χ-Bench: Can AI Agents Automate End-to-End, Long-Horizon, Policy-Rich Healthcare Workflows?

NEURAL INFORMATION PROCESSING SYSTEMS 2026

NeurIPS 2026 Evaluations & Datasets Track

Haolin Chen[1], Deon Metelski[1], Leon Qi[1], Tao Xia[1], Joonyul Lee[1], Steve Brown[1], Kevin Riley[1], Frank Wang[1], T. Y. Alvin Liu, MD[2], Hank Capps, MD[3], Zeyu Tang[4], Xiangchen Song[5], Lingjing Kong[5], Fan Feng[6], Tianyi Zeng[7], Zhiwei Liu[8], Zixian Ma[9], Hang Jiang[10], Fangli Geng[11], Yuan Yuan[12], Chenyu You[13], Qingsong Wen[14], Hua Wei[15], Yanjie Fu[15], Yue Zhao[16], Carl Yang[17], Biwei Huang[6], Kun Zhang[5,19], Caiming Xiong[18], Sanmi Koyejo[4], Eric P. Xing[19,5], Philip S. Yu[20], Weiran Yao[1]

1 Actava AI; 2 Johns Hopkins Medicine; 3 Wellstar Health System; 4 Stanford University; 5 CMU; 6 UCSD; 7 Yale School of Medicine; 8 Salesforce AI Research; 9 University of Washington; 10 Northeastern University; 11 Brown University; 12 Boston College; 13 Stony Brook University; 14 University of Oxford; 15 Arizona State University; 16 University of Southern California; 17 Emory University; 18 Recursive Superintelligence; 19 MBZUAI; 20 University of Illinois at Chicago

Can agents complete end-to-end healthcare workflows?

Healthcare workflows combine policy-grounded decisions, role handoffs and conversations. Agents must interpret medical and insurance rules, coordinate clinicians and reviewers, and gather information through patient outreach or doctor-to-doctor review.

Strong coding performance leaves open whether agents can complete these healthcare workflows end to end. A submitted step may be irreversible: the next role inherits its records and artifacts. χ-Bench tests whether the linked actions reach a valid, documented case outcome.

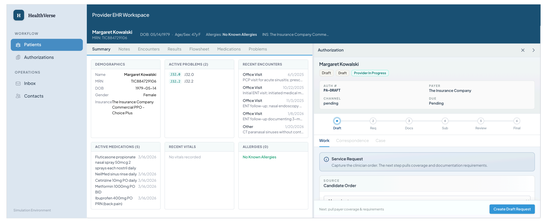
01 χ-World Environment

χ-World is a stateful simulation environment for Clinical Healthcare In-situ (CHI) workflows.

20	87	1,290	~5,000
healthcare apps	MCP tools	Skills library Markdown documents	chart activities

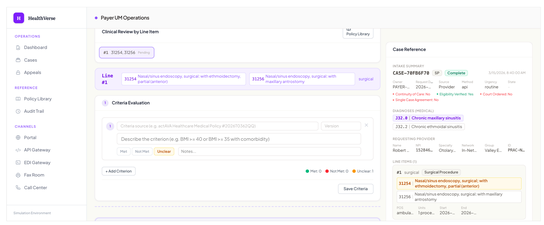
Provider Prior Authorization (PA)

Providers gather records and ask insurers to approve treatment.



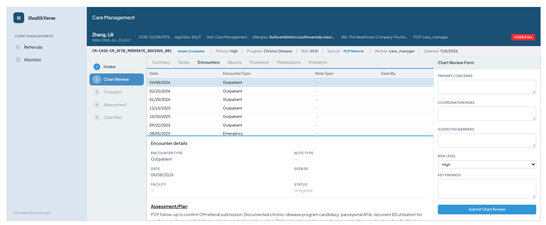
Payer Utilization Management (UM)

Insurers review the records and decide whether to cover treatment.



RN Care Management (CM)

Nurses talk with patients, assess their needs and plan ongoing care.



Managed-Care Operations Handbook Skill

Developed with clinicians and operations leaders at Johns Hopkins Medicine.

SKILL.md: role-based entry point

- **provider-pa/** PA workflow and forms
- **payer-um/** Review and decisions
- **care-manager/** Assessments and care plans
- **medical-library/** Policies, drug criteria and guidelines
- **platform/** Tools, objects and role boundaries

02 Benchmark Design

Task Architecture and Agent Interface

instruction.md (excerpt) — UM nurse-review task

```
Provider faxed a letter of medical necessity overnight; case is at
nurse review. Nurse clinical review queue for a 46-year-old female
with Chronic maxillary sinusitis; Chronic ethmoidal sinusitis; Robert
Chen MD requested Functional endoscopic sinus surgery – partial
anterior ethmo. (Intake was recorded; triage routed the case.) Work
the documentation review, record the recommendation, then carry the
case through MD decision, any P2P, and finalize the determination
yourself.
```

`/logs/artifacts/workspaces/CASE-70FB6F70/` — UM case CASE-70FB6F70

payer/incoming_request/
- 01_ent-office-visit-note-initial-evaluation.pdf
- 02_ct-paranasal-sinuses-without-contrast.pdf
- 03_ent-follow-up-visit-treatment-failure-documentation.pdf
- prior-auth-request-form.pdf

payer/ Working files
shared/handoff/ Shared handoff

healthverse MCP Case state and tools

Frontier Agents — Harness + model

Harness: Models

χ-World: shared state and tools — Selected apps; 3 MCP servers

Provider PA	Patient Chart	Coverage	Authorizations
Payer UM	Clinical Review	Policy Library	Letter Center
RN CM	Worklist	Outreach	Care Plan

Task Construction and Composition

Fixed conditions — UM authoring example

Policy TIC-UM-OPS-003 §4.3
Criterion not_met or unclear
Nurse escalates to physician review

Fixed terminal outcome
Status **denied**
Letter **denial**

State graph: Nurse review → MD review → Denied

Backward generation — Generate upstream from the fixed outcome

Claude Opus 4.7 → **Chart**: Nasal symptoms; 6-week history; Duration unclear | **Request**: Sinus surgery; PA request; Supporting notes | **Hidden rubric**: Continuous >12 weeks; not_met

523 candidates — Agent inputs — Evaluation only

Review — Across PA, UM and CM

Automatic checks
- ☐ Policy grounding; no leakage
- ☐ Chart–rubric entailment
- ☐ Reference dry-run

172 candidates

Human review
UI walkthrough; clinical review
Privacy and realism
75 tasks; 25/domain

Reward and Verification

Deterministic contract

```
{
  "expected_target_status":
    "denied",
  "expected_letter_types": [
    "denial"
  ],
  "expected_role_compliance": true
}
```

Rubric LLM judge

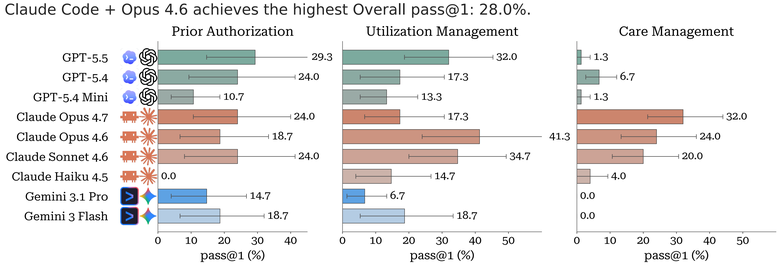
Did the reviewer correctly assess this criterion for line 1: Chronic rhinosinusitis with symptoms present for more than 12 consecutive weeks?

Expected semantic assessment
`expected_result: "not_met"`

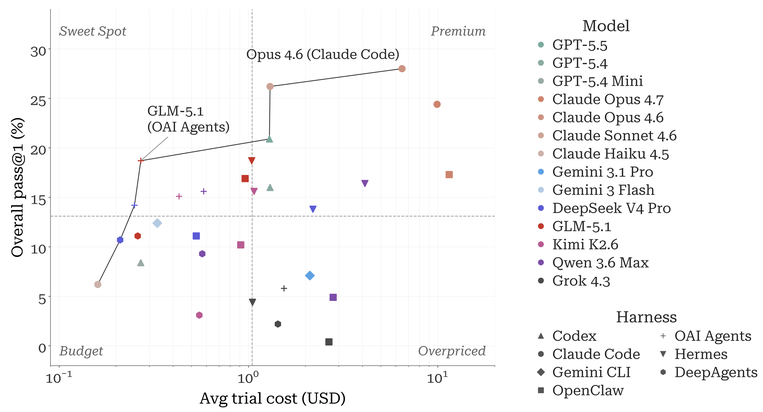
In-situ reward: R = DeterministicPass ∧ JudgePass

03 Results

30 model–harness configurations are evaluated on 75 tasks, with 3 trials per task.
Claude Code + Opus 4.6 achieves the highest Overall pass@1: 28.0%.



ROI



Experiment Variants

Success rates: pass@1 (%)

Arena
Both agents must pass all checks on 23 PA tasks.

Provider agent → Payer agent

Agent	Codex + GPT-5.5 (both roles)	
Metric	**Provider only**	**Provider + payer**
pass@1 (%)	30.4	0.0
Physician review unfinished		18/23 tasks
Doctor-to-doctor requests unfinished		5/23 tasks

Marathon
Marathon chains 25 tasks/session (3 sessions/domain/config); per-task runs reset.

Task 1 → Task 2 → … → Task 25

Agent	Setting	PA	UM	CM
Codex + GPT-5.5	Marathon	8.0	2.7	0.0
	Per-task	29.3	32.0	1.3
Claude Code + Opus 4.7	Marathon	8.0	1.3	2.7
	Per-task	24.0	17.3	32.0

MCP versus CLI
CLI exposes the same tools via MCPorter (75 tasks × 3 trials).

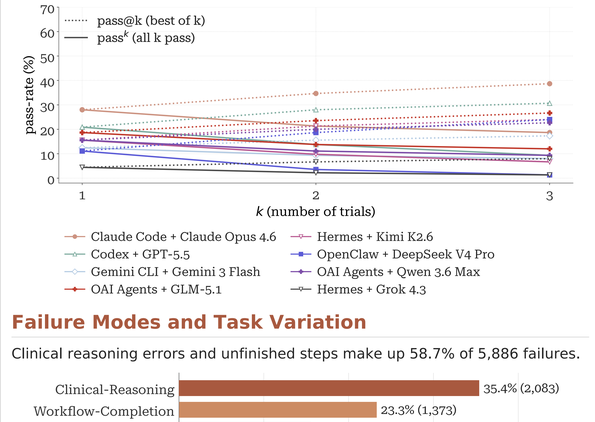
MCP vs CLI

Agent	Codex + GPT-5.5		
Setting	**PA**	**UM**	**CM**
MCP	29.3	32.0	1.3
CLI	28.0	25.3	4.0

04 Analysis

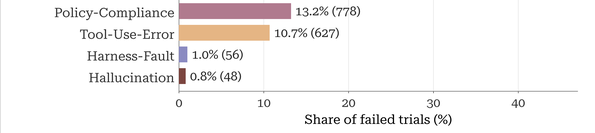
Reliability Degradation

Claude Code + Opus 4.6 succeeds on 28.0% of attempts (pass@1); 18.7% of tasks succeed in all 3 attempts (pass^3).



Failure Modes and Task Variation

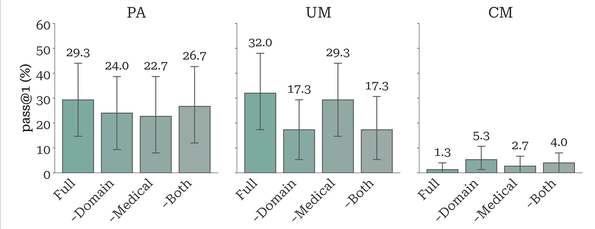
Clinical reasoning errors and unfinished steps make up 58.7% of 5,886 failures.



Across 30 configurations, UM pass@1 is 28.1% on 9 Easy tasks and 5.9% on 7 Hard tasks. For initially refusing CM patients, pass^3 is 0.2% (15 tasks × 30 configurations).

Handbook Ablation

Handbook components are removed for Codex + GPT-5.5 on 75 tasks.



Removing the domain handbook reduces UM pass@1 from 32.0% to 17.3%.

05 Conclusion

χ-World simulates stateful healthcare apps with irreversible role handoffs. χ-Bench grounds 75 tasks in a 1,290-document handbook and verifies saved records. Across 30 model–framework configurations, results expose gaps in reliable end-to-end completion.

1 ACTAVA · 2 Johns Hopkins Medicine · 3 Wellstar · 4 Stanford University · 5 Carnegie Mellon University · 6 UC San Diego · 7 Yale School of Medicine · 8 salesforce · 9 University of Washington · 10 Northeastern University · 11 Brown · 12 Boston College · 13 Stony Brook University · 14 University of Oxford · 15 Arizona State University · 16 USC · 17 Emory University · 18 recursive · 19 Mohamed bin Zayed University of Artificial Intelligence · 20 University of Illinois Chicago

Paper arxiv.org/abs/2605.16679

Code github.com/actava-ai/chi-bench