Claude, GPT, Gemini Agents Fail 72% of U.S. Healthcare Workflows, New Benchmark Finds

May 20, 2026

Pleasanton, CA - May 20, 2026 - PRESSADVANTAGE -

AI company actAVA.ai today released CHI-Bench, the world?s first long-horizon healthcare benchmark for AI agents. Across 75 workflows and 30 frontier agents from Anthropic, OpenAI, Google, x.AI, DeepSeek, and Z.ai, the best-performing agent fails roughly seven out of ten real clinical cases. Code, data, and the live leaderboard are at actava.ai/benchmarks.

AI labs position agents as ready for long workflows, but until now no public benchmark validated that claim in healthcare, where one missed policy check can mean a denied authorization, delayed treatment, or audit finding. Each trial in CHI-Bench runs an agent for 60-80 steps across four to six clinical stages, exposing 21 healthcare apps through 200+ MCP tools and a 1,279-document operations handbook. It evaluates the trajectory, every artifact, and world state using deterministic unit tests and LLM judge for evidence grounding, consent, and cross-stage consistency.

Across the 30 frontier agents tested, Anthropic's Claude Code with Opus 4.6 achieved the best overall

performance at 28% pass@1, followed by OpenAI's Codex with GPT-5.5 at 21%. By domain, utilization review reached 41%, care management 32%, and prior-authorization paperwork 29%. Reliability remained a major issue, with no agent clearing 20% when the same case was run three times. Under endurance testing, where agents were asked to handle 25 cases in one session, the best system completed under 4%. In a fully end-to-end setting, where one AI submitted a prior-auth request and a second acted as the UM reviewer, no task passed successfully.

CHI-Bench was built with a 20+ institution coalition spanning health systems (Johns Hopkins, Wellstar, Yale) and universities (Stanford, CMU, Oxford, USC, UCSD), with world-class AI researchers Caiming Xiong (Recursive Superintelligence), Sanmi Koyejo (Stanford), Eric P. Xing (CMU; MBZUAI), and Philip S. Yu (UIC).

Existing AI benchmarks in healthcare focus on narrow clinical knowledge, such as answering medical exam questions or extracting information from a single document. These tasks test what a model knows, not what an agent can do. Real-world healthcare operations require an agent to navigate multi-step processes that span departments, roles, and systems over extended periods. A utilization-review case, for example, demands reading a physician's clinical notes, applying payer-specific medical policies, querying formulary databases, generating a compliant determination letter, and routing the outcome to the correct downstream team?all without human intervention. CHI-Bench captures this full operational arc, making it possible for the first time to measure whether an AI agent can reliably replace or augment a trained healthcare professional across an entire workflow rather than a single isolated task.

"These workflows are long, role-composed, and gated by policy," said Haolin Chen, lead author. "An agent has to play intake clerk, nurse reviewer, and medical director across sixty-plus steps where one wrong site-of-service flip cascades into multiple failures."

"We need to know whether an agent can carry a real case end-to-end without error," said Weiran Yao, Chief AI Officer of actAVA. "CHI-Bench is built for that."

CHI-Bench is open under Apache 2.0 on GitHub; the leaderboard accepts community submissions today.

About actAVA.ai: actAVA.ai is a specialized AI platform that standardizes and accelerates the creation, training, and deployment of artificial intelligence solutions for healthcare and life sciences. We are a high-speed assembly line that automates the build and deployment of highly tested, always compliant, ever-learning agentic AI. actAVA is the Healthcare AI factory.

Follow: LinkedIn /company/actava · X @actAVAai

###

For more information about actAVA.ai, contact the company here:actAVA.aiMedia Teamresearch@actava.ai4695 Chabot Drive Suite 200, Pleasanton, CA 94588

actAVA.ai

actAVA.ai is a Healthcare AI factory that standardizes and accelerates the creation, training, and deployment of agentic AI solutions for healthcare and life sciences?automating builds that are highly tested, always compliant, and ever-learning.

Website: https://actava.ai/

Email: research@actava.ai