Biomni and the Code-Executing Biomedical Agent: What Happens to Drug Discovery, Clinical Genomics, an the Price of Scientific Labor When the Model Can Actually Run the Workflow

JUL 11, 2026 • PAID

1

🎧 Podcast episode for paid subscriber only. Also available on Spotify.

Thoughts on Healthcare Markets & Technology

Biomni and the Code-Executing Biomedical Agent: What Happens to Drug Discovery, Clinical Genomics, and the Price of Scientific Labor When the Model Can Actually Run the Workflow

Biomedical research has a gap between asking a scientific question and getting a defensible computational answer. That gap is filled by expensive human labor. A new agent called Biomni is trying to close it…

Listen now

17 minutes ago · Thoughts on Healthcare

To listen to paid episodes in Apple or Spotify, link your Substack subscription via the s settings on those platforms (instructions inside the Substack app under Subscriptions → Podcast).

Abstract

Biomni is a general-purpose biomedical AI agent that pairs language-model reas with an actual execution environment: 150 specialized tools, 105 software packa databases, code in Python and R, and the ability to run, inspect, and debug its ov work. Built on Claude Sonnet 3.7, it was assembled by mining roughly 1,900 can tasks out of 100 papers across each of 25 subject areas.

The short version of why it matters:

- It converts a research goal into a runnable sequence of actions instead of jus answering questions. Different product category from a chatbot.
- Benchmarks are respectable but honest: 74.4 pct on LAB-Bench DB QA (hu 74.7), 81.9 pct on sequence QA (human 78.8), and a humbling 17.3 pct on a h biomedical exam subset.
- The real proof was wet-lab, not leaderboard: it designed a working B2M gui into a lentiCRISPR v2 Blast vector, a human ran the protocol unmodified, colonies grew, sequencing matched.
- The economic target is the translation layer between scientific intent and executed computation. That layer is where most biomedical research labor actually goes.
- Moats are boring and durable: the maintained environment, workflow-speci evals, proprietary institutional data, governance, and outcome feedback. No model.
- Healthcare is harder than biology because liability, PHI, and prospective validation change everything.

What follows is a longer walk through all of it, with the caveats that separate a g demo from something a quality-assurance team will actually sign off on.

Table of Contents

What the thing actually is under the hood

The benchmarks, read like someone who has been burned before

The cloning run that mattered more than any leaderboard

What actually shifts inside pharma R and D

The repricing of scientific labor

Why the clinic is a different animal than the bench

Where the real moat sits

The failure modes nobody puts in the demo

What the next few years probably look like

Why the model was never the hard par

Biomedical research has spent about a decade quietly drowning in its own outpu Sequencing got cheap, multi-omics platforms multiplied, imaging went high-dimensional, and clinical data leaked out of the record in a dozen semi-usable sh Every lab now sits on a heap of spreadsheets named final, final_v2, and final_ACTUAL_use_this. The bottleneck stopped being data generation a while The bottleneck is the human labor it takes to turn a scientific question into a defensible chain of computation and, eventually, experiment.

That chain is ugly in a way that rarely makes it into papers. Someone has to find relevant literature, pick an analytical framework, locate the correct reference

database, map identifiers across builds, reconcile versions, install a package that not been maintained since the postdoc who wrote it left for a hedge fund, write t code, look at the output, notice the metadata are wrong, redo half the pipeline, regenerate the figures, and then explain to a program team what happened and v took two weeks. None of these steps is genius work. Stacked together, they eat a absurd amount of expensive time.

Language models chipped away at pieces of this. They can explain an assay, draft Python, float candidate genes, summarize a review, or produce a protocol that re beautifully. And beautifully is exactly the problem. Fluent is a dangerous adjecti this field. A protocol can be well written and specify the wrong restriction site. A plan can look clean and quietly leak labels across train and test. A variant call ca name the right disease and completely ignore that the allele frequency makes th proposed mechanism absurd. Fluent text does not pipette, does not run Scanpy, not query ClinVar, does not align a sequence, and above all does not tell anyone the library just silently dropped a third of the observations.

Biomni is a swing at that gap. It treats biomedical research as an action problem trivia problem. The model does not just need to know things. It needs a place to knowledge into operations. That sounds like a small distinction until it hits real Ask a model how to annotate a single-cell dataset and you get advice. Ask an age load the dataset, inspect its structure, normalize counts, correct batch effects, cl score markers, weigh candidate labels against each other, and hand back the intermediate files, and you get a workflow. One is a conversation. The other is th front edge of labor substitution, which is a much more interesting thing to underwrite.

Or purchase a paid subscription.

← **Previous**