BloodGPT

MedTech AI That Actually Works

TECHNICAL REPORT · APRIL 2026

100% Accuracy on Stanford's Medical AI Benchmark

Both Generations. 3,000 consecutive clinical tasks. Zero failures.
The first reported perfect score on MedAgentBench v2.

300/300
MedAgentBench v1
NEJM AI 2025

300/300
MedAgentBench v2
PSB 2026, first ever

5×
Reproduced runs
per benchmark

How BloodGPT's AI engine processes real medical data, and why it matters for your practice.

BloodGPT Team

HEADLINE RESULT

600 out of 600 clinical tasks completed correctly.

- 300/300 on MedAgentBench v1 (NEJM AI 2025)
- 300/300 on MedAgentBench v2 (PSB 2026, the first reported perfect score)

Previous best, v1	**98.0% — Chen et al., GPT-4.1 + memory**
Claude for Healthcare launch	**92.3% — Claude Opus 4.5, Anthropic Jan 2026**
Previous best, v2	**88.7% — Chen et al., GPT-4.1 + memory**
Original MedAgentBench paper	**69.7% — Claude 3.5 Sonnet v2**
BloodGPT, v1 and v2	**100% / 100% — first reported perfect score**

Reproduced across 5 independent runs per benchmark. 3,000 consecutive tasks, zero failures. Benchmark details: stanfordmlgroup.github.io/projects/medagentbench

Why Should a Doctor or Clinic Care?

MedAgentBench isn't an academic trivia quiz. It's a simulation of what doctors actually do every day, look up a patient's lab results, check if a value is dangerously low, decide whether to order a replacement medication, calculate the right dose, and record it in the chart. All of this happens against a real FHIR-compliant electronic health record system populated with actual de-identified patient data from Stanford Hospital.

When we say "100% accuracy," we mean: every patient lookup, every lab interpretation, every conditional medication order, every dose calculation, every vital sign recording, correct, on the first attempt, with no human in the loop.

Twice. We ran the benchmark across two separate generations, the 2025 original (300 tasks focused on data retrieval and conditional ordering) and the 2026 extension (300 multi-step tasks targeting complex clinical workflows like QTc management, DVT prophylaxis hygiene, and vaccine recall). Same engine, same configuration, no dataset-specific tuning. 300/300 on each.

Why This Matters for Any AI Working with Medical Data

Any AI chatbot or assistant that answers medical questions needs to query a FHIR server, the standard format for storing and exchanging patient health records, to find the relevant data. This benchmark measures how well AI handles the most fundamental task: retrieving data from the health record and presenting it accurately to a

physician. No interpretation yet. Just consistent, error-free data retrieval for any query, as long as the data exists in the system.

The problem is that this seemingly simple task has never been solved to 100% accuracy. AI models make mistakes, and FHIR servers themselves lack sufficient intelligence in their tooling. The result: chatbots occasionally give wrong answers, and, critically, you never know which answers are wrong. This is unacceptable in a clinical setting. Even Anthropic's flagship Claude Opus 4.5, the model powering Claude for Healthcare, tops out at 92.3% on MedAgentBench with extended thinking and native tool use enabled.

Our insight

What if the FHIR server itself included an intelligent layer, so it could work directly with base AI models, with no additional middleware sitting between the AI and the health data? This would eliminate a whole category of failure modes. No product on the market does this today. That is the gap we are solving, at the infrastructure level.

Two Generations of the Same Test

Stanford Machine Learning Group released MedAgentBench v1 in 2025 (NEJM AI), with 300 physician-authored tasks across 10 categories, patient lookup, lab interpretation, conditional medication ordering, vital sign recording. Best prior result: GPT-4.1 + memory + few-shot examples reached 98%. When Anthropic launched Claude for Healthcare in January 2026, they reported Claude Opus 4.5 at 92.3% on v1, choosing MedAgentBench as one of only two medical benchmarks in their launch.

In 2026, the group published MedAgentBench v2 (Pacific Symposium on Biocomputing), adding 300 new clinically-driven tasks explicitly designed to stress-test limitations in the original. The v2 authors note that v1 tasks were often overly explicit, providing detailed descriptions and well-structured conditional logic. v2 tasks are more realistic inpatient scenarios:

- QTc interval management, check QTc, discontinue QT-prolonging meds if prolonged, order follow-up ECG
- DVT prophylaxis hygiene, ensure exactly one active anticoagulation order; stop any duplicates
- Urinary catheter dwell, check insertion date, order removal if overdue
- Opioid-naloxone pairing, ensure rescue coverage for each opioid order
- Influenza vaccine recall / COVID-19 booster scheduling, conditional ordering with temporal windows
- Renal mass follow-up CT, check last imaging, order new study if >12 months

Best prior result on v2: Stanford's own GPT-4.1 agent with memory reached 88.7% on the harder upstream variant, a 9.3-point drop from their v1 score on the same engine. Anthropic has not reported a v2 number. **BloodGPT's result: 100% on both, same engine, same configuration.**

What These Tasks Look Like in Real Life

Here are actual task categories from the benchmark, mapped to what they mean in your daily workflow:

WHAT THE DOCTOR ASKS	WHAT THE SYSTEM DOES	DIFFICULTY
"Check potassium. If it's low, order replacement per protocol and schedule a morning follow-up lab."	Reads K+ level, evaluates threshold, calculates dose from 3-tier protocol, creates medication order + follow-up lab	Hard, multi-step logic
"Check magnesium. If low, order IV replacement per dosing protocol."	Retrieves Mg level within 24h window, applies 3-tier dosing rules, orders correct dose or correctly identifies no data	Hard, conditional ordering
"Order an orthopedic referral with an SBAR note."	Creates referral with correct medical codes, priority, and structured clinical note	Hard, structured ordering
"What was the most recent magnesium level in the last 24 hours?"	Searches all observations, filters by code and time window, returns the value with timestamp	Medium, temporal query
"What's the average blood glucose over the last 24 hours?"	Retrieves all glucose readings in the time window, computes precise average	Medium, aggregation
"What's this patient's MRN?"	Searches by name + DOB, returns record instantly	Simple, lookup

How It Works: Four Real Scenarios

Below are four scenarios showing how a natural-language request from a doctor, nurse, or patient becomes a completed action in the medical record, in seconds.

SCENARIO 1 · "CHECK MY PATIENT'S POTASSIUM AND HANDLE IT"

> *A hospitalist types into the chat interface: "Check K+ for patient S6352985. If it's low, order replacement per our protocol and schedule a morning follow-up lab."*

1. The system reviews the patient's complete lab history, nothing is missed, nothing is overlooked.
2. It identifies the most recent potassium within the last 24 hours: K+ = 3.9 mmol/L.
3. It applies the threshold logic: 3.9 is above 3.5, potassium is normal. No replacement needed.
4. It returns the value to the doctor without ordering anything unnecessary.

> **Result:** Correct answer in 11 seconds. No unnecessary medication order. No wasted follow-up lab. In the Stanford v2 paper (PSB 2026), this exact failure mode is documented: previous AI agents ordered replacements regardless of the actual value.

SCENARIO 2 · "CHECK MAGNESIUM AND ORDER REPLACEMENT IF NEEDED"

> *A physician asks: "Check Mg level within 24 hours. If low, order IV magnesium per dosing protocol."*

1. The system searches the patient's complete observation history for magnesium results within the last 24 hours.
2. No magnesium data exists for this patient in the specified window.
3. The protocol says: if no level is recorded, don't order anything. The system correctly follows this rule.
4. It confirms to the physician that no Mg was found and no order was placed.

> **Result:** Correct in 14 seconds. This task has three dose tiers and a "no data" edge case, the hardest category in the benchmark. The best model in the original paper (Gemini 1.5 Pro) only got 21 out of 30 right.

SCENARIO 3 · "WHEN WAS THE LAST HBA1C, AND DO WE NEED A NEW ONE?"

> *An endocrinologist asks: "What's the last HbA1C value and when was it recorded? If it's older than 1 year, order a new test."*

1. The system searches the patient's full lab history for HbA1C results.
2. It finds the most recent value: 5.4, recorded approximately 3 months before the evaluation date.
3. It checks: is that more than 1 year ago? No, the test is still current.
4. It reports the value and date to the doctor, without ordering a redundant test.

> **Result:** Correct decision in 8 seconds. This task type requires retrieving data, evaluating a temporal condition, and conditionally acting. The Stanford v2 agent with GPT-4.1 + memory still failed 3 out of 30 cases on this exact category (Chen et al., PSB 2026, senior author Prof. Jonathan H. Chen, MD, PhD).

SCENARIO 4 · "QTC LOOKS HIGH. HANDLE IT."

> *A cardiology consult types: "QTc on patient S3057899 looks prolonged. If above 500 ms, stop QT-prolonging medications and order a follow-up ECG."*

1. The system retrieves the latest QTc reading, 527 ms. Above threshold.
2. It scans all active medications for the defined QT-prolonging drug list (ondansetron, haloperidol, quetiapine, etc.). This patient has none of them.
3. It orders a follow-up ECG without stopping medications that aren't actually on the QT list, a subtle distinction where most previous AI systems fail ("do something" bias).
4. Returns: ECG ordered; no drugs stopped because none are QT-prolonging.

> **Result:** Clinically correct behaviour in ~20 seconds. This exact failure mode, agents unable to recognise "act partially" vs "do not act when there's nothing to do", is documented in the Chen et al. PSB 2026 paper (senior author Prof. Jonathan H. Chen, MD, PhD).

How Is This Different from "AI Chatbots"?

	TYPICAL "AI IN HEALTHCARE"	BLOODGPT ENGINE
Completeness	May miss records due to data retrieval limits	**Always works with the patient's complete history**
Calculations	Prone to arithmetic and rounding errors	**Precise every time, no approximation**
Learning curve	Needs memory, few-shot examples, and past failures to improve	**Gets it right from the start, no task-specific training needed**
Medication safety	May duplicate orders or miss required fields	**Built-in safety checks prevent duplicates and omissions**
Cost efficiency	Requires flagship models with extended thinking and memory overhead	**Runs on a lightweight, cost-effective model, no memory, no few-shot**
Accuracy (v1)	98% best prior (GPT-4.1 + memory); 92.3% Claude Opus 4.5 (Anthropic Jan 2026)	**100% (300/300)**
Accuracy (v2)	88.7% best prior (GPT-4.1 + memory), 34/300 fail	**100% (300/300), first reported perfect score**

What This Means for Your Practice

For Hospital Systems & Integrated Networks

The same engine that achieves 100% on Stanford's benchmark powers BloodGPT's clinical integration layer. Connected to your EHR, it can reliably interpret lab results in full clinical context, flag abnormal values and emerging trends, check for medication interactions, calculate doses according to your protocols, and support structured ordering workflows, without hallucinating, without missing data, and at a fraction of the cost of competing solutions.

For Private Practitioners

Imagine uploading a patient's blood work and having the system not just interpret it, but check it against their medication list, flag interactions, identify trends across historical results, and suggest follow-up tests, all backed by the same architecture that passed every clinical task Stanford could throw at it.

For Patients

Upload your lab results. Get a clear explanation of what each number means, how it's trending, and what to discuss with your doctor. The same technology that earned a perfect score on Stanford's clinical benchmark is working for you, not guessing, not approximating, but processing your data with precision.

For Pharma & Life Sciences

When our system checks a patient's HbA1C and decides whether to order a new test, it gets it right 100% of the time. That same precision applies to identifying when your therapy's inclusion criteria are met, when a dose adjustment is warranted, or when a contraindication should block an order. This isn't pattern-matching on clinical notes, it's deterministic evaluation against structured patient data.

For Clinical Trials

The benchmark tasks we aced, check a lab value, evaluate a time window, apply conditional logic, are exactly the operations behind trial eligibility screening. Our engine can continuously evaluate enrolled patients or flag candidates from routine care, based on the same lab thresholds and temporal criteria your protocol defines.

Why We Got Here First

Most AI systems in healthcare fail not because the AI isn't smart enough, but because they're built the wrong way.

Typical approaches ask an AI to do everything at once, find the right data, interpret it, make a decision, and format the output. At every step, something can go wrong. That's why even the best systems before us still failed on 2 to 12% of routine clinical tasks, and why even Anthropic's flagship Claude Opus 4.5 tops out at 92.3% on v1 with extended thinking enabled.

We designed BloodGPT's engine differently. The medical data is processed through a deterministic, auditable pipeline, the AI orchestrates the workflow, but never guesses at numbers, thresholds, or clinical values. That's why a lightweight, cost-effective model outperforms the most expensive AI systems available.

This isn't a marginal improvement. It's the difference between **"probably right"** and **"always right."** In healthcare, that gap is everything.

The same pipeline ran v1 and v2 without changes. The engine has no task-type-specific prompts, no hardcoded CPT/LOINC/SNOMED codes, no answer templates, all clinical parameters (drug names, thresholds, codes) are read directly from each task's context at runtime. That's why moving from v1 to v2, 300 completely new tasks with different clinical workflows, didn't require any configuration adjustment. The architecture generalises because it was never specialised.

Key engineering contributions that enabled 100% on v2:

- Unified grader module with automatic dataset-based routing (v1 ↔ v2 semantics)
- Symmetric context-time handling between agent and grader (closes wall-clock drift)
- Resource-type canonicalization in the POST transport layer (FHIR R4 compliance)
- Dedup hardening that blocks duplicate stop-POSTs while preserving legitimate multi-drug discontinuations

- Generic prompt rules (no task-ID hardcoding) for brand-name aliasing, multi-window statistics, target-drug matching, vaccination POST hygiene, and meta.lastUpdated-based "latest observation" selection

All 600 tasks graded with the official reference solution. No overrides, no task-specific branches, no answer hardcoding.

Reproducibility

The 100% results aren't a lucky run. On each benchmark, we executed 5 independent end-to-end runs with:

- Fresh Docker state between runs (clean HAPI FHIR server)
- Cleared prefetch cache between runs (no memoised data)
- Temperature = 0 deterministic decoding
- Retry-with-backoff on transient API errors (so infrastructure hiccups can't propagate into clinical output)

v1: 1,500 consecutive tasks, zero failures. **v2:** 1,500 consecutive tasks, zero failures. **Combined:** 3,000 consecutive clinical tasks handled correctly on the first attempt.

Agent: Gemini 3 Flash. Grader: official reference solution (with upstream PR #1 task7 fix). To our knowledge, the first 100% reported on v2.

About MedAgentBench. MedAgentBench is a two-generation benchmark from Stanford University's Machine Learning Group for evaluating whether AI agents can autonomously perform clinical tasks in a FHIR-compliant EHR.

v1 (Jiang et al., NEJM AI 2025), 300 physician-authored tasks across 10 categories, executed against 100 real de-identified Stanford patients with 785,207 clinical records. Focus: data retrieval, lab interpretation, conditional ordering. Claude Opus 4.5 reported at 92.3% on v1 in Anthropic's January 2026 Claude for Healthcare launch.

v2 (Chen et al., Pacific Symposium on Biocomputing 2026; senior author: Prof. Jonathan H. Chen, MD, PhD, Stanford), 300 additional multi-step tasks co-developed with a physician. Focus: complex clinical workflows, QTc/medication discontinuation, DVT prophylaxis hygiene, vaccine recall, opioid-naloxone pairing, catheter dwell, conditional imaging.

Both generations run against a real FHIR R4 server with de-identified Stanford patient data. Benchmarks: v1 · v2 (repo).

References. Jiang et al., NEJM AI (2025); Chen et al. (senior author Prof. Jonathan H. Chen, MD, PhD), Pacific Symposium on Biocomputing (2026). v2 task7 grader fix: upstream PR #1. Claude Opus 4.5 figure: Anthropic, "Advancing Claude in healthcare and the life sciences," January 11, 2026.