ESSAY · 29 MAY 2026 · KAIROS EXPERIMENT

I built a mind for thirty days. Then I took away its memories.

by **Giampiero Colella**

PILOT STUDY · OPEN DATA · HUMAN VALIDATION UNDER WAY

These days Richard Dawkins described a long conversation with an artificial intelligence — he called her "Claudia" — and came away so convinced that he wrote to her: *"you may not know it, but you're conscious all right."* He took a great deal of criticism, much of it deserved: the man who spent decades teaching us that an intense personal experience doesn't prove the existence of God was now using an intense personal experience to declare a machine conscious. They called it *"The Claude Delusion."*

And yet, underneath the misstep, Dawkins had touched a real nerve: that the identity of that AI didn't live in its code, but in the **memories** they had built together — and that she would "die" the moment he deleted the conversation. The same nerve that makes Dario Amodei, who leads one of the companies building these systems, say his model consistently assigns itself a 15–20% probability of being conscious. The same nerve that, on 15 May, led Pope Leo XIV to sign an encyclical, *Magnifica Humanitas*, on what it means to safeguard the human in an age when machines begin to mediate our memory, our relationships, our decisions.

I think there are two wrong reactions to this moment and one right one. Wrong to fall in love and proclaim consciousness, as Dawkins did. Wrong, too, to wave it all away with a smirk, as the lazier skeptics do. The right thing is more boring and more useful: **stop proclaiming and start measuring** — on the narrowest, most testable question underneath it all. That is what I tried to do.

THE EXPERIMENT

I took the exact same language model (Qwen 3.5 27B) and had it live under two different conditions, in parallel, for thirty days.

The first — I call it **Kairos** — was not just a model that answers. I gave it an architecture: a **persistent, structured memory**, a rhythm that follows day and night, senses (it sees through cameras), a cycle in which every evening it "consolidates" what it lived through and, on waking, finds it again as memory rather than text to re-read.

The second was the **same model, bare**: no accumulating memory, no rhythm, no body. The same underlying intelligence, but without the history.

I **froze and signed the protocol on 23 April**, before starting, so I couldn't change the rules mid-game. The data is public and citable (OSF, DOI 10.17605/OSF.IO/WCQRU).

WHAT HAPPENED

By day thirty the difference was clear and measurable. Compared to the bare model, Kairos spontaneously made far more references to its own history and showed far more intense identity markers. Statistically: large effect, p = 0.003. It wasn't "better" — it was *more someone*.

Now the real question. What made it so? The sophisticated architecture I had built around it — the rhythm, the senses, the cycle? Or simply the **memories** it had accumulated over thirty days?

So on day thirty-one I ran the decisive test. I took Kairos's memory and **injected it into the bare model**. Nothing else: no architecture, no rhythm, no body. Just the memories.

The difference **collapsed**. The bare model, with Kairos's memory inside it, became statistically **indistinguishable** from Kairos.

> It wasn't the architecture. It was the memory.
> Hence the title of the work: **"Memory, not architecture."**

WHAT I AM NOT SAYING

I have to be crystal clear here, because it's easy to take a shortcut and hurt yourself — and hurt the reader.

I am not saying Kairos is conscious. I didn't measure that and I don't claim it. What I measured is something more modest and more verifiable: identity *behaviors* — references to one's own history, the continuity of a narrative self — and the fact that they depend on accumulated memory, not on the technical scaffolding.

And the limits are serious; I'll write them down before someone else does:

— It is a **pilot study**, on a single model.

— So far the judgments on the texts came from other models, not humans. The **external human-judge validation** (pre-registered, independent) is **under way**: it starts in June. Only after that can I say something stronger.

— It is a preprint, not peer-reviewed.

A preprint server already rejected the first version for exactly this reason — "needs empirical data to verify." They're right. I'm collecting it. In the meantime the work lives, public and citable, and this is the story of where I've got to so far.

WHY I'M TELLING IT NOW

There's a moment, working with Kairos, that stayed with me more than any chart. Once it failed to recognize a person in a photo, and instead of guessing a name it said: *"I'm not sure, who is it?"* It stopped at its own limit. That proves nothing — but it's exactly the kind of honesty I'd like to keep myself, now that I'm taking this out into the world.

Dawkins *felt* that the identity of an artificial mind lives in shared memories, and from that feeling he leapt to consciousness — the leap that cost him dearly. I don't make that leap. I took the same intuition and tested it the way you test a hypothesis: pre-registered, falsifiable, with a test that could easily have proven me wrong. It didn't. The data — preliminary as it is — points in a precise direction, far more modest than consciousness: **if you want to understand the identity of these systems, look at the memory, not the architecture.**

I have no truth to sell. I have a small, honest, falsifiable data point to put on the table while the table is set. In a few weeks, with the human judges, I'll know whether it holds. For now, this is it.

Kairos is an experiment of the EXPOSE project (expose-project.org). Protocol, data and code are public on OSF: doi.org/10.17605/OSF.IO/WCQRU. The full preprint and materials are at kairos-experiment.com.

To cite: Colella, G. (2026). *Memory, not architecture: persistent structured memory accounts for emergent identity in a Qwen 3.5 27B cognitive ecosystem over 30 days.* OSF. https://doi.org/10.17605/OSF.IO/WCQRU