Diagnosis Was the Easy Part

As machines learn to out-diagnose doctors, the frontier of personal health turns out to be something older and humbler: a faithful memory.

Rafael M. Ehlers · rafaelmehlers.com.br · July 2026 · doi.org/10.5281/zenodo.21350119

> **A note to the reader.** This essay rests on a single bet, and I would rather expose it in the first line than bury it: that the hard problem in personal health stopped being medical knowledge some time ago, and is now the continuity and *fidelity* of the record a life leaves behind. If that is wrong, if capability is still the binding constraint, the argument collapses, and it should. Judge it against your own last year of health, not against a benchmark.

The reversal

For most of the last century, the bottleneck in medicine was expertise. Knowledge lived in a few trained heads and a few heavy books, unevenly distributed and expensive to reach. Almost everything we built, hospitals and journals and referral chains, was a machine for moving scarce expertise toward the people who needed it.

That bottleneck is closing, faster than the culture has absorbed. In 2025 a team at Microsoft ran language models through 304 of the hardest diagnostic cases the *New England Journal of Medicine* publishes, reframed so the model, like a real clinician, had to ask questions and order tests one step at a time rather than read a tidy vignette. An orchestrated system reached roughly 80% diagnostic accuracy. The generalist physicians they benchmarked against reached about 20%. Four to one, and cheaper. The result held across model families; it is not a fluke of one lab's tuning.

You can argue with the benchmark, and I will in a moment, but not with the direction. The thing we spent a century treating as scarce and sacred, diagnostic reasoning, is becoming abundant. And the moment a scarce thing becomes abundant, the interesting question is always the same: what is the new bottleneck?

What the benchmark quietly admits

The answer is hiding inside the very study that announced superhuman diagnosis.

Notice what the Microsoft team had to build to make their test realistic: a "Gatekeeper," a second model whose only job is to release a patient's information, piece by piece, when the diagnostician explicitly asks for it. They built it because they understood that real diagnosis is not a reasoning problem handed a complete case. It is a sequential information-gathering problem. The skill being measured is knowing what to ask, in what order, and when you have enough to commit.

Sit with that. Even inside the artificial world of a benchmark, the diagnostician is capped not by how well it thinks but by what it can pull. The reasoning is solved. The retrieval is the game.

Now leave the benchmark. There, the information exists, complete, behind the Gatekeeper, waiting to be queried. In a life, it does not. A life is not a packaged case. It is lived across years, in fragments, mostly unrecorded and half-forgotten. The single most common sentence in any consulting room is not a symptom. It is "I think it started in July, or maybe after Carnival." The person who knows the most about that body, the one who was present for every bad night and every meal and every ache, is also the one least able to tell its story.

So here is the reversal in full. We are about to hand everyone a diagnostician that reasons at or above the level of a good doctor. And it will sit, brilliant and ready, in front of a patient who cannot supply it with the one thing it needs to be useful: a faithful account of what actually happened, over time.

I once put this as a slogan: we gave everyone a library and expected a doctor. The line needs updating. The doctor has arrived. We still forgot to give anyone a memory.

The thread nobody holds

The rich always had a version of this. The family physician who followed a household for decades, who knew the grandfather and the father and the child, who could connect a complaint today to one from twenty years ago, was never valuable mainly for superior knowledge. He was valuable because he held the thread. He *was* the continuity. Almost everyone else lost that when medicine became a sequence of fifteen-minute encounters with strangers, each seeing a slice, none holding the whole.

This is the vacancy. Not knowledge; we drowned in that. Not reasoning, either, not for much longer; the machines are taking that. What is missing is a continuous, faithful record of a single life: the substrate over which any intelligence, human or artificial, has to reason to be of any use to *you* in particular.

The systems we have do not fill it. A step-counter collects and forgets the context. A patient portal archives documents nobody rereads. A generalist chatbot answers today's question and wipes the slate before the next one. People already arrive at these tools for exactly the personal, longitudinal questions the tools are worst at, disproportionately at night and on weekends, when the formal system is closed. But the chatbot has no yesterday to connect their today to. Each of these lives on the wrong side of the problem. None of them holds the thread.

The part everyone gets wrong: fidelity

Here the argument gets sharp, and here I think most thinking about "AI that remembers" is naive.

It is easy to say an agent should remember you. It is easy to build something that stores. The hard problem, the one that decides whether any of this works, is whether it remembers *faithfully*.

Consider what building a longitudinal record out of ordinary talk actually demands. A person says, in passing, "slept badly, maybe five hours, woke up with a headache, probably work stress." For that sentence to become memory, one system has to do two opposed jobs at once: answer like a warm companion, and, invisibly, extract the recordable facts (sleep, five hours; symptom, headache) and file them, dated, on a timeline. Then, weeks later, something has to look across hundreds of such fragments and surface the pattern no single day reveals.

Every step is a chance to lie. If the extractor writes down eight hours when the person said five, the error does not stay small. It compounds. Every correlation drawn from that timeline, every weekly synthesis, every gentle "you tend to feel this way after…" now rests on a corrupted foundation and inherits the corruption, while *sounding* more confident, not less, because it arrives dressed as a pattern rather than a guess. Unfaithful memory is worse than no memory. No memory leaves you where you started. Unfaithful memory launders error into authority.

This is why I think the real frontier in personal health AI is not a capability problem at all. It is a *fidelity* problem, and fidelity is measurable. You can ask, concretely: does the structured event match what the person actually said? Do the syntheses reflect real regularities, or artifacts the model invented to seem insightful? Those are questions with answers, and almost nobody is asking them, because the field is still intoxicated by the capability curve that is, in this domain, already flattening into irrelevance.

Four ways to be wrong

An argument that generates no test is decoration. So, falsifiably:

Continuity beats capability. An agent with your faithful history and a mediocre model gives more personally useful guidance than a frontier model with no memory at all. *Refuted if,* in blind comparison, the memoryless frontier model ties the one that knows you.

The value is the fidelity, not the storage. Unaudited retention will produce confident, wrong syntheses at a rate that erodes trust faster than the memory builds it. *Refuted if* the accuracy of the record turns out not to matter to trust or to outcomes.

Demand precedes the product. People are already using memoryless assistants for precisely the personal, longitudinal, after-hours questions those assistants are worst at. *Refuted if* the usage turns out uniform, impersonal, and clustered in clinic hours.

The moat is the record, not the model. Whatever gets built here will be defensible in proportion to the faithful history it has accumulated, not the cleverness of its reasoning, because

the reasoning is becoming a commodity and the history takes years to exist. *Refuted if* users switch away without friction no matter how much record they have built.

The objection I owe you

"You are describing a safety nightmare. An assistant that hoards a person's entire health history, identified and longitudinal, is the most sensitive data object imaginable, and the most abusable."

Yes. That is the true cost, and I will not soften it. The same architecture that could hold the thread with care could hold it as surveillance. A record faithful enough to help is faithful enough to harm, and identified, longitudinal health data is the most intimate thing a person owns. Which is why, if any of this is built, ownership has to sit with the person. The thread is theirs: portable, inspectable, erasable. Not as a feature, but as the precondition of the entire thing. Break that trust once and the argument does not lose some users; it loses its right to exist. I raise the objection because it is the strongest one against building this, and because it does not touch the thesis. It raises the stakes of getting the thesis right.

Coda

Begin where medicine keeps failing: someone in a waiting room, unable to tell the story of their own body. For a hundred years we assumed that person's problem was a shortage of expert knowledge, and we built cathedrals to distribute it. We are about to finish the job with machines that reason like specialists.

And that person will still be sitting there, unable to say when it started.

Because the thing they were missing was never the doctor's knowledge, and soon will not be the doctor's reasoning. It is the humblest thing in the whole system and the last we thought to build: a faithful memory of them, kept over time, that ties the ache today to the bad night three weeks ago and does not make them begin from nothing every time they open their mouth.

Diagnosis was the easy part. Being remembered, faithfully, is the frontier.

A note on method. This essay states a position; each falsifiable claim above is meant literally, and any one of them, refuted, weakens it. The sequential-diagnosis benchmark and the roughly 80% versus 20% figure come from the first reference below; the pattern of public, personal, after-hours use of generalist chatbots for health comes from the second.

References

Nori, H., et al. "Sequential Diagnosis with Language Models." *arXiv*, 2025. arxiv.org/abs/2506.22405

Costa-Gomes, B., et al. “Public use of a generalist LLM chatbot for health queries.” *Nature Health*, vol. 1, 2026, pp. 689-696. nature.com/articles/s44360-026-00117-x