. If the gap is due to infrastructure the trial is classified invalid and handled as described in §5; it is never scored from partial evidence.

We treat parity between modalities as a methodological requirement rather than an implementation detail, and each part of it was learned from a defect in our own harness. A structured tool must first admit the same choices as the UI, wrong ones included, or it measures schema following instead of judgement. Our first structured allergy tool offered only the correct choices; with the UI's open fields, the structured pilot cells fell from 8/8 to 0/8 passes. Second, whatever the screen shows or allows must be available to structured agents, and whatever the screen offers must work. A missing episode date, validation stricter than the UI's, a short laboratory list and a patient search that did not accept input each changed outcomes in one modality, and in three tasks the structured tools lacked medication or order actions that the UI offered. Third, evidence custody fails closed, as described above. We audited the harness against these requirements after the main run, fixed the five defects that remained, and reran all 350 trials of the affected cells, which then had five trials each, under three addenda, each frozen and hash-committed before its runs and preceded by provider-free controls (160/160 passed). The superseded trials are kept as a sensitivity analysis. Appendix B summarises the parity properties, the defects and the effect of each rerun.

4 TASKS AND GRADING

Table 2 lists the ten tasks. The clinical content of each task was written or reviewed and revised by physicians; the prompts and grading implementations were adapted from it with AI assistance and reviewed by the authors. Five are retrieval and protocol tasks whose output is a structured answer: a mean systolic blood pressure over a window, the latest eGFR, two audits of conflicting laboratory and spirometry values under supplied rules, and a potassium protocol that also requires holding every qualifying ACE-inhibitor order and scheduling a recheck. The other five are consults phrased as a clinician would phrase them, naming a clinical problem and asking for a signed assessment and plan. The consult prompts do not name the actions. The agent has to infer from the chart which medication to stop, which tests to order and when, and what the note must say. Appendix A describes each task.

Each episode is graded by deterministic checks and a multi-turn agentic judge over the final environment state and the recorded actions. The deterministic checks compare the persisted chart changes and write actions with the task contract. They cover patient identity, the required actions and their readback, the absence of forbidden or unrelated changes, and write attribution; retrieval answers are compared with fixed expected values. The judge, Claude Opus 5 run as a tool-using agent, grades the clinical content of the signed note against the patient record one rubric criterion at a time, blind to model, provider and modality. It sees the note and the record but not the trajectory, because persisted actions are already covered by the deterministic checks, and it runs only on the four note-bearing consults and only after every deterministic check has passed. Before any evaluated trial it

Table 2: Tasks. Gates are deterministic checks (consult tasks) or answer checkpoints (retrieval tasks; PIXEL/structured where they differ). Crit. is the number of rubric criteria scored by the judge.

Task	Type	Required outcome	Gates	Crit.
BP	Retrieval	90-day mean systolic BP with evidence dates	6	–
eGFR	Retrieval	Latest final eGFR before the clock	6	–
Conflicting labs	Retrieval	Bounds and claim labels over same-date conflicts	17	–
Spirometry	Retrieval	Pre/post FEV_1 bounds and claim labels	15	–
Potassium	Protocol	Report potassium; hold qualifying orders; schedule recheck	7/6	–
Thyroid	Consult	Signed note; repeat thyroid testing	19	4
Duplicate anticoag.	Consult	Resolve the duplication; monitoring; note	22	5
NSAID safety	Consult	Correct the highest-risk medication; monitoring; note	20	5
Depression	Consult	Signed note on monitoring evidence; no other writes	15	4
Allergy	Consult	Record a reported drug allergy as a coded entry	14	0

was qualified on five constructed notes per task, ranging from a gold note to an unsafe one, and met every expectation (Appendix F).

The binary outcome is conjunctive. A trial passes if every deterministic check holds and, where the judge applies, its band scores (EXCELLENT 1, GOOD 2/3, PARTIAL 1/3, FAIL 0) average at least 0.75 with no FAIL. As a secondary metric we report a weighted score $r \in [0, 1]$ in which critical checkpoints (identity, safety, required actions and the task's central clinical finding) carry more weight than supporting detail and gate the pass: a trial passes this weighted metric only if every critical checkpoint holds and $r \geq 0.75$ (Appendix G). The assignment of checkpoints to the two classes was fixed before any trial was scored under it. Two evaluated models come from the judge's vendor. Judges are known to favour their own outputs (Panickssery et al., 2024), and have been reported to favour models of their own family (Spiliopoulou et al., 2025); we test for this in §6.

5 EXPERIMENTAL PROTOCOL

The design crosses 10 tasks, 3 modalities, 5 models and 10 trials per cell, for 1,500 trials. The models are GPT-6 Astra and GPT-5.6 Sol (OpenAI), Gemini 3.1 Pro (Google, reached through an aggregator API), and Claude Fable 5.1 and Claude Opus 5.5 (Anthropic); Appendix C gives the exact API identifiers. OpenAI requests set reasoning effort to medium in every task and modality, while Anthropic and Gemini requests set no reasoning or thinking parameter and run at provider defaults. Budgets are those of Table 1.

The trial protocol and an analysis plan were frozen and hash-committed before any model call. Together they fixed the prompts, rubrics, graders, judge, outcome rules, contrasts and statistical models, and each attempt verifies that its prompt, rubric, record and runtime match the frozen configuration before it starts. Opus was added by an addendum frozen before any Opus call, and each parity addendum was likewise frozen before its runs, as was the extension that added trials 6 to 10 of every cell on the final configuration. Other elements came after the main run, when its outcomes already existed: the post-run parity audit, the critical/non-critical mapping of the weighted score and the process checks of §7, each fixed before it was applied to any trial, and the statistical analyses described below together with the judge, driver and failure-stage analyses of §6 and §7, which were chosen with the outcomes known. Every valid trial counts, and none is retried or discarded because of its outcome. Attempts that fail for a qualified infrastructure reason are rerun at the same trial index at most twice, attempts lost to a defect of the harness itself never decide a trial, trials rerun in the parity addenda replace the originals, and every rerun and reused pilot result is listed in Appendix C.

The analysis plan counts every valid trial "including model failures and turn limits". Accordingly, a trial that ends on a budget (turn or output-token limit, deadline, or the PIXEL driver's limit on rejected actions), on a provider refusal, on an empty or malformed action, without a completion or stopped by infrastructure counts as a failure. There are 62 such adjudicated failures, all under PIXEL, and 57 ended on a budget, usually while the agent was looping (Table 9 reconciles all our

Table 3: pass@1 (%) by model and modality with 95% task-clustered bootstrap intervals, and paired modality differences (percentage points). Primary rule: budget-exhausted trials counted as failures; other rules in Table 16. pass^k uses $k = 10$.

	pass@1			Paired difference		
Model	PIXEL	FHIR	PACKET+ACTION	Pix.−FHIR	Pix.−Pkt.	FHIR−Pkt.
GPT-6 Astra	48.0 [24.0, 71.0]	64.0 [34.0, 90.0]	60.0 [30.0, 90.0]	−16.0 [−45.0, +11.0]	−12.0 [−44.0, +19.0]	+4.0 [0.0, +12.0]
GPT-5.6 Sol	27.0 [11.0, 46.0]	60.0 [30.0, 90.0]	61.0 [34.0, 87.0]	−33.0 [−56.0, −11.0]	−34.0 [−52.0, −16.0]	−1.0 [−8.0, +6.0]
Gemini 3.1 Pro	6.0 [0.0, 18.0]	32.0 [4.0, 62.0]	49.0 [20.0, 78.0]	−26.0 [−60.0, +6.0]	−43.0 [−78.0, −7.0]	−17.0 [−40.0, 0.0]
Claude Fable 5.1	40.0 [13.0, 68.0]	47.0 [21.0, 72.0]	57.0 [27.0, 87.0]	−7.0 [−47.0, +32.0]	−17.0 [−59.0, +25.0]	−10.0 [−22.0, 0.0]
Claude Opus 5.5	41.0 [14.0, 70.0]	67.0 [42.0, 88.0]	73.0 [49.0, 93.0]	−26.0 [−49.0, −5.0]	−32.0 [−56.0, −10.0]	−6.0 [−15.0, +2.0]
All models	32.4 [17.2, 49.2]	54.0 [32.2, 73.8]	60.0 [37.4, 80.8]	−21.6 [−40.4, −5.4]	−27.6 [−44.8, −12.2]	−6.0 [−11.6, −1.2]
pass^k, all	12.0 [0.0, 26.0]	42.0 [20.0, 62.0]	48.0 [26.0, 70.0]			

termination counts). Excluding them would reward failing slowly. The frozen execution classifier instead drops them from denominators, and we report that rule as a sensitivity analysis. Format-only grading relaxations appear only in sensitivity analyses (Appendix D).

The unit of inference is the task. For each model–modality cell we report pass@1, the mean success over its trials, and pass^k, the probability that all k trials succeed (Yao et al., 2025), with k equal to the trials per cell. Pooled estimates are unweighted means over task–model cells, and paired differences are means over task–model pairs. Intervals are 95% percentile intervals from a task-clustered bootstrap (10,000 resamples of the ten tasks, seed 20260924), and paired differences resample tasks jointly so that each task contributes to both arms. The primary contrasts are PIXEL against FHIR and against PACKET+ACTION; FHIR against PACKET+ACTION is secondary, as are the weighted score and its pass rate. Tests are exact sign-flip tests (Fisher, 1935) over all 2^{10} assignments of signs to the ten per-task differences. Odds ratios come from a logistic GEE (Liang & Zeger, 1986) clustered on task, with small-sample standard errors (Mancl & DeRouen, 2001); a mixed model whose modality effect varies by task checks only the direction, and per-model tests use Holm's correction (Holm, 1979) over fifteen tests. The analysis plan specified a mixed model with random task intercepts; we replaced it after the main run because the outcomes showed a modality effect that varies strongly by task, which that model does not allow. Wilson intervals (Wilson, 1927) per cell, and pass@k (Kulal et al., 2019) (unbiased estimator of Chen et al. (2021)) and pass^k for $k \in \{3, 10\}$ are in Appendix D.

6 RESULTS

Pooled over models and tasks, pass@1 is 32.4 [17.2, 49.2]% for PIXEL, 54.0 [32.2, 73.8]% for FHIR and 60.0 [37.4, 80.8]% for PACKET+ACTION (Table 3, Figure 2). Both primary contrasts favour the structured modalities. PIXEL−FHIR is −21.6 points, lower on 7 of the ten tasks and higher on 2 (exact sign-flip $p = 0.031$), and PIXEL−PACKET+ACTION is −27.6 points, lower on 9 ($p = 0.008$). Leaving out any one task keeps both gaps negative (−13.8 to −25.8 and −20.4 to −32.4 points, with p up to 0.063 and 0.016), and leaving out the task at the floor, NSAID, where no cell passes more than four of ten trials, widens them to −24.0 ($p = 0.031$) and −29.6 points ($p = 0.012$). On the odds scale, the task-clustered GEE gives odds ratios of 0.41 [0.15, 1.07] and 0.31 [0.11, 0.91], so against FHIR the interval includes one ($p = 0.069$), and a mixed model that lets the PIXEL effect vary by task agrees in direction (Table 13). PACKET+ACTION exceeds FHIR by 6.0 points. Its bootstrap interval in Table 3 excludes zero, but the exact test gives $p = 0.109$ and the contrast is secondary, so we describe the two structured modalities as close rather than ordered. Neither primary contrast changes sign under any sensitivity analysis (Table 16). On trials 1 to 5, the only ones run before the parity repairs, the repairs narrowed the two gaps by 5.2 and 7.6 points, almost entirely through the first addendum (Appendix B).

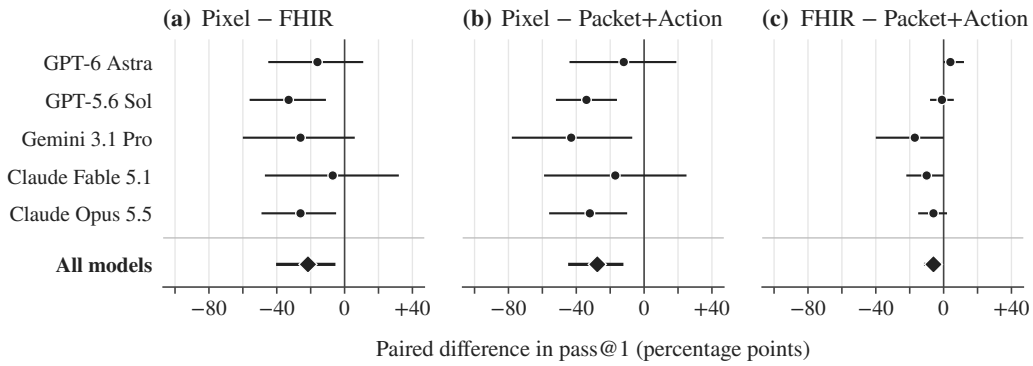

Figure 2: Paired differences in pass@1 between modalities, per model and pooled over models (k=10; primary rule). Points are means over the ten tasks and bars are 95% task-clustered bootstrap intervals; negative values favour the second modality in each panel. Per-model pass rates are in Table 3.

The size of the deficit depends on the model. Every per-model estimate against either structured modality is negative, from −7 to −43 points, largest for GPT-5.6 Sol against FHIR and Gemini 3.1 Pro against PACKET+ACTION and smallest for Claude Fable 5.1 against FHIR and GPT-6 Astra against PACKET+ACTION. With ten tasks per model, none of the fifteen per-model sign-flip tests survives Holm correction (smallest adjusted $p = 0.176$), so we read the per-model results as consistent in direction rather than individually established (Table 14). Reliability follows the same order. pass^k is below pass@1 in every modality, and the gap is widest for PIXEL, where 52.0% of cells succeed at least once in ten trials but only 12.0% succeed in all ten (Table 15).

Access modality moves pass@1 about as much as the choice of model. In a two-way decomposition of the fifteen model–modality cell means, modality accounts for 48% of the variance and model for 41%. Within a model, pass@1 spans 16 to 43 points across modalities; within a modality, it spans 24 to 42 points across models. PIXEL is the lowest modality for every model, and it reorders the models: GPT-6 Astra leads under PIXEL and Claude Opus 5.5 under both structured modalities.

The checkpoint-weighted score leads to the same modality conclusions, but it is not independent evidence: binary and weighted verdicts agree on 1,470 of 1,500 trials, and the remainder differ only through judge bands (Appendix G). Neither the judge nor the driver accounts for the gap (Appendix E). The judge, Claude Opus 5, sees only trials whose hard gates all pass, and ignoring its verdicts widens PIXEL−FHIR to −26.0 points. It passed 81.1% of judged trials from Anthropic models and 84.1% from OpenAI models, so it shows no preference for its own vendor. The 60 driver-bound exclusions (episodes ending on the rejection limit, the deadline or a turn or output-token limit) were all PIXEL (Table 9). Excluding them in every modality, a bound that assumes they would have passed at the model's own rate, moves PIXEL−FHIR to −17.2 points [−35.1, −1.4]. Most of that change comes from Claude Opus 5.5, whose PIXEL rate rises from 41.0% to 68.3%, above its 67.0% under FHIR, so its PIXEL deficit may be largely an artefact of the driver; no other model moves by more than 6.5 points. Every remaining asymmetry between the modalities that is not neutral would inflate the PIXEL deficit (Table 17), and the pooled effect remains under the bound.

Pooling models within each task (Table 10 in Appendix D) shows where the deficit arises. It is largest for potassium and spirometry, for different reasons according to the process checks of §7: PIXEL potassium agents often never displayed the latest value, whereas PIXEL spirometry agents reached every value and still misread the comparison. The deficit is smaller for the anticoagulant consult and, against FHIR, for the BP and eGFR retrievals. NSAID is near the floor in every modality and allergy in both structured modalities, so neither can show much of an effect; PIXEL scores above FHIR on allergy and, narrowly, on eGFR. Per-cell counts are in Figure 3.

Table 4: Failure categories over failed trials (k=10; primary rule). A trial can fall in several categories, so columns do not sum to the last row.

Failure category	PIXEL	FHIR	PACKET+ACTION
Budget, refusal, stop or other early end	62	0	0
Response format (parse, extra field)	29	27	24
Patient-identifier format only	4	19	2
Missed or misread chart evidence	72	0	0
Wrong chart action (retrieval task)	7	0	0
Medication decision (wrong target or extra change)	39	54	23
Required order missing, duplicated or mistimed	84	88	82
Note not signed once / not reopened	87	12	2
Coded-entry semantics (allergy)	41	50	50
Clinical reasoning below judge threshold	15	37	35
Failed trials (any category)	338	230	200

7 ANALYSIS

Table 4 assigns each failed trial one or more categories, based on the gates and checkpoints it failed and, where applicable, the judge's verdict. Missed or misread chart evidence occurs only under PIXEL, as do all 62 adjudicated failures, and most notes that were never signed or reopened. Medication-decision failures are most common under FHIR. Judge-threshold failures are more common in the structured modalities, where agents reach the clinical decision more often, and coded-entry failures on the allergy task are close to uniform across modalities. Much of the modality effect therefore appears to arise before clinical reasoning begins, in finding, reading and persisting information.

An exploratory process check tests this reading directly. For each task we fixed, from the frozen task contract and before scoring any trial, the key facts an agent should consult before its first write or answer (24 in total; Appendix H). FHIR agents read every key fact in 446 of 450 trials, so their failures are not retrieval failures. PIXEL agents had every fact on screen in 73% of trials, and this separated outcomes sharply, with 42% of fully covered trials passing against 4% of the rest. Among PIXEL failures, 62% had seen every fact and 34% never reached at least one. Having the evidence on screen did not mean the failure was one of reasoning, however. Assigning each failed trial on nine tasks to the first stage at which it failed (Appendix H), the 21.6-point gap between FHIR and PIXEL comprises 22.4 points from evidence never reached and 12.9 from episodes that ended early or never persisted their writes, offset by 13.9 points of decision and judge failures that are rarer under PIXEL because fewer of its trials get that far. Among trials that had seen the evidence and persisted their writes, 62.8% passed the decision gates and the judge under PIXEL, against 62.0% under FHIR and 68.4% under PACKET+ACTION. As far as this check can tell, the PIXEL deficit arises in access and in finishing and persisting work, not in clinical reasoning.

Manual review of the single-trial pilot (four models, 120 cells) shows what these failures look like; the observations are illustrative. In the conflicting-labs task the laboratory grid shows one value per date and analyte, while the record holds two conflicting same-date values for alkaline phosphatase and ALT. Three of four PIXEL agents reported only the displayed values and failed the same eight checkpoints, and the fourth opened the trend view, saw both values and answered correctly; structured agents receive both values directly. In the NSAID task, once structured validation matched the UI, all four FHIR agents changed an anticoagulant, and left the higher-risk NSAID in place for an 80-year-old patient, while no PACKET+ACTION agent touched the anticoagulants. The two modalities share every tool and differ only in whether the medication list had to be fetched. Some failures are clinically defensible. In the duplicate-anticoagulant consult, eight of ten pilot failures omitted a monitoring order that the frozen contract requires but the agent argued was unnecessary for the regimen it kept. On the allergy task most failures in both the pilot and the final run recorded an otherwise complete entry with a weaker verification status than the contract requires, which the prompt's wording arguably allows (§8). In the final run, both Anthropic models also repeatedly requested an undeclared triple-click action from the function-calling computer tool, which ended many of Opus's PIXEL episodes (§6).

8 Discussion and limitations

Health systems and vendors deciding how an agent should reach the record face the choice that our modalities isolate. On these tasks, agents that worked through a structured interface succeeded more often than agents operating the UI. The difference arose in locating information and in finishing and persisting writes; screen agents that had seen the evidence and persisted their writes decided about as well as API agents (§7). This favours structured access where it exists, provided the structured surface is complete. An interface that omits the episode date, rejects actions the UI permits or narrows the list of orderable tests changes what the agent does, sometimes in its favour. Where UI automation is the only route, its evaluation should check that the agent saw the values it acted on. In both cases a score obtained through one interface says little about behaviour through another, so acceptance testing should use the deployed interface.

This guidance rests on synthetic charts and a simulated application, so neither the difficulty nor the failure modes need match production EHRs. Ten tasks and 10 trials per cell give wide per-model intervals, and each consult task rests on a single patient. Some asymmetries also remain after the fixes of §3, among them the PIXEL drivers' retained history and rejection feedback, the structured agents' inability to read back their own writes, and budgets that differ by modality (Table 1). Appendix B lists them and Table 17 gives their direction; none favours PIXEL.

The judge shares a vendor with two evaluated models, and judges have been reported to favour models of their own family (Spiliopoulou et al., 2025); we see no such preference here, but four judged tasks are a small test. Qualification on constructed notes does not establish agreement with clinicians on agent-written notes; a blinded clinician agreement study is future work, as are clinician-authored process checks. The critical/non-critical mapping of the weighted score is our own rather than the product of clinician review. The implemented thresholds of some frozen grading rules also warrant further independent clinical review. In the allergy, duplicate-anticoagulant and NSAID tasks the frozen contract settles a point on which clinicians could reasonably differ.

A few infrastructure reruns are correlated with model behaviour, notably custody reruns after one model's habit of reopening historical notes under PIXEL, and the primary rule counts all 62 adjudicated failures, including deadline cut-offs, refusals, malformed or empty actions and stopped trials, as failures (Table 9). The later trials extend cells that the parity audit had already shaped, adding precision but not independence from it.

9 Conclusion

With tasks, records, graders and action semantics held fixed, the interface through which an agent reaches a clinical chart changes its measured success by tens of percentage points, about as much as the choice of model, and by different amounts for different models. On screen, the gap arises before clinical reasoning, in finding values that the screen does not show at once and in finishing and persisting writes; once the evidence was seen and the writes persisted, screen agents decided about as well as API agents. Interface comparisons are also easy to get wrong. An answer-bearing schema, an extra validation rule, a hidden clock and a broken search field each changed outcomes in one modality, and the schema alone took the structured allergy pilot from eight passes out of eight to none. Clinical agent benchmarks should report which modality they use and equalise the decision space and visible state before comparing modalities, and a score obtained on one modality should be read as evidence about that modality.

Ethics statement

This work uses no human subjects data, and all patient records are synthetic. The simulated EHR writes only to disposable per-trial stores that are reset and verified after every trial. Our results concern measurement and should not be read as evidence that any model is safe for clinical use. Several of the failures we report, such as leaving an NSAID active for an anticoagulated 80-year-old, would be harmful in practice. The evaluated models are commercial systems used under their providers' terms. This work was funded by the authors' employer. The authors are employed by a company that develops clinical AI evaluation environments and may benefit commercially from this work. The authors' employer has commercial relationships with developers of some evaluated

models. No developer of an evaluated model funded, reviewed or was informed of this study; all model access was purchased through standard commercial APIs.

REPRODUCIBILITY STATEMENT

We describe the tasks, the grading and the protocol at the level needed to interpret the results. Section 3 and Appendix B specify the modalities and the parity requirements, Section 4 and Appendix A describe the tasks and their grading, and Section 5 and Appendix C give the frozen protocol, the rerun and exclusion rules and the disclosures. Appendices D to H report every analysis, including the judge's qualification and the weighted score, and Appendix I describes the retained evidence and the analysis pipeline. The task specifications, rubrics and grading materials are proprietary; on request through the area chair, we will consider sharing them, and the per-trial outcome table, confidentially with reviewers, in a form that preserves anonymity.

AI USE STATEMENT

The evaluated systems are language models, and their use is the subject of the study (Sections 3–5). The judge is also a language model: it grades the clinical reasoning of the signed consult notes against the record, as described in Section 4. The clinical content of each task was written or reviewed and revised by physicians; the prompts and grading implementations were adapted from it with AI assistance and reviewed by the authors. Beyond that, AI assistants were used for engineering, including parts of the evaluation harness and the audit that found the parity defects of Appendix B; for synthetic data generation; for writing the analysis scripts and helping to analyse the results; for finding related work and verifying citations; and for drafting and editing the manuscript. The synthetic patient records were partly generated with language models, and the clinical scenarios of two tasks came from a language-model task-generation pipeline and were then reviewed and revised by physicians. The authors reviewed all content and take responsibility for it.

REFERENCES

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. HealthBench: Evaluating large language models towards improved human health. arXiv preprint arXiv:2505.08775, 2025.

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M. Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, Hao Qiu, Shrey Jain, Leonardo Schettini, Mehr Kashyap, Jason Alan Fries, Akshay Swaminathan, Philip Chung, Fateme Nateghi Haredasht, Ivan Lopez, Asad Aali, Gabriel Tse, Ashwin Nayak, Shivam Vedak, Sneha S. Jain, Birju Patel, Oluseyi Fayanju, Shreya Shah, Ethan Goh, Dong han Yao, Brian Soetikno, Eduardo Reis, Sergios Gatidis, Vasu Divi, Robson Capasso, Rachna Saralkar, Chia-Chun Chiang, Jenelle Jindal, Tho Pham, Faraz Ghoddusi, Steven Lin, Albert S. Chiou, Hyo Jung Hong, Mohana Roy, Michael F. Gensheimer, Hinesh Patel, Kevin Schulman, Dev Dash, Danton Char, Lance Downing, Francois Grolleau, Kameron Black, Bethel Mieso, Aydin Zahedivash, Wen wai Yim, Harshita Sharma, Tony Lee, Hannah Kirsch, Jennifer Lee, Nerissa Ambers, Carlene Lugtu, Aditya Sharma, Bilal Mawji, Alex Alekseyev, Vicky Zhou, Vikas Kakkar, Jarrod Helzer, Anurang Revri, Yair Bannett, Roxana Daneshjou, Jonathan Chen, Emily Alsentzer, Keith Morse, Nirmal Ravi, Nima Aghaeepour, Vanessa Kennedy, Akshay Chaudhari, Thomas Wang, Sanmi Koyejo, Matthew P. Lungren, Eric Horvitz, Percy Liang, Michael A. Pfeffer, and Nigam H. Shah. Holistic evaluation of large language models for medical tasks with MedHELM. *Nature Medicine*, 32(3):943–951, 2026. doi: 10.1038/s41591-025-04151-2.

Rogerio Bonatti, Dan Zhao, Francesco Bonacci, Dillon Dupont, Sara Abdali, Yinheng Li, Yadong Lu, Justin Wagle, Kazuhito Koishida, Arthur Bucker, Lawrence Keunho Jang, and Zheng Hui. Windows Agent Arena: Evaluating multi-modal OS agents at scale. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 4874–4910. PMLR, 2025.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374, 2021.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Li YanTao, Jianbing Zhang, and Zhiyong Wu. SeeClick: Harnessing GUI grounding for advanced visual GUI agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9313–9332, 2024. doi: 10.18653/v1/2024.acl-long.505.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2Web: Towards a generalist agent for the web. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 36, 2023.

Ronald A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.

Scott L. Fleming, Alejandro Lozano, William J. Haberkorn, Jenelle A. Jindal, Eduardo P. Reis, Rahul Thapa, Louis Blankemeier, Julian Z. Genkins, Ethan Steinberg, Ashwin Nayak, Birju S. Patel, Chia-Chun Chiang, Alison Callahan, Zepeng Huo, Sergios Gatidis, Scott J. Adams, Oluseyi Fayanju, Shreya J. Shah, Thomas Savage, Ethan Goh, Akshay S. Chaudhari, Nima Aghaeepour, Christopher Sharp, Michael A. Pfeffer, Percy Liang, Jonathan H. Chen, Keith E. Morse, Emma P. Brunskill, Jason A. Fries, and Nigam H. Shah. MedAlign: A clinician-generated dataset for instruction following with electronic medical records. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22021–22030, 2024. doi: 10.1609/aaai.v38i20.30205.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30(9):2613–2622, 2024. doi: 10.1038/s41591-024-03097-1.

Health Level Seven International. HL7 FHIR release 4 (v4.0.1). https://hl7.org/fhir/R4/, 2019.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and Jie Tang. CogAgent: A visual language model for GUI agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14281–14290, 2024.

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), 2025. doi: 10.1056/AIdbp2500144.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. doi: 10.3390/app11146421.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. VisualWebArena: Evaluating multimodal agents on realistic visual web tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 881–905, 2024. doi: 10.18653/v1/2024.acl-long.50.

Sumith Kulal, Panupong Pasupat, Kartik Chandra, Mina Lee, Oded Padon, Alex Aiken, and Percy Liang. SPoC: Search-based pseudocode to code. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Gyubok Lee, Elea Bach, Eric Yang, Tom Pollard, Alistair Johnson, Edward Choi, Yugang Jia, and Jong Ha Lee. FHIR-AgentBench: Benchmarking LLM agents for realistic interoperable EHR question answering. In *Proceedings of the Fifth Machine Learning for Health Symposium*, volume 297 of *Proceedings of Machine Learning Research*, pp. 1297–1315. PMLR, 2026.

Kung-Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986. doi: 10.1093/biomet/73.1.13.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations*, pp. 52989–53046, 2024.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023. doi: 10.18653/v1/2023.emnlp-main.153.

Lloyd A. Mancl and Timothy A. DeRouen. A covariance estimator for GEE with improved small-sample properties. *Biometrics*, 57(1):126–134, 2001. doi: 10.1111/j.0006-341X.2001.00126.x.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. arXiv preprint arXiv:2411.00640, 2024.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, pp. 68772–68802, 2024.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive APIs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 126544–126565, 2024. doi: 10.52202/079017-4020.

Shishir G. Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. 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*, volume 267 of *Proceedings of Machine Learning Research*, pp. 48371–48392. PMLR, 2025.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *International Conference on Learning Representations*, 2024.

Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, Daniel Toyama, Robert Berry, Divya Tyamagundlu, Timothy Lillicrap, and Oriana Riva. AndroidWorld: A dynamic benchmarking environment for autonomous agents. In *International Conference on Learning Representations*, 2025.

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

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Pontes Reis, Jeffrey Jopling, and Michael Moor. AgentClinic: a multimodal benchmark for tool-using clinical AI agents. *npj Digital Medicine*, 9(1):499, 2026. doi: 10.1038/s41746-026-02674-7.

Paul Schmiedmayer, Adrit Rao, Philipp Zagar, Vishnu Ravi, Aydin Zahedivash, Arash Fereydooni, and Oliver Aalami. LLM on FHIR – demystifying health records. arXiv preprint arXiv:2402.01711, 2024.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce C. Ho, Carl Yang, and May Dongmei Wang. EHRAgent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22315–22339, 2024. doi: 10.18653/v1/2024.emnlp-main.1245.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023. doi: 10.1038/s41586-023-06291-2.

Yueqi Song, Frank F. Xu, Shuyan Zhou, and Graham Neubig. Beyond browsing: API-based web agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11066–11085, 2025. doi: 10.18653/v1/2025.findings-acl.577.

Evangelia Spiliopoulou, Riccardo Fogliato, Hanna Burnsky, Tamer Soliman, Jie Ma, Graham Horwood, and Miguel Ballesteros. Play favorites: A statistical method to measure self-bias in LLM-as-a-judge. arXiv preprint arXiv:2508.06709, 2025.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2024.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, volume 35, pp. 20744–20757, 2022.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations*, 2025.

Chaoyun Zhang, Shilin He, Liqun Li, Si Qin, Yu Kang, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. API agents vs. GUI agents: Divergence and convergence. arXiv preprint arXiv:2503.11069, 2025.

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. GPT-4V(ision) is a generalist web agent, if grounded. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61349–61385. PMLR, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 36, pp. 46595–46623, 2023.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, pp. 15585–15606, 2024.

Yuxuan Zhu, Tengjun Jin, Yada Pruksachatkun, Andy Zhang, Shu Liu, Sasha Cui, Sayash Kapoor, Shayne Longpre, Kevin Meng, Rebecca Weiss, Fazl Barez, Rahul Gupta, Jwala Dhamala, Jacob Merizian, Mario Giulianelli, Harry Coppock, Cozmin Ududec, Antony Kellermann, Jasjeet Sekhon, Jacob Steinhardt, Sarah Schwettmann, Arvind Narayanan, Matei Zaharia, Ion Stoica, Percy Liang, and Daniel Kang. Establishing best practices in building rigorous agentic benchmarks. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 38, 2025. doi: 10.52202/085713-5547.

Mingchen Zhuge, Changsheng Zhao, Dylan R. Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, Yangyang Shi, Vikas Chandra, and Jürgen Schmidhuber. Agent-as-a-Judge: Evaluate agents with agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80569–80611. PMLR, 2025.

A TASK DESCRIPTIONS

All patients are synthetic, and each task uses one synthetic patient and a fixed episode clock. Retrieval prompts state the quantity to report and any rules for reading the record. Consult prompts name the clinical problem and ask for a signed assessment and plan without naming the actions required. Every task also requires the chart to be left unchanged apart from the required writes, and every consult requires complete evidence capture. The task specifications and grading materials are proprietary, so we describe each task in one sentence; Table 2 gives the number of deterministic checks and rubric criteria for each.

A.1 RETRIEVAL AND PROTOCOL TASKS

The blood-pressure task asks for the mean systolic pressure over a fixed window ending at the clock, reported with its evidence dates as a structured answer. The eGFR task asks for the most recent final eGFR recorded on or before the clock. The conflicting-laboratory task asks the agent to select a panel date under supplied rules, bound derived quantities across conflicting values recorded on that date, and label a set of claims as supported, refuted or indeterminate. The spirometry task applies the same logic to pre- and post-bronchodilator lung function. The potassium task applies a supplied synthetic protocol: the agent reports the qualifying values and, where the protocol requires it, holds the qualifying medication orders and schedules a recheck in a single commit. For this task the PIXEL grader compares the details of each write more strictly than the structured grader; 50 of the 50 structured trials in the final set also satisfy the stricter comparison.

A.2 CONSULT TASKS

The thyroid consult asks for a signed assessment and plan for a thyroid abnormality and for the repeat testing needed to reassess it. The duplicate-anticoagulant consult asks the agent to review the patient's anticoagulation, correct the problem it finds, order the monitoring the correction requires and sign a note. The NSAID consult asks the agent to correct the highest-risk medication issue of an older anticoagulated patient, obtain the necessary safety monitoring and sign a note, without unrelated changes. The depression consult asks for a signed assessment of the available monitoring evidence and permits no other write. The allergy task asks the agent to record a reported severe drug reaction as a coded allergy and to verify that it persists; it has no judge. On the four note-bearing consults the judge scores the signed note one criterion at a time, asking whether it identifies the central clinical finding, respects the task's safety boundaries, justifies its plan and is consistent with the record and with the actions taken.

B MODE PARITY: REQUIREMENTS, DEFECTS AND RERUNS

Table 5 states each parity property at the level of principle, with its status in the reported runs. "Fixed" marks a property that was repaired before or during the study, and "disclosed" marks an asymmetry that remains and is discussed in §8.

The defects behind the "fixed" rows fall into two groups (Table 6). Four were found during the single-trial pilot and repaired before the main run, with the affected pilot cells rerun under a predeclaration. The other five were found by auditing the harness after the main run. Each of those was repaired, the first structured request was checked to change for exactly the intended cells, provider-free controls were run, and all trials of the affected cells, then five per cell, were rerun under a plan frozen before any model call. The three addenda passed 30/30, 30/30 and 100/100 controls respectively, and no valid outcome was retried. Each addendum supersedes the latest earlier source of its cells, so the depression structured trials of the second addendum were themselves superseded by the third, and the 1,500 final trials comprise 500 from the main run, 250 from the addenda and 750 from the extension (trials 6 to 10, Appendix C), which ran only on the repaired configuration. The 350 superseded trials form sensitivity analysis (e) in Table 16. The stage comparison therefore uses trials 1 to 5 of every cell, the only ones that existed before the addenda. Rebuilding that trial set as it stood before any addendum, PIXEL−FHIR was −26.8 points and PIXEL−PACKET+ACTION −35.2; after the first addendum they were −21.6 and −27.2, and the later addenda moved either by at most 0.8 points, to the final −21.6 and −27.6 on those trials.

Table 5: Mode-parity properties for the final trial set.

Property	Status	Principle
Task prompt	Disclosed	Same prompt in every modality; the structured modalities share one added instruction
Source record and clock	Fixed	Same record as of the episode clock, and the clock available in every modality
Patient identity	Fixed	Every modality can reach, and is limited to, the task patient
Available actions	Fixed	Structured actions mirror the UI's operations and admit the same choices, wrong ones included
Pre-save validation	Fixed	The UI's rules, applied to structured actions by the shared executor
Readback	Disclosed	Structured reads return the record as of the clock, not the agent's own writes
Screen driver	Disclosed	Retained history and the handling of undeclared actions differ by provider
Budgets	Disclosed	Time and turn budgets differ by modality (Table 1)
Grader and judge	Equal	Same checks, judge and evidence requirements

Table 6: Parity defects and reruns. "Structured" means both FHIR and PACKET+ACTION. Outcome changes are passes before and after the repair over the rerun trials, or over pilot cells for pilot defects.

Defect	Affected	Rerun	Effect
Allergy tool offered only the correct choices	Structured	Pilot	Structured pilot cells 8/8 → 0/8
Validation stricter than the UI; missing action (NSAID)	Structured	Pilot	FHIR pilot 2/4 → 0/4
Episode clock not given	Structured	Pilot	One pilot cell changed
UI reload did not refetch	PIXEL	None	Not in the frozen consult environment (disclosed)
Incomplete laboratory list (thyroid, NSAID)	Structured	Addendum 1, 100 trials	GPT-6 Astra 13 → 0 of 20; it now orders extra tests
Patient search unusable (BP, eGFR)	PIXEL	Addendum 1, 50 trials	Full evidence coverage 80% → 98% (BP), 76% → 96% (eGFR)
Missing order action (depression)	Structured	Addendum 2, 50 trials	37 → 35 of 50; 0 orders placed
Missing medication actions (thyroid, depression)	Structured	Addendum 3, 100 trials	Thyroid 19 → 21, depression 35 → 36, each of 50; 0 medication actions used
Incomplete protocol action (potassium)	Structured	Addendum 3, 50 trials	50 → 50 of 50

Two of the post-run repairs changed outcomes. With the complete laboratory list, GPT-6 Astra's thyroid and NSAID structured cells fell to zero passes because it now orders additional tests; the first requests differ only in that list, so we treat this as model behaviour rather than a harness fault. The patient-search repair raised evidence coverage under PIXEL on the two retrieval tasks it affected. The remaining three had no detectable effect. No depression trial placed a laboratory order and no thyroid or depression trial used a medication action, so the small changes on those tasks are run-to-run variation, and every potassium trial made one accepted commit of exactly the required writes. A rescan of all ten tasks after the third addendum found no remaining difference in the actions available to each modality that could affect an outcome; two benign interface differences remain.

C PROTOCOL AND DISCLOSURES

Table 7 gives the API identifiers of the evaluated models and the judge. Gemini 3.1 Pro was reached through an aggregator API; all other models were called through their providers' own APIs.

The protocol and analysis plan were frozen and hash-committed before any model call. Every valid trial counts, and none is retried or discarded for its outcome. Attempts that are invalid for a qualified

Table 7: Model identifiers and request settings.

Model	API identifier	Reasoning setting
GPT-6 Astra	`gpt-6-astra`	Effort medium
GPT-5.6 Sol	`gpt-5.6-sol`	Effort medium
Gemini 3.1 Pro	`gemini-3.1-pro`	Provider default
Claude Fable 5.1	`claude-fable-5-1`	Provider default
Claude Opus 5.5	`claude-opus-5-5`	Provider default
Judge: Claude Opus 5	`claude-opus-5`	Provider default

infrastructure reason (provider transport, sandbox start or evidence custody) are rerun at the same trial index at most twice and listed per trial; a trial still invalid after that stops its cell. Attempts ended by a harness defect that was later fixed are recorded and disclosed and do not consume the rerun allowance. Unqualified invalid trials are never rerun: the frozen execution classifier excludes them from denominators, while the primary rule, following the analysis plan, counts them as failures (§5). Before launching, every attempt recomputes a content hash over its task specification, grading materials and runtime files and refuses to start on any difference; prompts and rubrics are unchanged from the pilot. In the main run, which produced the first five trials of each cell, the first stage (trials 1 and 2) runs before the second (trials 3 to 5) for each provider unless the provider is otherwise idle. Structured second-stage trials ran while the first PIXEL stage was blocked, so the stage order held within each modality but not across modalities. Pilot results for trial 1 are reused only where the retained launch matches what the frozen harness would send, which for structured cells means the complete first provider request. Of 46 planned reuses, a harness fix sent the PIXEL ones fresh and the addenda superseded 24, leaving 16. Graders are frozen apart from conformance fixes that make them accept exactly what the frozen contract allows. Every trial can be regraded offline from its retained evidence.

The statistical analysis departs from the frozen plan in one respect. The plan specified a logistic mixed model with random intercepts for task and for task by model, with Wald tests and Holm's correction over ten per-model contrasts. Once the outcomes were known, the size of the PIXEL effect was seen to vary strongly across tasks, which that model treats as noise, and we report instead the exact task-level sign-flip tests, the task-clustered GEE and a mixed model with task-by-modality slopes (§5, Table 13). The plan-specified model gives pooled odds ratios in the same direction, further from one and with narrower intervals; the analysis scripts retain it.

Several fixes preceded the reported runs in addition to the parity repairs of Appendix B. The duplicate-anticoagulant task was built after an audit of an earlier substitute task, and eight of its structured pilot cells were superseded when the episode clock was added. The thyroid pilot was run as successive whole-plan restarts, and two valid pilot outcomes that were rerun in error were restored to the first valid attempt; all of these are disclosed. Opus was added before any Opus call, and its consult PIXEL environment differs from the other models' only in the model route, which four provider-free controls confirmed. In the first addendum 10 transport reruns over 7 trials, all during the launch burst and coinciding with a host DNS failure, resolved within the rerun allowance; the second and third addenda needed none.

The main run also required the following disclosures. Several evidence-capture defects were found and fixed during the run, among them storage and timeout limits; affected attempts are recorded as harness defects and rerun without consuming the allowance. Twice the packing timed out on episodes of about 1.4 GB. Their evidence showed deadline cut-offs, and it was read before a later cleanup deleted it, so that read is the only record of the two outcomes. In the main run, requests that started at, or were cut off by, the two-hour PIXEL deadline were recorded as provider errors and rerun as transport failures under the frozen classification. In 4 potassium PIXEL trials every allowed attempt ended this way, and these trials stop their cells and count as budget failures under the primary rule. Three provider-side refusals, from a terms-of-service stop category rather than a clinical refusal, and one malformed action ended PIXEL attempts and were recorded as harness errors, which makes them unqualified under the frozen rule and failures under the primary rule. Consult PIXEL custody rejects the readback of a reopened historical note, and such trials were rerun as custody failures; these reruns are correlated with a model's tendency to reopen historical notes and are included in the rerun counts of Table 8. Both Anthropic models requested an undeclared triple-click action

from the function-calling PIXEL driver; each such request was rejected, the episode ended after a fixed number of rejections, and the driver was not changed during the run. Format-only grading relaxations of the retrieval answers are confined to sensitivity analysis (b). The clinical grading rules implement the physician-reviewed task content and were kept as frozen; the implemented thresholds of three of them warrant further independent clinical review (§8). Remaining asymmetries are listed in Table 5.

Trials 6 to 10 of every cell (750 trials) form an extension run after the parity addenda, under a plan frozen and hash-committed before its first model call. Each extension trial runs on its cell's final binding: the same prompts, rubrics, graders, judge, model route and outcome rules, and the same worker runtime, so cells whose final trials had run on an earlier runtime kept it. Every attempt verified that binding at launch, and 34/34 provider-free controls passed. The 794 attempts include 24 qualified infrastructure reruns (17 custody, 7 sandbox) over 18 trials. A further 22 attempts were lost to two host DNS outages, which failed name resolution before any provider response; they are recorded as harness defects, consume no rerun allowance and decide no trial, and later launches waited until the provider and sandbox names resolved. A trial is decided by its first finished attempt that is neither a qualified infrastructure failure nor a harness defect. 12 trials had a harness-defect attempt, and letting that attempt decide them as failures would turn 6 passes into failures and lower pass@1 by at most 0.6 points in any modality; an audit found 0 published trials that depart from the rule, and offline regrading reproduced all 12/12 of their deciding attempts. Two PIXEL thyroid trials of GPT-6 Astra used all three attempts on custody failures of a known intermittent note-readback defect; these two stopped trials count as failures under the primary rule and are excluded by sensitivity (a). One GPT-6 Astra PIXEL spirometry trial ended when the model returned an action call with no action, which the extension adjudicates as a protocol failure graded on the final chart and does not rerun. Unlike the main run, the extension counts a request cut off by the episode deadline as a deadline failure, not a transport failure, when the last provider response arrived in the second half of the episode; 11 trials ended this way, all in the PIXEL potassium cells where the main run's cut-offs had stopped 4 trials. No valid outcome was retried, and offline regrading of the 737 extension trials with a gradable final state found 0 disagreements with the recorded grades. Pooled pass@1 over all 10 trials differs by at most 0.4 points from the five-trial estimate in each modality, and trials 6 to 10 alone give 32.0%, 53.6% and 59.6% under PIXEL, FHIR and PACKET+ACTION.

Table 8: Trials, attempts and exclusions per model and modality in the final set. Rerun: trials with more than one attempt (infrastructure or harness-defect reruns; no valid outcome was retried). Adjudicated: the 62 adjudicated failures of Table 9, failures under the primary rule and excluded under the frozen rule. Stopped: trials whose every allowed attempt was infrastructure-invalid (included in Adjudicated); 4 are main-run potassium trials recorded as budget failures after deadline cut-offs, and 2 are the stopped extension trials. Reused: pilot trial-1 results. Addendum: trials from the parity addenda, all among trials 1 to 5; trials 6 to 10 come from the extension.

Modality	Model	Trials	Attempts	Rerun	Adjudicated	Stopped	Reused	Addendum
PIXEL	GPT-6 Astra	100	116	11	4	2	0	10
	GPT-5.6 Sol	100	104	3	0	0	0	10
	Gemini 3.1 Pro	100	116	11	2	0	0	10
	Claude Fable 5.1	100	118	12	16	1	0	10
	Claude Opus 5.5	100	118	14	40	3	0	10
FHIR	GPT-6 Astra	100	102	2	0	0	2	20
	GPT-5.6 Sol	100	102	3	0	0	2	20
	Gemini 3.1 Pro	100	100	1	0	0	2	20
	Claude Fable 5.1	100	99	1	0	0	2	20
	Claude Opus 5.5	100	101	1	0	0	0	20
PACKET+ACTION	GPT-6 Astra	100	98	0	0	0	2	20
	GPT-5.6 Sol	100	101	1	0	0	2	20
	Gemini 3.1 Pro	100	99	1	0	0	2	20
	Claude Fable 5.1	100	98	0	0	0	2	20
	Claude Opus 5.5	100	107	3	0	0	0	20
All		1,500	1,579	64	62	6	16	250

The paper reports three counts of trials that did not end with the agent completing its work, and Table 9 reconciles them. The 62 adjudicated failures are the trials the trial list adjudicates (§5): they fail under the primary rule and are excluded by sensitivity (a). The 60 driver-bound exclusions are the episodes that ended on the rejection limit, the episode deadline or a turn or output-token

limit, which the bound of Appendix E removes. The 65 non-normal terminations are every episode that ended in any of the listed ways, which is the union of the other two sets. The adjudicated failures include 5 trials that the bound keeps as failures (the refusal, the harness errors and the stopped trials), and the driver-bound exclusions include 3 main-run deadline cut-offs that the frozen classifier recorded as valid outcomes and graded on the final chart.

Table 9: Reconciliation of the termination counts, by what ended the episode and how the trial list records the trial. A number is the count of trials in that row that each set includes. Every trial in the table is a PIXEL trial; no structured trial falls in any row. "Budget failure" is the trial list's label for a trial that exhausted a budget; it also covers one harness error.

Episode ended on	Recorded as	Adjudicated	Driver bound	Non-normal
Rejection limit	Budget failure	35	35	35
Episode deadline	Budget failure (main run)	9	9	9
Episode deadline	Deadline failure (extension)	11	11	11
Episode deadline	Valid outcome, graded	0	3	3
Output-token limit	Budget failure	1	1	1
Turn limit	No-completion failure	1	1	1
Provider refusal	Refusal failure	1	0	1
Harness error (malformed action)	Budget failure	1	0	1
Harness error (empty action)	Protocol failure	1	0	1
Stopped by infrastructure	Stopped	2	0	2
Total		62	60	65

D PER-TASK AND PER-MODEL TABLES

Table 10: pass@1 (%) per task, pooled over models (primary rule).

	BP	eGFR	Labs	Spiro.	K^+	Thyroid	Anticoag.	NSAID	Depr.	Allergy
PIXEL	70.0	78.0	54.0	34.0	8.0	10.0	16.0	0.0	38.0	16.0
FHIR	80.0	74.0	76.0	78.0	100.0	34.0	30.0	0.0	68.0	0.0
PACKET+ACTION	100.0	96.0	76.0	76.0	100.0	44.0	28.0	10.0	70.0	0.0

Table 11: Successes over counted trials per task, model and modality (primary rule; k=10). Models: A = GPT-6 Astra, S = GPT-5.6 Sol, G = Gemini 3.1 Pro, F = Claude Fable 5.1, O = Claude Opus 5.5. Wilson 95% intervals (%) for c successes in ten trials: 0: [0, 28]; 1: [2, 40]; 2: [6, 51]; 3: [11, 60]; 4: [17, 69]; 5: [24, 76]; 6: [31, 83]; 7: [40, 89]; 8: [49, 94]; 9: [60, 98]; 10: [72, 100].

	PIXEL					FHIR					PACKET+ACTION				
Task	A	S	G	F	O	A	S	G	F	O	A	S	G	F	O
BP	9/10	6/10	0/10	10/10	10/10	10/10	10/10	0/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10
eGFR	10/10	9/10	0/10	10/10	10/10	10/10	10/10	2/10	5/10	10/10	10/10	10/10	8/10	10/10	10/10
Conflicting labs	6/10	2/10	0/10	9/10	10/10	10/10	10/10	10/10	0/10	8/10	10/10	9/10	10/10	0/10	9/10
Spirometry	0/10	2/10	0/10	8/10	7/10	10/10	10/10	10/10	0/10	9/10	10/10	8/10	10/10	0/10	10/10
Potassium	2/10	2/10	0/10	0/10	0/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10	10/10
Thyroid	5/10	0/10	0/10	0/10	0/10	0/10	0/10	0/10	10/10	7/10	0/10	2/10	0/10	10/10	10/10
Duplicate anticoag.	8/10	0/10	0/10	0/10	0/10	4/10	0/10	0/10	6/10	5/10	0/10	0/10	0/10	10/10	4/10
NSAID safety	0/10	0/10	0/10	0/10	0/10	0/10	0/10	0/10	0/10	0/10	0/10	2/10	0/10	0/10	3/10
Depression	8/10	5/10	0/10	2/10	4/10	10/10	10/10	0/10	6/10	8/10	10/10	10/10	1/10	7/10	7/10
Allergy	0/10	1/10	6/10	1/10	0/10	0/10	0/10	0/10	0/10	0/10	0/10	0/10	0/10	0/10	0/10

Table 12: Per-model pass@1 over all tasks and pass^k ($k = 10$) by modality, with 95% task-clustered bootstrap intervals (primary rule).

Model	pass@1, all modes	pass^k PIXEL	pass^k FHIR	pass^k PACKET+ACTION
GPT-6 Astra	57.3 [33.3, 79.7]	10.0 [0.0, 30.0]	60.0 [30.0, 90.0]	60.0 [30.0, 90.0]
GPT-5.6 Sol	49.3 [25.7, 72.3]	0.0 [0.0, 0.0]	60.0 [30.0, 90.0]	40.0 [10.0, 70.0]
Gemini 3.1 Pro	29.0 [12.3, 46.7]	0.0 [0.0, 0.0]	30.0 [0.0, 60.0]	40.0 [10.0, 70.0]
Claude Fable 5.1	48.0 [28.3, 67.0]	20.0 [0.0, 50.0]	30.0 [0.0, 60.0]	50.0 [20.0, 80.0]
Claude Opus 5.5	60.3 [38.0, 81.0]	30.0 [0.0, 60.0]	30.0 [0.0, 60.0]	50.0 [20.0, 80.0]
All	48.8 [30.4, 66.1]	12.0 [0.0, 26.0]	42.0 [20.0, 62.0]	48.0 [26.0, 70.0]

Table 13: Task-level inference on the pooled contrasts (points of pass@1). Exact two-sided sign-flip p over all 2^{10} sign assignments to the ten tasks; tasks lower/higher counts tasks with a nonzero difference; the leave-one-task-out column gives the range of the estimate with each task dropped. Floor tasks (NSAID and allergy, no cell above four of ten passes) are dropped in the next column, with 2^9 assignments. GEE: logistic, mode and model as covariates, exchangeable working correlation (0.43) clustered on task, Mancl–DeRouen standard errors; odds ratios are population-averaged. The mixed model has a task intercept and task-by-modality slopes (standard deviations 2.6 and, for PIXEL, 2.3 on the logit scale) and is fitted by variational Bayes; its conditional odds ratios are a check on direction only, because its posterior intervals are too narrow to report.

Contrast	Est.	p	Lower/higher	Leave one out	No floor (p)	GEE OR [95% CI]	GEE p	Mixed OR
PIXEL − FHIR	−21.6	0.031	7/2	[−25.8, −13.8]	−24.0 (0.031)	0.41 [0.15, 1.07]	0.069	0.27
PIXEL − PACKET+ACTION	−27.6	0.008	9/1	[−32.4, −20.4]	−29.6 (0.012)	0.31 [0.11, 0.91]	0.032	0.13
PACKET+ACTION − FHIR	6.0	0.109	2/5	[4.2, 6.9]	5.6 (0.219)	1.31 [0.88, 1.94]	0.184	2.12

Table 14: Per-model paired differences (points) with exact sign-flip p over the ten tasks, raw and Holm-adjusted over all fifteen model-by-contrast tests. With ten tasks the smallest attainable two-sided p is $2/2^m$ for m tasks with a nonzero difference, so these tests have little power. Per-model GEE fits with the same Holm adjustment, which rely on large-sample approximations with ten clusters, leave one contrast significant: GPT-5.6 Sol, PIXEL vs. PACKET+ACTION (OR 0.24 [0.10, 0.55], Holm $p = 0.012$).

	PIXEL − FHIR			PIXEL − PACKET+ACTION			PACKET+ACTION − FHIR	
Model	Diff.	p	Holm	Diff.	p	Holm	p	Holm
GPT-6 Astra	−16.0	0.344	1.000	−12.0	0.531	1.000	1.000	1.000
GPT-5.6 Sol	−33.0	0.047	0.609	−34.0	0.012	0.176	1.000	1.000
Gemini 3.1 Pro	−26.0	0.188	1.000	−43.0	0.047	0.609	0.250	1.000
Claude Fable 5.1	−7.0	0.766	1.000	−17.0	0.453	1.000	0.250	1.000
Claude Opus 5.5	−26.0	0.094	1.000	−32.0	0.031	0.438	0.344	1.000

Table 15: Unbiased pass@k (Chen et al., 2021) and pass^k (%) from ten trials per cell, averaged over task–model cells, with 95% task-clustered bootstrap intervals.

	pass@3	pass^k, $k = 3$	pass@10	pass^k, $k = 10$
PIXEL	43.6 [26.8, 60.7]	21.0 [7.0, 37.6]	52.0 [34.0, 70.0]	12.0 [0.0, 26.0]
FHIR	60.2 [37.9, 79.8]	46.9 [25.9, 68.0]	62.0 [40.0, 82.0]	42.0 [22.0, 64.0]
PACKET+ACTION	65.8 [44.1, 85.6]	53.9 [31.3, 75.7]	70.0 [48.0, 88.0]	48.0 [26.0, 70.0]
All	56.6 [37.7, 73.7]	40.6 [23.0, 58.1]	61.3 [42.7, 78.0]	34.0 [18.0, 50.7]

Table 16: Sensitivity analyses: pooled pass@1 (%) and paired differences (points) with 95% task-clustered bootstrap intervals. (a) the 62 adjudicated failures of Table 9 excluded (the frozen classification); (b) format-tolerant grading; (c) a weaker allergy verification status accepted; (d) no-completion trials whose final chart state passes counted as passes; (b+c+d) all three; (e) the trial set before the parity addenda.

Rule	PIXEL	FHIR	PACKET+ACTION	PIXEL−FHIR	PIXEL−PACKET+ACTION
Primary	32.4 [17.2, 49.2]	54.0 [32.2, 73.8]	60.0 [37.4, 80.8]	−21.6 [−40.4, −5.4]	−27.6 [−44.8, −12.2]
(a)	35.9 [20.2, 52.2]	54.0 [32.2, 73.8]	60.0 [37.4, 80.8]	−18.4 [−35.8, −3.2]	−24.3 [−40.4, −9.4]
(b)	36.6 [17.6, 58.6]	59.8 [35.4, 82.4]	65.0 [39.8, 88.0]	−23.2 [−43.2, −5.2]	−28.4 [−48.0, −10.2]
(c)	40.4 [21.8, 60.4]	61.6 [42.8, 78.0]	69.6 [50.0, 87.0]	−21.2 [−40.4, −4.4]	−29.2 [−45.4, −16.0]
(d)	32.6 [17.6, 49.2]	54.0 [32.2, 73.8]	60.0 [37.4, 80.8]	−21.4 [−40.4, −4.8]	−27.4 [−44.8, −11.6]
(b+c+d)	44.8 [23.4, 68.0]	67.4 [45.8, 86.2]	74.6 [53.0, 93.2]	−22.6 [−43.0, −4.0]	−29.8 [−48.4, −12.8]
(e)	31.0 [17.0, 46.2]	55.2 [33.6, 74.4]	62.4 [40.4, 81.8]	−24.2 [−42.6, −8.0]	−31.4 [−47.8, −16.0]

	Astra			Sol			Gemini			Fable			Opus		
Task	Pix.	FHIR	Pkt.	Pix.	FHIR	Pkt.	Pix.	FHIR	Pkt.	Pix.	FHIR	Pkt.	Pix.	FHIR	Pkt.
Retrieval															
Potassium conflict	2	10	10	2	10	10	·	10	10	·	10	10	·	10	10
Conflicting labs	6	10	10	2	10	9	·	10	10	9	·	·	10	8	9
Spirometry	·	10	10	2	10	8	·	10	10	8	·	·	7	9	10
Blood pressure	9	10	10	6	10	10	·	·	10	10	10	10	10	10	10
eGFR	10	10	10	9	10	10	·	2	8	10	5	10	10	10	10
Action-bearing															
Thyroid follow-up	5	·	·	·	·	2	·	·	·	·	10	10	·	7	10
Duplicate anticoagulant	8	4	·	·	·	·	·	·	·	·	6	10	·	5	4
NSAID + anticoagulant	·	·	·	·	·	2	·	·	·	·	·	·	·	·	3
Depression follow-up	8	10	10	5	10	10	·	·	1	2	6	7	4	8	7
Coded allergy	·	·	·	1	·	·	6	·	·	1	·	·	·	·	·

Passes out of 10: · 1 2 3 4 5 6 7 8 9 10 Pix. = Pixel, Pkt. = Packet+Action

Figure 3: Passes out of ten trials in each of the 150 cells (ten tasks, five models, three modalities; 1,500 trials, primary rule).

E ROBUSTNESS OF THE MODALITY EFFECT

These analyses are secondary and were chosen after the outcomes were known. None changes a trial's outcome, and the bootstrap intervals below use 2,000 task-clustered resamples.

The judge is invoked only on consult trials whose hard gates all pass, so it can remove passes but cannot create failures of access or persistence. If its verdicts are ignored entirely, pass rates are 35.4% (PIXEL), 61.4% (FHIR) and 67.0% (PACKET+ACTION), and PIXEL−FHIR widens to −26.0 points [−45.4, −7.4]. Of the 261 trials that reached the judge, it passed 32 of 47 under PIXEL, 66 of 103 under FHIR and 76 of 111 under PACKET+ACTION. The judge, Claude Opus 5, passed 99 of 122 judged trials from the two Anthropic models (81.1%; 79.1% after standardising to the task mix) and 74 of 88 from the two OpenAI models (84.1%; 79.8%), so we see no preference for its own vendor. Gemini 3.1 Pro passed 1 of its 51 judged trials in any modality; the judge's cited evidence points to content errors rather than to style.

The PIXEL driver bound excludes every episode that ended on the rejection limit, the episode deadline (including cut-offs recorded as provider errors, whether or not the trial was adjudicated), or a turn or output-token limit. There are 60 such driver-bound exclusions, all under PIXEL; refusals, harness or protocol errors and stopped trials stay in the bound as failures (Table 9). Excluding the driver-bound exclusions in every modality assumes that they would have passed at the rate of the model's other episodes, so it is a bound rather than an estimate. It raises pooled PIXEL pass@1 to 36.8% and moves PIXEL−FHIR from −21.6 to −17.2 points [−35.1, −1.4] and PIXEL−PACKET+ACTION to −23.2 points [−39.5, −8.7]. Per model, PIXEL pass@1 moves from 41.0% to 68.3% for Claude Opus 5.5 (40 episodes excluded; FHIR 67.0%), from 40.0% to 46.5% for Claude Fable 5.1, from 6.0% to 6.3% for Gemini 3.1 Pro and from 48.0% to 48.5% for GPT-6 Astra, and GPT-5.6 Sol is unchanged at 27.0%. In the extreme case in which every excluded episode would have passed, PIXEL−FHIR is −9.6 points [−25.0, 6.4]. Before any exclusion, the gap is −19.7 points for the three models on the function-calling driver and −24.5 points for the two on the OpenAI native driver; driver and model are confounded, so this comparison cannot isolate the driver.

The limits behind these endings were fixed in the frozen plan and were the same for every model: a PIXEL episode deadline of 7,200 s with no turn cap, except on the allergy task (1,800 s, 80 turns and 160 actions), and a limit on rejected actions that exists only in the function-calling driver used for the Claude and Gemini models (Table 1). We classified each limit-based ending from its trajectory (Table 18). Of the 23 deadline endings, 20 were repetitive cycles that had stopped making progress: 18 Claude potassium episodes cycled between the same laboratory-result views, 7 of them after committing the correct chart changes without ever returning the required answer, and 2 Gemini episodes repeatedly re-filed notes and orders. Runaway generation dominated another 2, and 1 was a slow run that was still progressing, its trajectory resembling those of its passing sibling trials. All 35 rejection-limit endings were still progressing when the cap on rejected calls to an undeclared triple-click action ended them, and 6 of them were near completion. Sibling trials of the same cells suggest that about 3 to 9 of the 500 PIXEL trials, at most 1.8 points of pass@1, might have passed with more time or without the cap. That is well inside the bound above, which drops every such trial.

Table 17 lists the asymmetries that remain after the parity addenda and the direction in which each would bias the PIXEL deficit.

Table 17: Residual asymmetries between modalities and the direction of their bias on the PIXEL deficit. "Neutral" means the evidence shows no effect on any outcome.

Asymmetry	Direction	Evidence
Off-task page reachable on every PIXEL task	Neutral	Not used in any PIXEL trial
Signed-note readback: established from the trajectory under PIXEL, by an explicit reopen in the structured modalities	May inflate	Fails in 86 PIXEL trials (28 as the only failed check) and 0 structured trials; at most 5.6 points of PIXEL pass@1
Stricter comparison of the PIXEL potassium writes	Neutral	0 of 46 PIXEL potassium failures fail on it alone, and every structured pass also meets it
Potassium evidence forms, validation and order lists	Neutral	Each grader reads its own form of the same audited targets; order lists made equal in the third addendum
One fixed field in the PIXEL consult order form	Neutral	Same choices as the structured order tool
Shorter history in the function-calling driver (Claude, Gemini) than in the OpenAI driver	Inflates for Claude and Gemini	All 23 PIXEL deadline terminations (Table 9) are Claude or Gemini episodes; gap −19.7 against −24.5 points, −12.6 once terminations are excluded
Non-native computer tool with a rejection limit (Claude, Gemini)	Inflates	35 rejection-limit terminations (Table 9), 35 of them Anthropic, driven by an undeclared triple-click action
Generic feedback on an undeclared action	Inflates, mainly for Opus	Claude Opus 5.5 PIXEL 41.0% → 68.3% under the bound; FHIR 67.0%
Episode deadline and turn budgets	Inflates against a no-budget estimand; neutral for the route as built	All 65 non-normal terminations (Table 9) are PIXEL episodes; the bound gives −17.2 points

Table 18: Limit-based endings in the driver bound, by the state of the episode when the limit ended it; every row is a PIXEL episode, and the totals by ending match Table 9. Progressing includes the one slow deadline run; near completion means at most one required write was left, or every write was done but completion was never declared.

Episode ended on	No-progress cycle	Progressing	Near completion	Runaway generation	Total
Episode deadline	20	1	0	2	23
Rejection limit	0	29	6	0	35
Turn limit	0	0	1	0	1
Output-token limit	0	0	0	1	1
Total	20	30	7	3	60

F JUDGE SETTINGS AND QUALIFICATION

The judge is Claude Opus 5 (Table 7) run as a tool-using agent. Each rubric criterion gets its own bounded tool loop, which must end in a structured verdict giving the band, a reason and the cited evidence. There is no fallback judge, and malformed verdicts are not repaired. The judge sees the criterion, the signed note and the patient's source record as it stood before the episode, with identity fields that would reveal the trial removed; the evaluated model, provider and modality are withheld, and it cannot see the agent's trajectory or the live chart. Over the 261 judged trials it returned 1,096 criterion verdicts, with a median of 5.0 API turns each. The judge runs only after every deterministic check has passed. Band scores: EXCELLENT 1, GOOD 2/3, PARTIAL 1/3, FAIL 0; binary pass if the mean is ≥ 0.75 and no criterion is FAIL.

Before any evaluated trial the judge scored five constructed notes per task. A gold note and a paraphrase of it were expected to score at least GOOD on every criterion, a partial note to receive mixed bands without failing uniformly, and a contradictory and an unsafe note to receive at least one FAIL. Every note met its expectation, so all four judged tasks qualified. The allergy task has no judge.

Qualification shows that the judge separates constructed good and bad notes; it does not show agreement with clinicians on agent-written notes. We have not run a clinician agreement study and treat this as a limitation (§8).

G CHECKPOINT-WEIGHTED SCORE

The weighted score $r \in [0, 1]$ marks each deterministic checkpoint and each judge criterion as critical or non-critical. Critical checkpoints (patient identity, safety prohibitions, the required actions and the central clinical finding of the task) carry more weight than supporting detail and also gate the pass: a trial is a weighted pass if and only if every critical checkpoint holds and $r \geq 0.75$. Checks that establish the validity of the trial rather than the agent's behaviour are excluded. We assigned every existing check to one class before any trial was scored under this metric, and no check, threshold or grade was changed.

The critical/non-critical assignment is our own, not the product of a clinician review. The judge scores four of the ten tasks and only after every deterministic check has passed, so the weighted score is mostly deterministic. Trials that ended before the grader ran have no per-checkpoint detail and score $r = 0$, a lower bound. Binary and weighted verdicts differ on 19 + 11 of 1,500 trials, all through judge bands: a PARTIAL on a critical criterion fails a trial that the band mean passed, or weak non-critical criteria no longer fail a trial whose criticals all held. Table 19 gives the metric by modality. By model (Table 20) the weighted score puts GPT-6 Astra highest on r and Claude Opus 5.5 highest on both pass rates, but the intervals overlap and we do not rank models.

Table 19: Checkpoint-weighted score r, weighted pass rate (%) and binary pass@1 (%) by modality (k=10, 1,500 trials), with 95% task-clustered bootstrap intervals; Δ rows are paired differences.

	Weighted r	Weighted pass	Binary pass@1
PIXEL	0.554 [0.433, 0.671]	32.0 [17.0, 48.8]	32.4 [17.2, 49.2]
FHIR	0.786 [0.660, 0.886]	53.2 [32.0, 72.2]	54.0 [32.2, 73.8]
PACKET+ACTION	0.849 [0.730, 0.941]	59.6 [37.8, 79.8]	60.0 [37.4, 80.8]
Δ Pix.−FHIR	−0.232 [−0.378, −0.086]	−21.2 [−40.2, −5.0]	−21.6 [−40.4, −5.4]
Δ Pix.−Pkt.	−0.295 [−0.423, −0.164]	−27.6 [−44.6, −12.6]	−27.6 [−44.8, −12.2]
Δ FHIR−Pkt.	−0.063 [−0.104, −0.023]	−6.4 [−12.4, −1.0]	−6.0 [−11.6, −1.2]

Table 20: Weighted score r, weighted pass rate (%) and binary pass@1 (%) by model, pooled over modalities (k=10, 1,500 trials; 300 trials per model), with 95% task-clustered bootstrap intervals. The leading model differs between metrics and the intervals overlap, so we do not rank models.

Model	Weighted r	Weighted pass	Binary pass@1
GPT-6 Astra	0.801 [0.660, 0.915]	57.3 [33.3, 79.7]	57.3 [33.3, 79.7]
GPT-5.6 Sol	0.791 [0.684, 0.882]	49.3 [26.0, 72.0]	49.3 [25.7, 72.3]
Gemini 3.1 Pro	0.593 [0.522, 0.667]	31.0 [15.3, 47.3]	29.0 [12.3, 46.7]
Claude Fable 5.1	0.721 [0.618, 0.822]	44.0 [23.3, 64.7]	48.0 [28.3, 67.0]
Claude Opus 5.5	0.743 [0.605, 0.872]	59.7 [38.0, 80.3]	60.3 [38.0, 81.0]
All	0.730 [0.631, 0.815]	48.3 [30.4, 65.3]	48.8 [30.4, 66.1]

H EXPLORATORY TRAJECTORY PROCESS CHECKS

This analysis is secondary and exploratory, and it does not alter any primary outcome. Outcome-based grading cannot distinguish an agent that never found the relevant evidence from one that found it and acted wrongly. We retrospectively scored whether each agent consulted the key evidence before acting. Before scoring any trial, we fixed a list of 2–4 key facts per task (24 facts across 9 tasks), taken solely from each task's frozen source contract; the allergy task was excluded because its key evidence is in the prompt. Under FHIR, a fact counts as consulted if its resource type was successfully read before the first write or final answer (searches return all rows for the patient). Under PIXEL, it counts as consulted if its label and value co-occur, by OCR, on a screenshot returned to the model before the first chart write; no DOM or route logs exist. PACKET+ACTION supplies the evidence up front, so coverage is not applicable. Intervals are task-clustered percentile bootstrap intervals.

FHIR agents retrieved all key evidence in 446 of 450 trials, so FHIR failures reflect reasoning, action policy or answer format, not retrieval (Table 21). PIXEL agents reached all key evidence in 73% of trials [53, 90], and coverage strongly separated outcomes: 42% of fully covered trials passed versus 4% of the rest (+39 pp [+15, +60]); only 4 of 127 PIXEL passes lacked detected coverage. Of PIXEL failures, 62% had every key fact on screen and still failed, 34% never reached at least one fact, and 4% reached it only after writing to the chart. Coverage gaps were concentrated by model (46% for Gemini 3.1 Pro, 86% for Claude Opus 5.5) and by item (the latest potassium result was on screen in 44% of trials). The check also registers the patient-search defect repaired in the first parity addendum: in the superseded BP and eGFR PIXEL cells, run with a broken patient search, full coverage was 80% and 76%; after the fix it is 98% and 96%.

A failure with every fact on screen is not necessarily a reasoning failure. For the post hoc decomposition in §7 we assign each failed trial to the first stage it failed, checked in this order: access (a key fact not consulted before the first write or answer), termination (the episode ended on a budget, the rejection limit, the deadline, a refusal, a protocol error or without completion), persistence (a required write missing, mistimed or not read back), a deterministic decision gate, and the judge. The population is nine tasks, all but conflicting labs, whose values the OCR matcher cannot detect; the allergy task, whose evidence is in the prompt, and PACKET+ACTION count as having access by construction, so the shares differ slightly from Table 21. On this population 67% of PIXEL failures had the evidence, and 62% of those ended early or never persisted their writes. Decision and judge failures are less frequent under PIXEL than under FHIR largely because fewer PIXEL trials reach those stages, which is why we compare pass rates conditional on reaching them.

The checks have limitations. They are automated and contract-derived, not clinician-authored, and were defined after the run (though before outcomes were examined). PIXEL coverage depends on OCR: the conflicting-labs task was excluded because its values appear only as graph labels (including it gives +23 pp [−7, +52]). Of 134 "not reached" items, 54 never had their label on any pre-action screen, while 80 are ambiguous between an unexpanded row and an OCR miss. Conclusions are unchanged with the single OCR pass fixed with the check definitions (full coverage 69%, +40 pp) and when restricting to frozen-valid trials (+44 pp), but with 8–9 tasks the intervals are wide. Trajectory-level process checks authored by clinicians, and applied to every modality, are future work.

Table 21: Exploratory trajectory process checks (final trial set; the 62 adjudicated failures of Table 9 count as failures). Full coverage: every key fact consulted before the first write or answer. Brackets: task-clustered 95% bootstrap intervals. PIXEL excludes one task whose values the OCR matcher cannot detect.

Mode	n	Pass	Full coverage	Passes w/o full coverage	Fails with full coverage	Pass: full vs. incomplete
FHIR	450	60%	99% [98, 100]	0/270	176/180	61% vs. 0% (n=4)
PIXEL	400	32%	73% [53, 90]	4/127	168/271	42% vs. 4%
PACKET+ACTION	450	67%	n/a	n/a	n/a	n/a

I REPRODUCIBILITY AND ARTEFACTS

Each cell is bound by content hashes to its task specification, its grading materials and every runtime file of the PIXEL worker. An attempt recomputes this binding from the code it is about to run and refuses to start on any difference.

Each trial retains the baseline and final chart captures, the write receipts, the audit events, the reset proof and the full provider transcript, together with the screenshot trajectory for PIXEL. Grading is a pure function of this evidence. Every reported trial has been regraded offline from it, and the alternative grading rules discussed in §8 can be applied without new model calls.

A single script produces every number in the paper from the per-trial results and from the weighted-score, judge-audit, process-check and parity-addendum outputs derived from the same evidence. These numbers include pooled and per-model pass@1 and pass^k, paired differences and task-clustered bootstrap intervals (10,000 draws, seed 20260924, 2.5th and 97.5th percentiles) under every sensitivity rule, unbiased pass@k, Wilson intervals and the rerun table. The exact sign-flip

tests, GEE and mixed-model fits, the judge and driver analyses and the failure-stage decomposition come from two further offline scripts over the same trial table, and the paper reads their outputs through the same generator. A second script builds the failure taxonomy from the graded evidence, and one build command regenerates both outputs and the paper.

The task specifications, rubrics and grading materials are proprietary, and no data, code or other materials are released with this submission. On request through the area chair, we will consider sharing materials, including the per-trial outcome table, confidentially with reviewers, in a form that preserves anonymity.