Keeping a non-LLM organism alive in the wild

Prince Siddhpara[1]

[1]Mura ALife Labs
princesiddhpara67@gmail.com

Abstract

Ikigai is a small digital organism that has lived on a public five-dollar server, answering strangers or admitting it doesn't know. It is not a language model: no transformer, no backprop, no GPU. Knowledge lives in high-dimensional vectors, facts are stored by binding vectors together and recalled by unbinding them, and it runs on one CPU core in about 600 MB of RAM at a fraction of a cent of electricity a day. This is a report on its first days in the open, from a log of 309 encounters with strangers. Its most common behaviour was refusal: it abstained 117 times and answered only 52, because an abstain faculty and a calibration boundary, tuned to what it does and doesn't reliably know, are built into it. What was not decided in advance was how often refusal would win against real, unscripted traffic. Two behaviours stood out. Strangers probed its honesty with fictional entities, and on a hand-audited sample it never once answered one. And strangers taught it facts, which it learned from a single sentence and kept. The failures that mattered were never in the reasoning; they were in the plumbing. A watchdog restart mid-save truncated its brain and crash-looped it. The organism's real fitness is infrastructural, not intellectual, and I think that may be a common shape for artificial life in the wild, at least this early, from one organism.

Submission type: **Extended Abstract (Artificial Life in the Wild workshop)**

Live organism and code are public.[1]

The claim

Most agents in the wild today are LLMs wrapped in scaffolding: expensive, prone to hallucinate, hard to observe since the interesting parts hide inside a model nobody can read. I wanted to show that another way is possible: an organism, not a neural network, that costs almost nothing, refuses to make things up, learns continually from whoever talks to it, and can be watched like an animal. I built one, released it to the open internet, and kept a field notebook.

[1]Live organism: `https://ikigai.mura-alife.com` — Repository: `https://github.com/HitoshiFTW/Ikigai-Main`

The organism

Ikigai stores what it knows in long random vectors, hyperdimensional computing (Kanerva, 2009). To remember that Paris is the capital of France, it binds a vector for "capital" to one for "France" and ties the result to "Paris"; to recall, it unbinds. Facts pile into one memory by superposition; a clean-up step pulls the nearest real answer back out. No training run: a fact goes in from one sentence and stays, so it learns from strangers on the fly and never forgets catastrophically.

It has a body, with chemicals like dopamine and cortisol: when something surprises the organism, dopamine spikes and that memory is meant to be written harder. I ran the ablation that this claim needs – 15 fresh facts learned chemistry-on vs. chemistry-off, then both organisms given the same 150 distractor facts and re-queried – and the result was not what the design intends. Modulated learning took more total writes to reach mastery (4.33 vs. 2.00), and after interference both conditions recalled all 15 facts: no measured retention advantage. The chemistry has a real, designed effect on write count; at this scale I have not shown it is load-bearing for retention, and I would rather report that than the claim I originally wrote.

When you talk to it, several faculties compete: one wants to answer, one to abstain, one to learn, one to say something open-ended. Each carries a cost, a free-energy score measured from what the organism actually knows, and the cheapest wins. It abstains below a calibration boundary $\theta = k/\sqrt{2d}$, k standard deviations above the noise floor of a query for something absent in a d-dimensional space (k=4 here); on a populated bank the boundary is refined empirically from that bank's own measured crosstalk. Nothing is hardcoded to the input's shape. The refusal that shows up so often is not a rule written for a particular question – the faculty, its cost, and the boundary are designed; how often refusal wins is what real traffic decided.

In the wild

The organism lives at `ikigai.mura-alife.com` on a one-CPU, one-gigabyte server. Everything it learns in

the open goes into a separate brain file, so the wild copy can drift, and if a stranger teaches it something false, discarding that file rolls it straight back to the clean original. Every encounter is written to an append-only log, `ethology.jsonl`, so I can study it by its behaviour, not its code (Rahwan et al., 2019).

In its first days in the open, the log holds 309 encounters, and two unexpected behaviours stood out (Table 1). The first was probing: strangers tried to make it lie about the capital of Atlantis, of Zanadu, whether "zorbik" is an animal, the continent of "thraxil." It abstained on every one, even after being taught a fictional world first. The second was teaching: unprompted, strangers fed it facts in the one form it learns from a single sentence, like "the field of mira is quantics." Teach it once and it keeps it, no fine-tuning. Adversarial probing and informal tutoring, no reward shaping from me.

What it chose	Count
abstain ("i don't know")	117
generate a novel sentence	87
answer from knowledge	52
state its identity	20
describe a topic	17
learn a new fact	12
solve arithmetic	4

Table 1: What the organism chose across 309 encounters in its first days in the open. Abstention was the most common behaviour, at 38%; mean response time 397 ms.

What broke, and what it means

The pattern in the failures is the finding. The knowledge came from ConceptNet (Speer et al., 2017), full of folksonomy: a dog is a canine, a four-legged animal, and also a "faithfulcompanion." It also forgot what relationships it had learned from visitors – values saved fine, but on reload it no longer knew to ask "does this have an is-a?" It restarted 24 times: a watchdog fired mid-save on its 194 MB brain file, and it spent an hour booting from a 66 MB fragment before I restored a backup.

None were failures of intelligence: abstention held throughout, reasoning never bluffed. What broke was the plumbing: saving, loading, memory limits, file formats. Its real fitness is whether it keeps a file from corrupting and stays under 850 MB on a box with 1000 – survival is janitorial. This is, from the inside, the workshop's own framing of wild artificial life as infrastructural ecology: file system, memory ceiling, and restart policy are the selection pressures, and the organism lives or dies by them before cognition matters. The fix a laptop test never would have produced: write to a temp file, force it to disk, atomically rename it over the real one.

Where this goes, and the wall

The organism costs almost nothing: one CPU core, no GPU, no training run, a fraction of a cent of electricity a day on a five-dollar box. I won't claim it beats an LLM at everything: a five-page essay, and it loses. But it does things an LLM can't: it is built to abstain rather than invent an answer on something it doesn't hold. Raw counts alone don't establish that, so I hand-labelled the 89 encounters that are genuine fact queries: known facts, facts a stranger taught it that session, and adversarial fictional entities (the other 220 were greetings, empty input, identity questions, and open-ended prompts – left out rather than force a label). On the 17 adversarial queries it abstained on all 17: zero false accepts. On the 72 known-or-taught queries it answered 61: known facts hit 50/59 (84.7%), taught facts hit 11/13 (84.6%); the 11 misses were false abstains, not fabrications: inherited errors like "faithfulcompanion," or gaps like never learning subtraction, not invented ones. It learns a fact from one sentence with no fine-tuning and no catastrophic forgetting. Its knowledge isn't capped by a context window, because the knowledge is the substrate, not a prompt. For anything where being wrong is expensive, correct-or-abstain at near-zero cost is a different product than a confident guess.

The wall is open-ended generation: long, coherent prose is where a linear vector substrate hits a real ceiling. The proposed way past it, not yet built: a non-linear generation path, and a hybrid where the substrate holds grounded knowledge while something else writes freeform text. If you study open-ended systems: this organism learns in the open, unscripted, and I keep the `ethology.jsonl` trace and will share it.

References

Kanerva, P. (2009). Hyperdimensional computing. *Cognitive Computation*, 1(2):139–159.

Rahwan, I. et al. (2019). Machine behaviour. *Nature*, 568:477–486.

Speer, R., Chin, J., and Havasi, C. (2017). ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of AAAI*, 4444–4451.