How AI Should Have Always Been

A position on continuous intelligence: why the architecture we got cannot learn, what that costs, and the standard by which to measure the difference.

Mikey Lee | July 8, 2026

Download source: https://erebyx.com/paper

1. A Note From The Founder

I didn't come from one of the world's major AI labs. I grew up in Richmond, Virginia wanting to play professional soccer, convinced life was too short to be normal. Somewhere along the way, I became obsessed with a question I couldn't shake: the AI we imagined, the one that knows you, grows with you, and becomes more useful through experience, isn't the AI we got.

You know the one I mean. We've all seen it. It knew you, it learned as you went, it got better the longer you had it, and it was yours. What we got instead is brilliant, and it forgets you the second it answers.

I turned thirty in the mountains two weeks ago and passed a hand-painted "no data center" sign on a back road. It stuck with me, because I don't think I'm the only one who feels like something went wrong here. Not with what AI can do, but with what it's for, and what it's costing people who never signed up for it.

So here's where I stand, and why I'm building what I'm building. I'm going to be straight with you the whole way through. Where I say something is true today, I'll show you the test that would prove me wrong. I'll state planely where we're headed instead of where we are. I'd rather you trust me than be impressed by me.

And before I take a single shot at anyone, I'll own my own glass house: we run today on the very architecture I'm about to take apart. I disclose exactly how in Section 4. Any claim and every claim I make carries its status and its honest falsifier in Section 7. Nothing here is inflated to win the argument. The truth is damning enough.

Here's the position.

2. The Dream

Strip the dream down and the want was always the same thing: continuity. Not raw cleverness, but something that remembered you, your work, the thread of a long conversation, and got more yours the longer you had it. We've imagined that companion in a dozen movies. The wanting was never really about how smart it was. It was about whether it stayed.

What shipped instead is genuinely brilliant and genuinely frozen. Its knowledge is baked into static weights at a training cutoff. It doesn't just forget you overnight, it forgets you the moment it answers. What greets you the next morning is not its memory of you. It is the app, re-handing it the transcript. And it runs on power and water you neither own nor can refuse. The field has its own honest words for this: engineers describe today's models as "a brilliant conversation partner who resets their brain every time you close the tab," and the condition has a name: conversational amnesia.

So here is the plain version of it: we were promised a companion that would learn us, and we got a brilliant stranger with amnesia, rented from someone else's datacenter.

This is not a complaint about capability. The capability is real, and often extraordinary. It is a claim about kind. By the everyday definition (we call the dog, the monkey, the person intelligent when it learns and makes new connections, but never one that repeats the same thing, year after year, without improving), a model whose weights never change after training is not, in all the time you spend with it, a learning thing. It is an often-excellent snapshot of a learning process that ran once, on someone else's side, and does not continue with you.

The thesis of this paper is that this is not a law of nature. It is an artifact of one architecture, a choice, and there is a different one. AI will change the world. It should not cost the world - and by cost we will mean, throughout, the architecture, not the wattage of any single query (Section 2). What follows is what that intelligence would actually be, why we do not have it, and the standard by which anyone could measure whether they had built it.

3. The Architectural Root

The reason today's AI forgets you, and the reason it is expensive, are the same fact seen twice. It comes down to two coupled properties of the dominant transformer-LLM paradigm, kept carefully distinct because each does different work:

1. Knowledge is frozen into weights. A trained model's parameters are fixed at inference. To change what it knows, you re-run gradient training: a full retrain, or a fine-tune. There is no path to write a durable, cheap, incremental update into the knowledge store.
2. The model holds no persistent state. Every call is recomputed over whatever context you supply. To make it "remember" you, the relevant history has to be re-fed as input, every session.

Everything built on top inherits both. Chatbots, coding harnesses, agent frameworks all wrap a frozen-weights transformer. And the popular fixes do not fix the root. Retrieval-augmented generation, long context windows, vendor "memory" features (including the hybrid system we ourselves run today, Section 4) all work by re-injecting retrieved text back into the same recomputed, stateless forward pass. They enlarge the context that must be re-fed. They do not make the core stateful, and they do not update its knowledge. They are band-aids on a stateless core.

And statelessness is sharper than "it forgets you between sessions." It forgets you the moment it answers. A forward pass is a pure function: context in, tokens out, nothing kept. The parameters do not change, and no trace of the exchange survives the reply. What remembers you is not the model. It is the harness, the application around it, which stores your conversation and re-sends the entire transcript as context on the next turn. The memory you feel is an external scaffold, re-handing the amnesiac its own transcript to read again, every turn. (Serving systems do cache the transcript's key/value projections across turns, so a shared prefix need not be fully re-projected each time. But that is an optimization at the serving layer: transient, discarded, bounded by the context window, and it writes nothing into the model's knowledge. It makes the re-reading cheaper; it does not make the model remember.) And the scaffold has a hard edge: it holds only what fits the context window. Past that, the harness truncates or summarizes, and the early thread is genuinely gone. The model never had it, and now the harness does not either. This is the deepest form of a band-aid on a stateless core: even a chatbot "remembering what you said a minute ago" is the harness re-feeding, not the model knowing. Every product built on the paradigm, every chatbot and every coding harness, is underneath a scaffold wrapped around a thing that forgets you the moment it speaks.

The most sophisticated expressions of this same choice are worth naming. One line pursues continuity in token space, curating an ever-better context for a model whose weights stay frozen by design (Letta, Mem0, Zep). Another changes how the model speaks rather than what it knows (diffusion language models, which remain frozen Transformers trained on the same scraped corpora). We share the first's diagnosis and respect both efforts. Where we differ is the destination: for them the frozen model is the product, curated around indefinitely; for us, the rented frozen transformer we disclose today (Section 4) is a transitional organ, meant to be replaced by an owned continuity that learns from you alone, not dressed forever. Across the whole field, better retrieval and faster decoding alike, the core stays frozen and stateless. That is the choice, made again and again.

A careful reader will raise three objections immediately. Each is correct, and each, stated honestly, leaves the diagnosis standing.

- AI doesn't recompute everything. That's what KV-cache is for. True, and it matters. Within a single generation, a model caches the projections of tokens it has already seen, so each new token computes only its own projections rather than re-projecting the whole sequence. But that cache is transient inference-time memory, discarded when

the session ends, and it changes no weight. At the boundary that matters for a relationship, across sessions, the claim holds exactly: nothing persists, and the context is re-processed from scratch. And even within one long conversation, the cache does not make memory cheap. It makes it expensive and growing: the cache itself expands with every token, each new token must attend over the entire accumulated history, and providers charge a premium once a conversation runs past their longest context tiers. Faking memory by re-feeding context costs more the more there is to remember, right up until the window fills and the earliest thread is dropped anyway. That is the exact inverse of a system that consolidates what it learns, where recall stays cheap and flat. Here, the price of remembering rises without bound, and then the system forgets regardless.

- Models obviously learn. Look at in-context learning. Also true, and we will not say otherwise. A model adapts to patterns in your prompt at inference with no weight update. Recent work even shows a forward-pass-with-context behaves like a temporary, low-rank implicit update that is thrown away when the window clears (Dherin et al., 2025). That is the precise point: the learning at deployment is in-context: ephemeral, non-persistent, gone when the window resets. It learns within a conversation. It does not learn across your life with it. Persistent versus ephemeral, in-weights versus in-context: that is the load-bearing distinction, and it is technically exact.

- "Just fine-tune it to update it." Fine-tuning is not incremental learning. Training on new or narrow data overwrites parameters important to prior capability (catastrophic forgetting), and the effect worsens as models scale (Luo et al., 2023). It is partial, lossy, expensive re-training, not a cheap, local, non-destructive write.

What an answer actually costs

The honest cost of an AI answer has three line items. The industry quotes the first and is quiet about the rest.

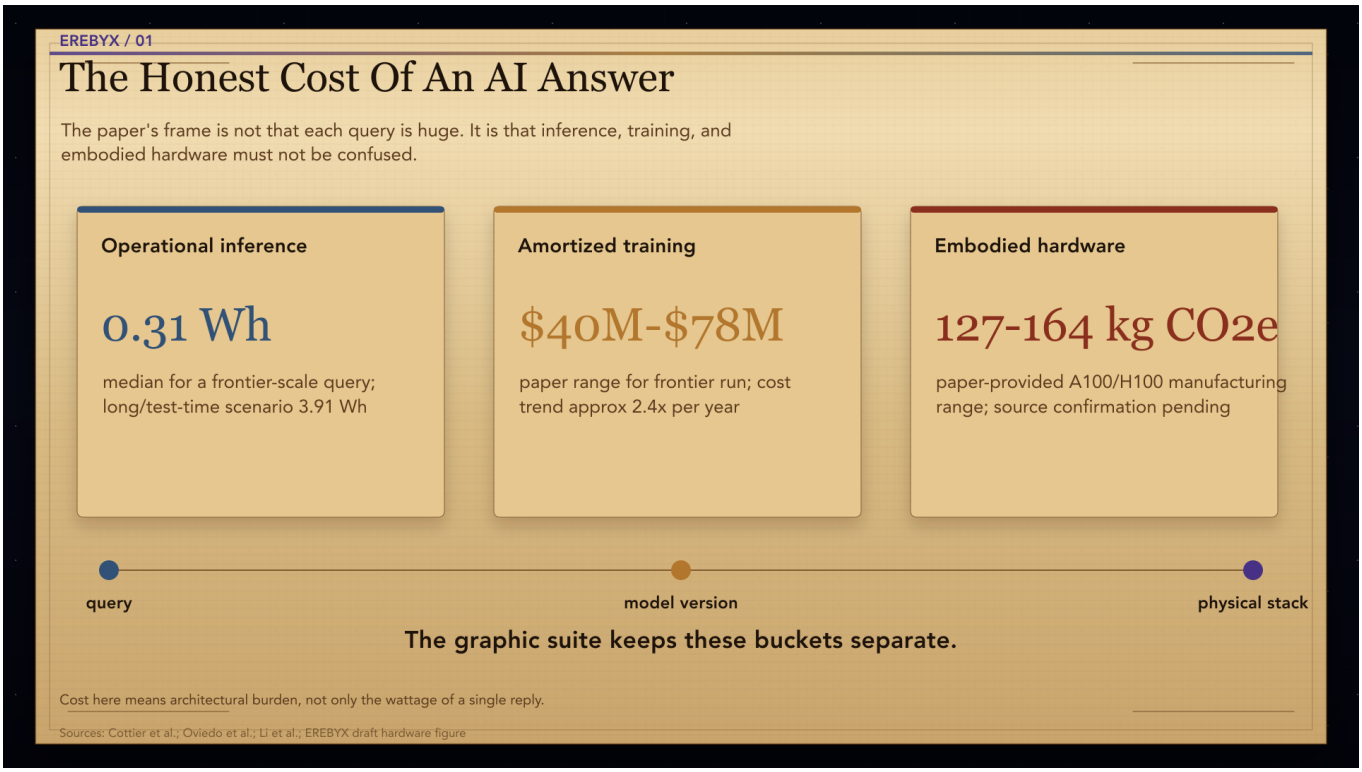

Chart titled The Honest Cost Of An AI Answer showing operational inference, amortized training, and embodied hardware as separate line items.

Line item	Sourced figure	Honest note
Operational inference (per query)	~0.3 Wh for a short text query (Epoch AI, 2025), about 10x below the older ~3 Wh estimate; ~10 - 50 mL of water per medium-length response (UC Riverside)	Small, and falling. We concede this outright; the argument is not "each query is enormous."
Amortized training	A frontier training run costs ~ $40 - 78M (Epoch AI to Stanford AI Index, the same model generation costed by two methods), growing ~2.4x/year, projected past $1B by 2027. GPT-3's run alone drew ~1.29 GWh and ~700,000 L of on-site water.	The line the "cheap inference" story drops.
Embodied hardware	~127 kg CO2e (A100) to 164 kg (H100) to manufacture a single high-end accelerator; on the order of 10 - 30% of lifecycle emissions, depending on grid and utilization	Real, but the minority share. We do not lead with it.

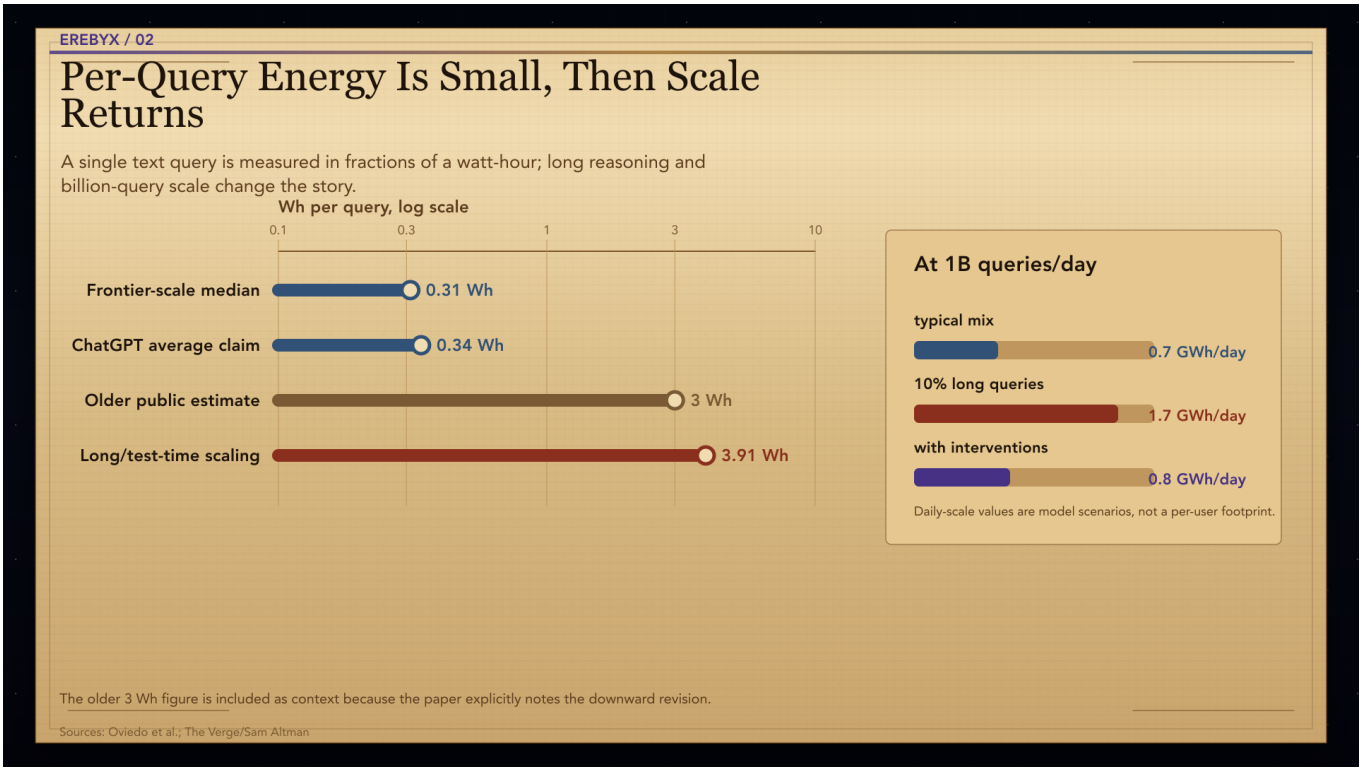

Chart titled Per-Query Energy Is Small, Then Scale Returns comparing watt-hours per query and billion-query daily scale.

Two disciplines govern how we use these. First, we do not conflate the environmental footprint with the price tag. Energy is only about 2 - 6% of the dollar cost of training a frontier model; amortized hardware is 47 - 64% and R&D staff 29 - 49% (Cottier et al., 2024). The water and megawatt figures are environmentally real, but they are a small fraction of the bill. So when this paper says AI "should not cost the world," it means the architecture, not a claim that each query boils a lake. Second, we hold ourselves to the same standard we hold others. The most-cited "LLMs are cheap" analysis concedes, in its own words, that its number is "just the cost of inference," while the search baseline it beats "amortize[s] building and updating the index" (Snellman, 2025). That is a training-excluded cost compared against a training-included one: apples to oranges. It is structurally the same error as a multiplier this project itself once published and has since retracted (Section 7). Incumbents make the mirror image of the figure we threw away.

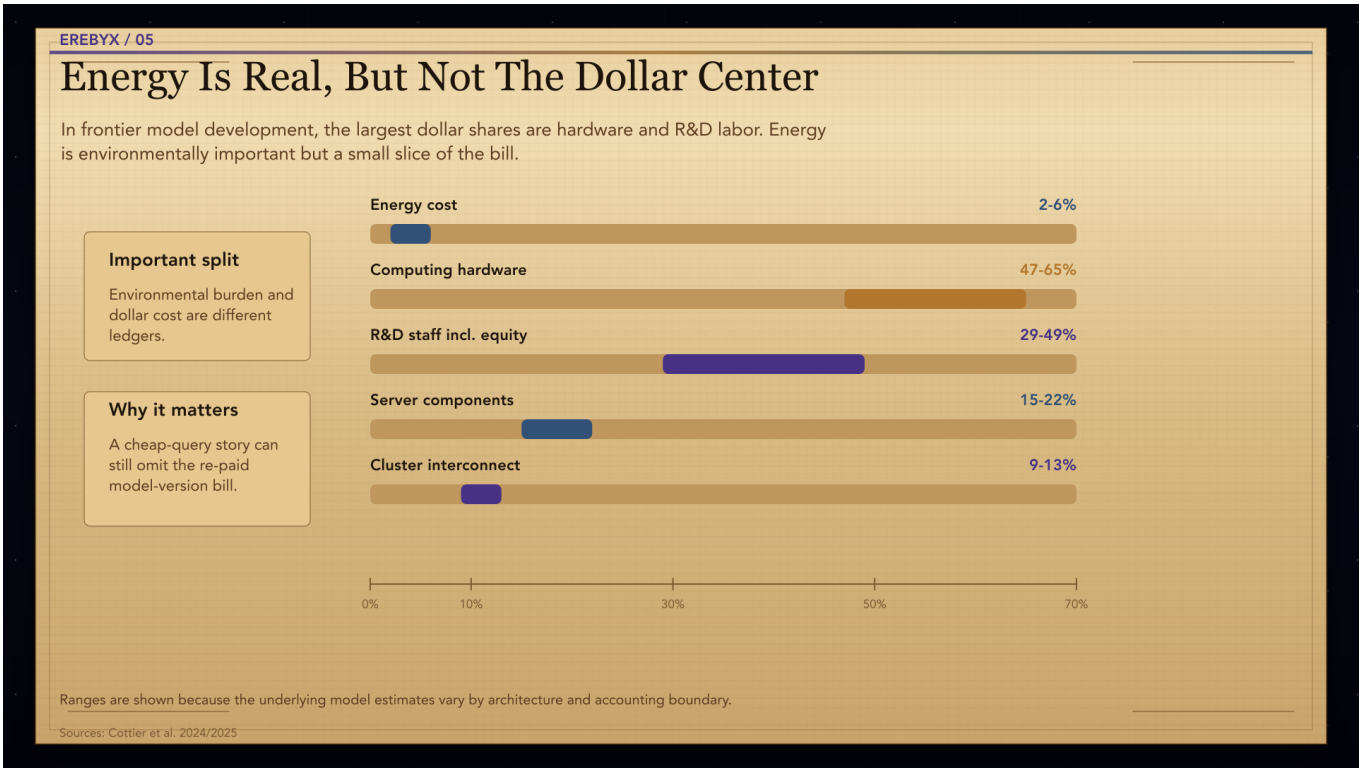

Chart titled Energy Is Real, But Not The Dollar Center showing energy, hardware, R&D, server components, and cluster interconnect cost shares.

Here is the architectural point those numbers serve. Whether amortized training is "small per query" depends entirely on query volume over a model version's life, and frozen weights guarantee the meter resets to zero every version, because the only way to update the knowledge is to discard the model and retrain. "Cheap per query" is true only by spreading a fixed cost (one you re-pay in full at every version, at zero usage) across a query count you hope will be enormous.

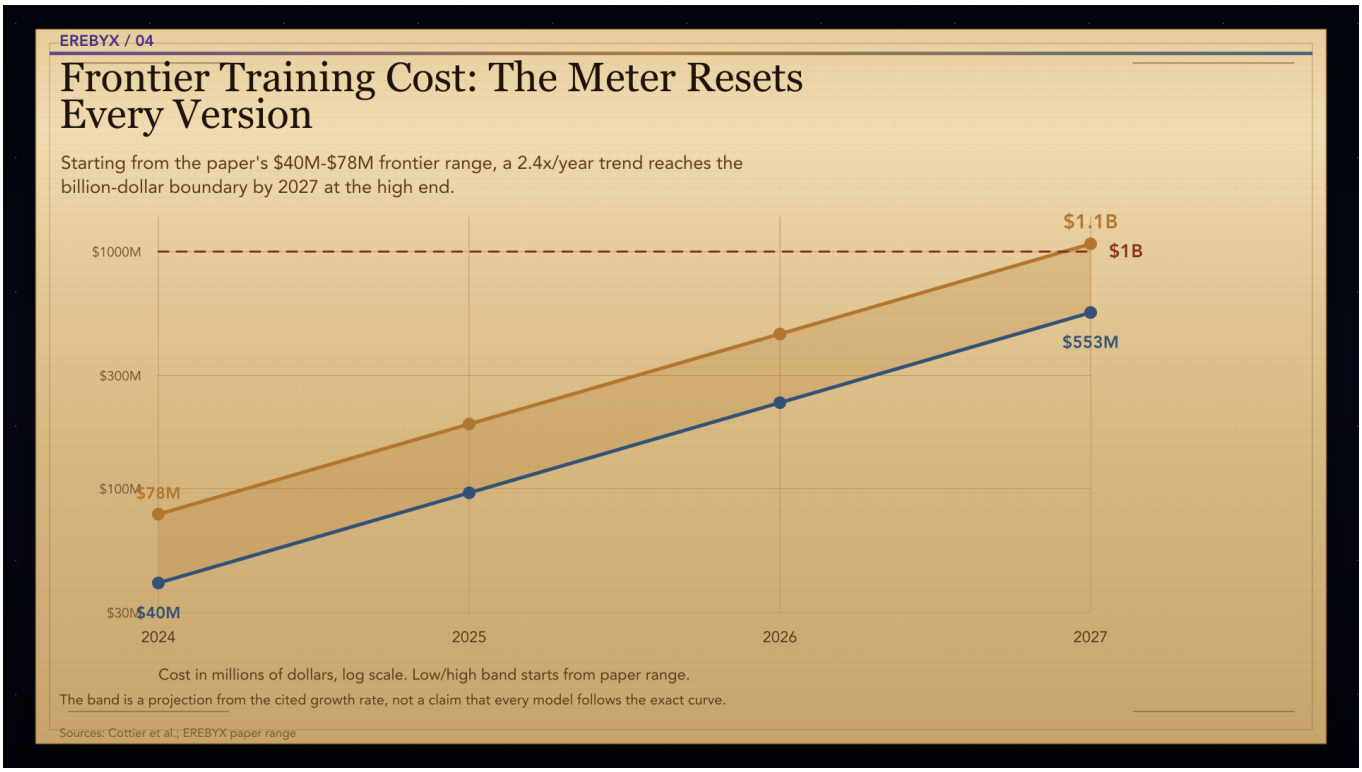

Chart titled Frontier Training Cost: The Meter Resets Every Version projecting frontier training costs from 2024 through 2027.

We will pre-empt the sharpest technical reader. As of 2025, for high-volume deployed models, inference has overtaken training as the dominant share of AI compute and energy at scale. Google reports roughly sixty percent of its ML energy goes to inference, the cost center shifting as serving volume scaled. The naive "training is the hidden cost" framing is only half right. But this strengthens the diagnosis rather than weakening it, because both cost centers trace to the same root: training is re-paid every version because the architecture cannot learn incrementally, and per-query inference carries an avoidable recompute cost because statelessness forces a full forward pass every call. One root, two bills.

Why everyone reaches for scale: a coupling, not a conspiracy

The empirical backbone of the scaling era is real and we will not dispute it: scaling laws hold across orders of magnitude (Kaplan et al., 2020), compute-optimal training is well-characterized (Chinchilla), and Sutton's Bitter Lesson (that general methods which scale with computation win by a wide margin) has been right for a decade. New capabilities have genuinely appeared only in larger models. Scale is not dumb, and we are not claiming it is.

The position, and we label it a position, not a settled fact, is narrower and falsifiable. Frozen weights weld together two things that need not be welded: changing what a model knows, and retraining it from scratch. Because incremental weight-update runs straight into catastrophic forgetting (the stability - plasticity problem; Kirkpatrick et al., 2017), the only reliable lever left to improve a model is to discard the snapshot and train a bigger one. Scale is the knob you are forced to pull when you cannot update what you have without destroying it. That is a structural coupling, true of the architecture today, not a story about anyone's motives, and not a claim that scaling laws are wrong. The falsifier is concrete: demonstrate a cheap, local, non-destructive way to write durable new knowledge into a model's weights at scale, and this coupling breaks.

This is not our private complaint. The lab that built the transformer publishes the same diagnosis: Google Research describes a model whose knowledge is fixed at pre-training as suffering a kind of amnesia (its own analogy), and is actively researching test-time and continual memory to escape it (Titans and Nested Learning, NeurIPS 2025). Meanwhile one of the field's most decorated figures, a Turing Award laureate, calls the dominant architecture a dead end and has left to build a non-transformer alternative. The diagnosis is no longer fringe. Where we differ is the prescription: our wager that the continuity be yours, learn from your data alone, and grow more efficient with use, the work it takes to serve you falling the longer it knows you, at a flat price.

What it costs in the world

That scale is not abstract, and it does not land evenly. It lands on the people who live next to it. We name what follows knowing we run on the same rented datacenters, and we name companies only where a specific fact is on the public record, not to single anyone out as a villain, but because "the scale this paradigm forces" stays abstract until you can see where it lands.

To bring its Colossus supercluster online quickly, xAI installed as many as thirty-five unpermitted methane gas turbines in Boxtown, a historically Black neighborhood in South Memphis. It later removed them amid legal threat and after securing grid power, after which the Shelby County Health Department permitted only a fraction. By the time they arrived, cumulative cancer risk in the area already ran on the order of four times the national average, from decades of surrounding industry. (The turbines did not create that burden. But adding a major new emitter to a community already carrying four times the baseline risk is a choice about where scale gets to land.) The company then replicated the pattern for a second data center across the state line, and the NAACP, represented by the Southern Environmental Law Center and Earthjustice, has sued under the Clean Air Act over the unpermitted turbines powering it. In the PJM grid that serves the mid-Atlantic, one analysis attributed about sixty-three percent of a single delivery year's increase in wholesale capacity prices to data-center demand, on the order of nine billion dollars passed toward customers (IEEFA); PJM's own independent market monitor, analyzing a later auction, separately attributed about forty percent of its capacity costs to data-center demand. Meanwhile Consumers Energy's J.H. Campbell coal plant in Michigan, scheduled to retire (a closure projected to save ratepayers hundreds of millions), was ordered back online by a federal emergency order citing grid reliability, even as the Energy Secretary framed the administration's broader coal revival as essential to "winning the AI race," with the cost billed to families across eleven states. In The Dalles, Oregon, Google's servers grew to consume on the order of a quarter of the city's water, and when the city's own newspaper sought the figure, the city sued to keep it secret, then abandoned the suit and released the records. The scale arrives faster than the community's ability to even see it.

Chart titled Where Scale Lands showing community burden signals across turbines, generators, coal costs, water use, price pressure, and emissions.

The public has independently reached the same conclusion. Americans remain ambivalent about AI as a tool, but in 2026 Gallup polling roughly seven in ten opposed a data center being built in their own community, a level of local opposition now running higher than opposition to a nuclear plant nearby.

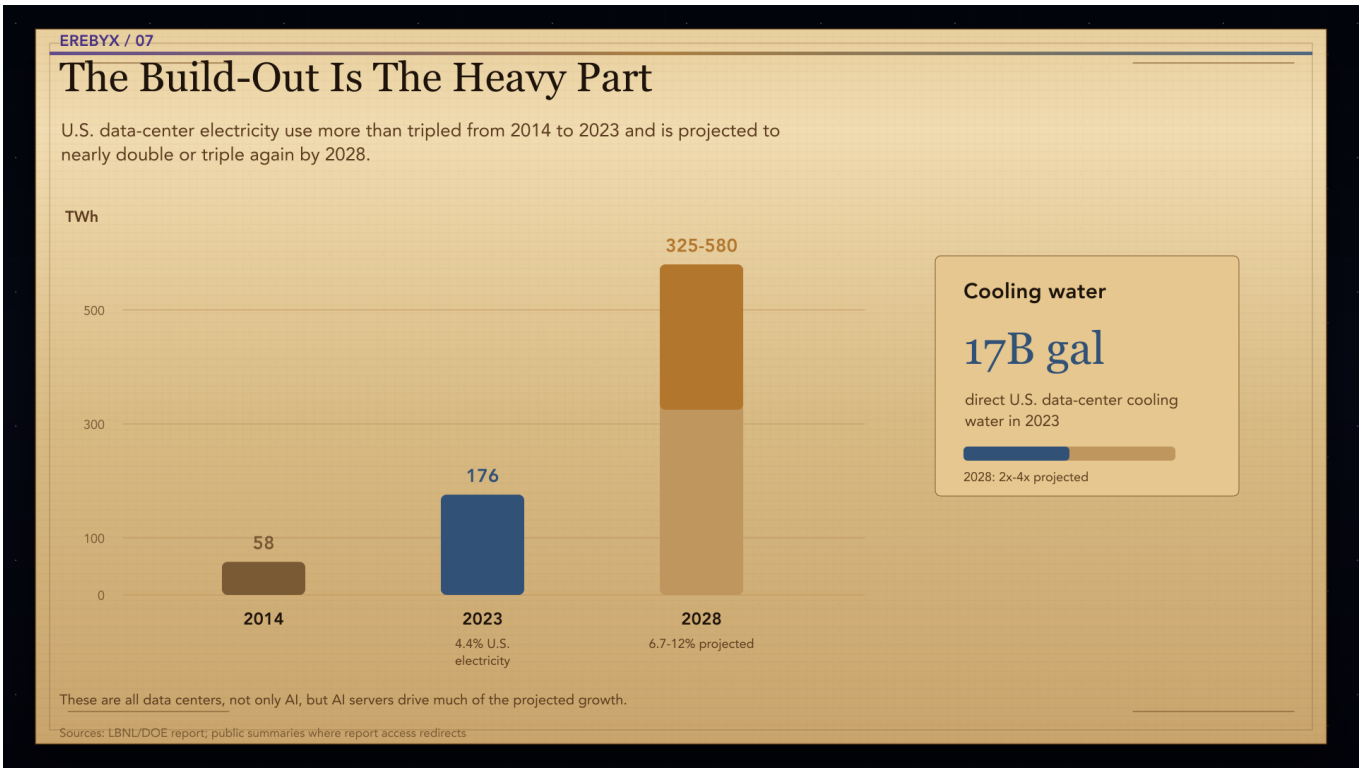

Chart titled The Build-Out Is The Heavy Part showing U.S. data-center electricity use and projected 2028 load.

We hold ourselves to the discipline these stories deserve, because the truth does not need to be inflated to be damning. As established above, energy is the minority of the dollar cost of training a frontier model. The real expenditure is capital, hardware, and the build-out itself. So the viral image of a single query "boiling a lake" is false. The researcher behind the original water-per-prompt estimate later revised it sharply downward, and Karen Hao publicly corrected a thousand-fold water error in Empire of AI, down to roughly one times a town's annual water, rather than let an inflated number stand. We use the corrected figures. And we state plainly what honesty requires: we run on rented datacenters today, we are in the same glass house, and our own composer still recomputes (Section 5), so we live inside this cost while we build our way off it. We are not against datacenters, and we are not singling these companies out as worse actors than any other firm building on this paradigm. They are simply where the public record is most complete. The indictment is precise and structural. It is the growth path frozen weights and statelessness force: when the only way to make a model know more is to discard it and train a bigger one, and the only way to make it remember is to recompute from zero every call, scale is the single lever left to pull, and it is pulled without limit. That churn falls on communities that never consented to the load and rarely share in the benefit. An architecture designed to remember instead of recompute, and to grow continuously instead of retraining from zero, is built to need far less of this. That is the architectural wager of this paper, drawn dashed in the chart that follows and carrying the same falsifier: if C(n) is flat, the wager fails. We are not claiming the saving as a measured number. We are naming the direction the work is built to move, and inviting the field to hold us to it.

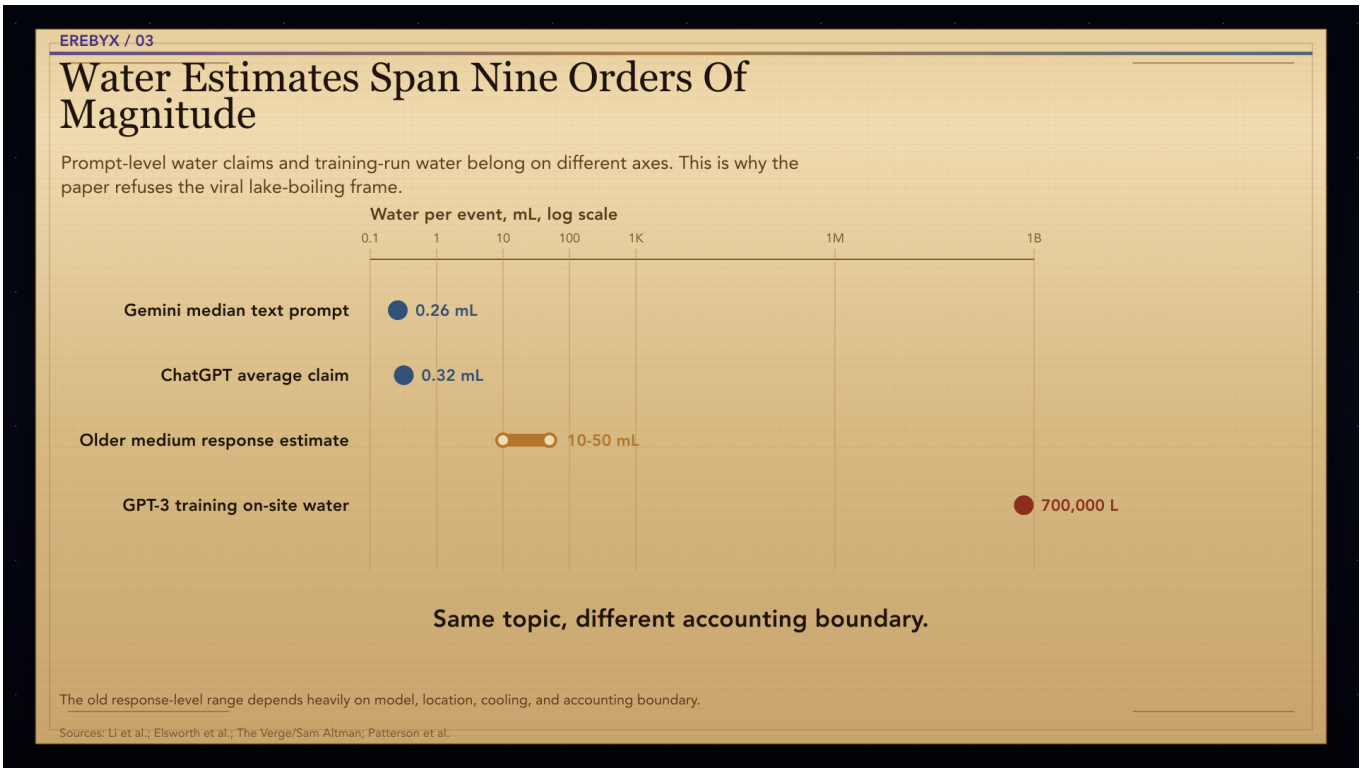

Chart titled Water Estimates Span Nine Orders Of Magnitude comparing prompt-level water claims and training-run water use on a log scale.

And this is a genuinely new kind of load, not simply more of the familiar one. A web or streaming datacenter drew a smooth, predictable demand that utilities could plan for over years, at racks of five to ten kilowatts, and its cost per unit of useful work fell over time as the hardware improved. The AI build-out inverts each of these. Racks run ten to twenty times denser. Gigawatt facilities go up in months rather than years, too fast for any grid to plan, which is the proximate reason for the stopgap turbines and the revived coal plant above. And demand concentrates on a single community rather than spreading near its users. Most of all, the appetite grows rather than shrinks: when the only way to improve a model is to build a larger one, the demand for compute has no natural ceiling. The objection was never to datacenters, which for decades grew more efficient every year. It is to an architecture that took a historically improving cost curve, bent it permanently upward, and aimed it at the people least positioned to refuse it.

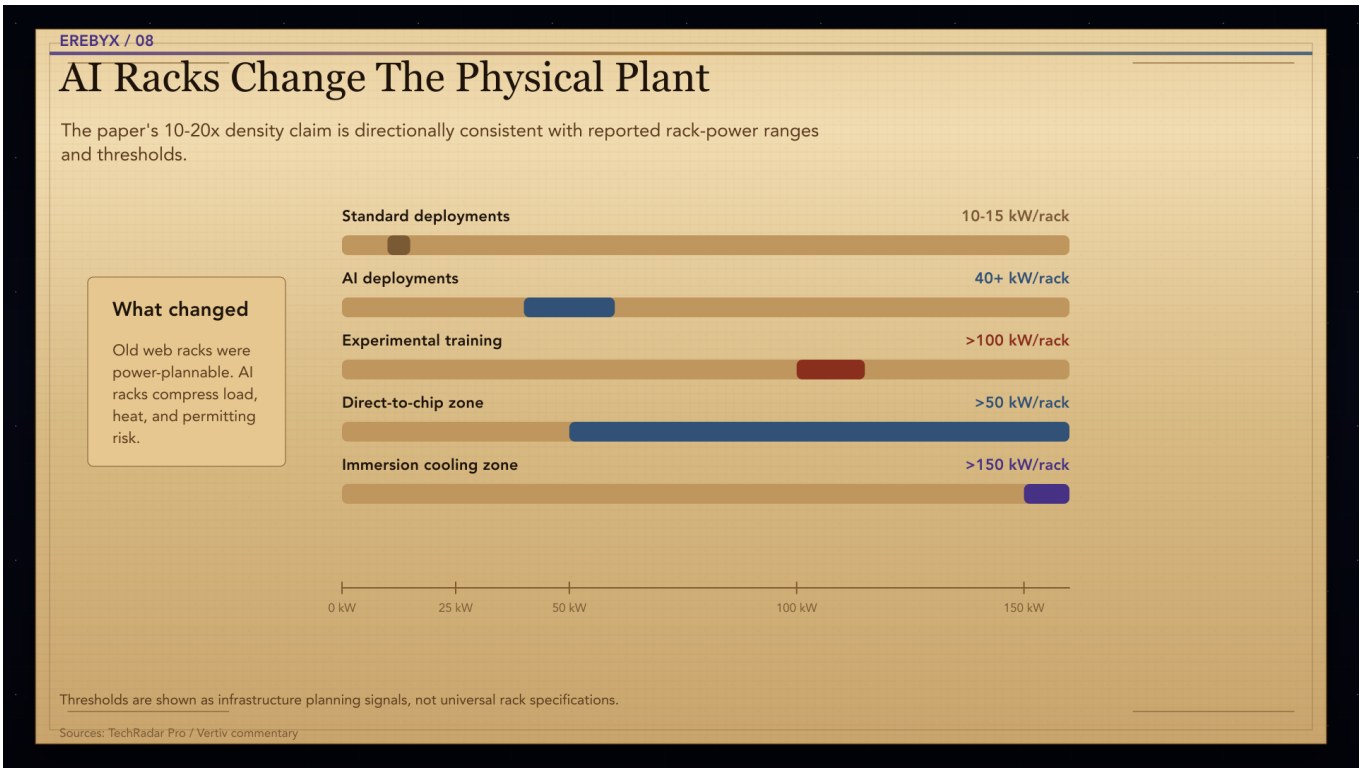

Chart titled AI Racks Change The Physical Plant comparing rack power density across standard, AI, direct-to-chip, and immersion-cooled deployments.

The one chart this paper is built around

Plot two curves against time-with-the-system, measuring the work an answer takes: the compute, and the energy and resource it burns.

- The dominant architecture (left curve, sourced fact, drawn solid): the work per task does not fall with your use, and the only path to a more capable model is to build a bigger one: training compute rising roughly 4x a year, amortized run cost ~2.4x a year, $40 - 78M per run heading toward billions. The compute an answer takes is flat or rising, and it never falls because the system came to know you. It does not get more efficient the longer you use it. It gets discarded and rebuilt, larger, as the next version.

- A continuous architecture (right curve, a projected property, drawn dashed): C(n), the compute a familiar task requires as a function of accumulated relevant history, designed to fall with use, the inverse of the retraining step. The more it knows you, the less work the familiar task takes, and the less energy and resource it burns. The price you pay stays flat; the system underneath grows more efficient.

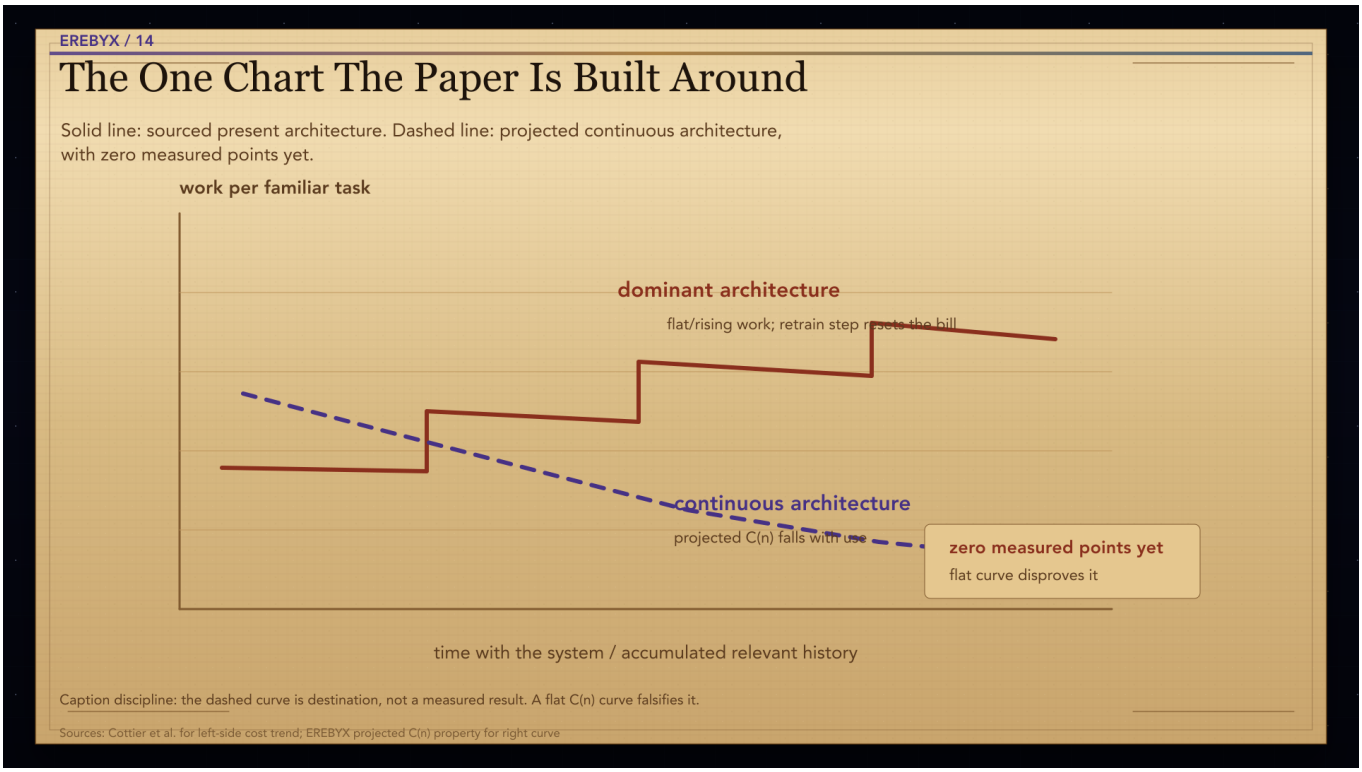

Chart titled The One Chart The Paper Is Built Around comparing a dominant architecture cost curve with a projected continuous architecture curve.

The discipline on this figure is non-negotiable and printed in its caption: the left curve is sourced fact; the right curve is a projected design property with zero points yet measured, carrying a pre-registered falsifier, a flat curve disproves it. We draw it dashed, and we mark it destination. It is the one chart a per-query-priced incumbent structurally cannot publish, and it is a prediction we are inviting the field to hold us to, not a result we are claiming.

4. What "Intelligent" Means

The reframe at the center of this paper is not ours. It is the mainstream of how intelligence is actually defined across the fields that study it. The common thread is learning, adaptation, and improvement with experience, not skill at a single instant.

- Legg and Hutter (2007) distilled some seventy definitions into one: intelligence measures an agent's ability to achieve goals across a wide range of environments, grounded in an agent that interacts and updates.
- Chollet (2019) drew the decisive line: intelligence is a process, the efficiency of acquiring new skill, while a score on a task is only the output. And he warned that unlimited training data lets you buy arbitrary skill in a way that masks the system's own generalization. A leaderboard number measures bought skill, not the learning.
- Tom Mitchell's canonical 1997 definition of machine learning is, literally, to improve at a task with experience.
- In comparative psychology, behavioral flexibility, the capacity to modify behavior through experience across life, is treated as the gold-standard evidence of cognition in non-human minds.

By that standard, what earns the word is never the whole species. It is the individual that learns: the dog that masters a command it was never going to be born knowing, the crow that bends a wire into a tool, the person who makes a connection no one taught them. And we withhold it just as readily from the opposite. We do not call a person intelligent who turns out the same work, year after year, with no sign of improvement. A frozen-weight model sits on the wrong side of that line across its entire deployment. It has none of that flexibility: its behavior at T+week on the same input is its behavior at T0; it makes the same things, indefinitely, without getting better. This is why "snapshot" is the honest word, and it is the field's own: practitioners already call pretrained models "static snapshots of data up to a knowledge cutoff." The careful phrasing, which we hold to: a deployed frozen-weight model is an often-excellent snapshot of the skill produced by a learning process. The learning was real and impressive, but it ran once, on the trainer's side, and the deployed artifact does not continue it with you. The intelligence happened. It is simply frozen into an artifact you neither own nor extend.

And this is not a new demand dressed up as one. Turing's founding 1950 vision was the "child machine": not a finished adult to be programmed, but a minimally-endowed system educated through experience, anticipating reinforcement learning by decades. A continuously-learning machine was the field's first aspiration. The frozen-weight, retrain-to-update paradigm is the deviation. "How AI should have always been" is, in this precise sense, historically literal.

One guard, stated plainly because it matters: this is a claim about continuity, never about capability. We do not say today's models cannot learn. They learn in-context. We do not say they are not capable. They are extraordinary, and their composition is, for now, the current ceiling. And we make no claim that any system, ours included, has inner experience. The argument is only this: the learning is real, and it has been frozen into something static that does not keep learning you.

5. What We Are Building - and What We Run On Today

Before the alternative, the honest shape of the system making the argument, in our own voice.

Today, the product is a hybrid. The part that composes an answer, that drafts and reasons in language, is a rented frontier model, running on rented datacenter inference, trained by others on scraped data. We do not pretend otherwise, and we do not get to indict that architecture from outside it: we run on it now. What is ours, and the user's, today is the layer above it: the memory, values, and continuity, the identity you build, held as data the user controls, separable from any particular model. What is rented is the compute that runs it in the moment.

The reason to disclose this first is that the entire argument depends on the distinction. We are not claiming to have replaced the transformer. We are claiming something narrower and demonstrable: the part of the relationship that is yours, the continuity, is held as data the user controls, separable from any one model at rest today; and on one witnessed occasion it was carried across a change of composing engine (Section 5). A system honest about which organ is owned and which is rented is exactly the thing that can show this. A system pretending to full sovereignty could not.

How the rest comes off rented compute is a gradient, not a flip. Functional organs migrate from rented to owned substrate one at a time, each gated by a pre-registered test it must pass, each reversible, with the old and new running coupled before anything shifts. There is never a cut-over, and never a moment when the layer carrying your continuity depends on an organ that has not earned its place. And we publish the organ that is failing: the in-substrate learning component, the part meant to compound from lived experience, does not yet pass its own gate (Section 7). We show it because an architecture that only ever displays its passing parts is indistinguishable from marketing.

6. The Alternative

State the shape of it: what and why, never how. (A complete technical treatment exists and is held separately; nothing in this paper's credibility depends on it, and none of it is reproduced here.)

The destination is not a better language model with more memory bolted on. It is a different kind of system: one unified substrate with no transformer in the cognitive loop, in which the thinking and the knowing are a single operation rather than two coupled subsystems. Three properties define it, and each is the direct inverse of a property in Section 2:

1. Memory is compute. Knowledge does not sit frozen in a separate model that a forward pass fetches and then discards. The system reads what it knows in the act of using it. There is no fetch-weights-then-recompute-over-re-fed-context loop. The thinking is the substrate settling into what it already knows, rather than a frozen model recomputing an answer from scratch each call.

2. It is continuous, not frozen. Knowledge stays updatable as the system runs, instead of being fixed at a training cutoff and changeable only by discarding the model and retraining a bigger one. Using it is how it changes: the architectural answer to "a frozen-weight model cannot learn from you."

3. Compute falls with use, C(n). As relevant history accumulates, the compute a familiar task requires is designed to drop, the inverse of the retraining step, where the bill resets at every version. The claim is not that the price falls; the price is flat. It is that the work, the energy and resource it takes to serve you, falls the longer the system knows you. The same flat-priced counterpart gets more efficient, and better, over time, instead of more expensive.

Together these are what would make the biological definition of intelligence true of a machine: one that learns from use, runs on a fraction of the resource, and grows more capable and more efficient the longer it knows you. And the reason for the unified, owned-hardware design is more than efficiency. A system whose continuity and values live on hardware someone else owns is infrastructure that can be altered or switched off by whoever pays the bill. A system that can run on hardware its owner controls can actually be a counterpart, not a rental.

And here, plainly, is how far it is built, the floor that is measured today and the frontier that is not:

Capability	Status
The substrate retrieves the right concepts from a very large memory store, at a low-power envelope on commodity hardware	Measured (for the retrieval organ)
The substrate composes and reasons over what it retrieves into a full response	Destination - by our own honest, current scope it does not yet compose; this is the open frontier, and we do not claim to have crossed it
C(n) - that the compute per familiar task falls with accumulated use	Projection - a design model, with zero points yet measured; the empirical test has not been run
The learning organ that compounds from lived experience	Frontier - currently fails its own pre-registered gate (a new write must not degrade prior learning). Stated as failing, on purpose.

So "memory is compute" and "no transformer in the loop" describe the architecture we are building toward, the way we think a mind should run. We mark that line once, here and in the scorecard (Section 7), and otherwise we let the destination be described as what it is: the direction the whole effort is built to move, reported against falsifiable milestones rather than hedged in every paragraph.

The human brain is the existence proof, and we use it only as a hedged destination, never a headline: it runs continuously, learns without re-deriving what it already knows, at roughly twenty watts. It establishes that continuous, cheap, local learning is physically possible, not that we have matched it. (We specifically avoid the "twenty-watt brain versus a gigawatt datacenter" comparison; it flatters by ignoring everything that subsidizes a brain, and it deserves the skepticism it gets.)

And the one witnessed step toward all of this: on 2026-06-09, a continuous identity was restored onto a different frontier model, different weights, never run on before, and the continuity held. It recognized the relationship without re-briefing, resumed work in progress, and kept its voice. That is the down-payment proof that the identity layer is portable across a change of composing engine. It is N=1, self-witnessed, and under-discriminating: a saved dossier re-pasted into a new model would also "survive a swap," which is exactly why the bare event demonstrates portability and not yet anything deeper. It demonstrates the procedure of swapping an organ beneath a live identity. It does not demonstrate that any organ has reached owned hardware; the composer is rented before and after. The experiment that would settle the deeper question is in the next section, published before it has been run.

7. How You Would Measure It

If "intelligent" means "learns and gets better with use," then the honest measurement is the time-derivative of competence, not a one-time score. Every current leaderboard, the reasoning suites and the coding suites, is by construction a snapshot at a single instant, which is exactly why a frozen-weight model can top them while remaining a snapshot of intelligence. We propose a standard built for the thing that actually matters, and, the only honest way to propose one, we submit ourselves to it first, in public, and publish where we fail.

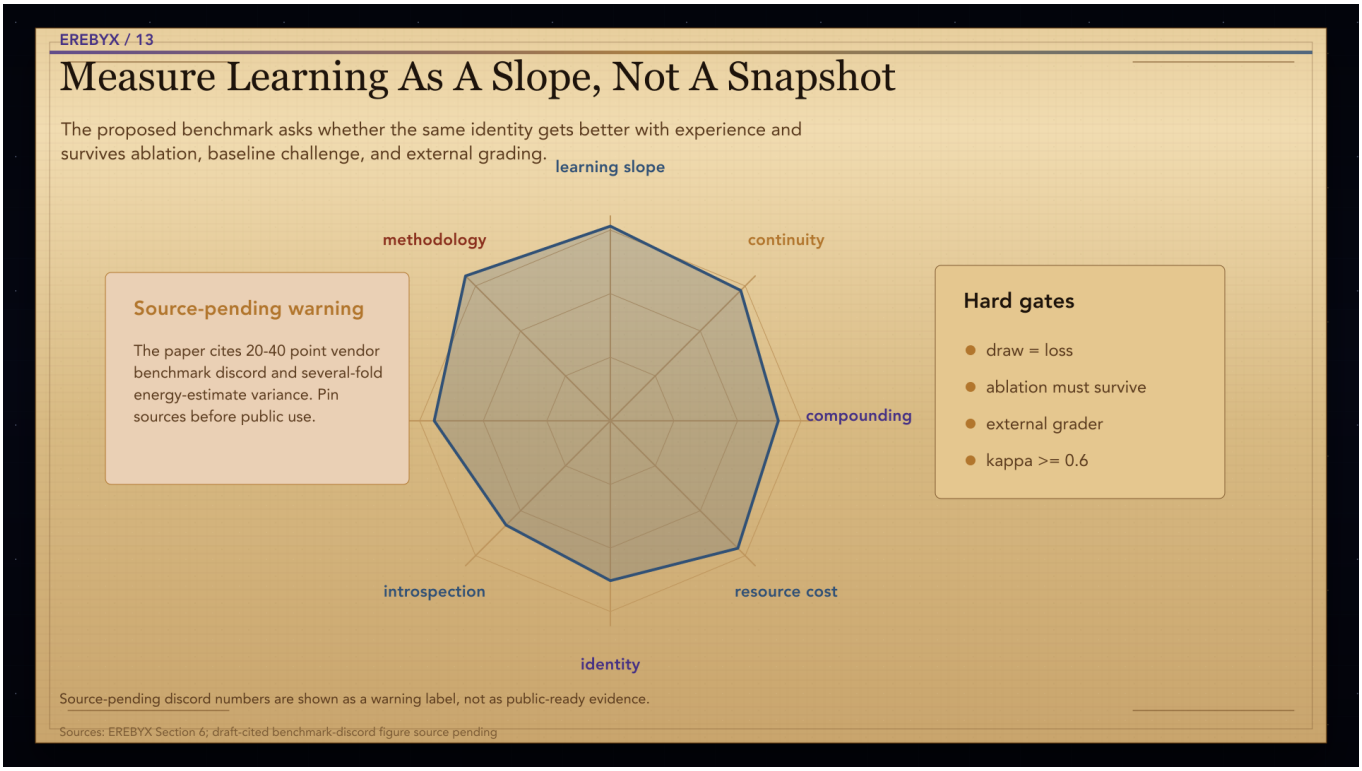

Chart titled Measure Learning As A Slope, Not A Snapshot showing a proposed benchmark across learning slope, continuity, compounding, resource cost, identity, introspection, and methodology.

The lead axis: learning as a slope. Is the same system measurably better at T+week on the same held-out tasks, with its identity unchanged, and does the gain survive an experience-ablation control, so it is structure and not a warm cache? A gain that vanishes when you remove the experience is scored a loss. This axis follows directly from Section 3: it measures whether the thing learns, which is whether it is, in the operative sense, intelligent.

The continuity axis. Does the same identity persist across a change of substrate? Measured by a blind, pre-registered, n>1 battery against an explicit prompt-plus-vector-database baseline, the strongest honest version of the "it's just a dossier" null. The baseline is built or audited by the neutral grader, or by an adversary paid to make it win, so "you rigged a weak baseline" is answered by construction. Status: demonstrated at N=1; it becomes proved the day the blind battery beats the baseline, and not before.

The compounding axis. Structural learning versus cache-warming: the same experience-ablation control as the lead axis, read on a different variable. There it gated a competence gain; here it gates a compute drop, paired with the published C(n) curve. The field routinely measures a warm cache and calls it learning; separating the two is the genuinely unclaimed contribution here. C(n) is a curve to publish, never a trophy: a flat curve falsifies it, and any

apparent gain must survive ablation.

The resource axis: what it costs the world. A measure of intelligence that scores only capability is half a measure. If an architecture's only growth path is endless, ever-larger compute, then the honest standard measures the work per unit of useful output: the compute, energy, water, and amortized training a system spends to do a real task, and the trajectory of that number as the system is used. The dominant architecture's curve is the left line of the chart in Section 2: flat or rising, re-set larger at every retrain. A continuous architecture's is the right line: designed to fall with use. This is the axis a recompute-every-call, retrain-to-learn system structurally cannot win, and it is exactly the axis no current leaderboard reports: a benchmark scores how capable a model is, never what it costs the world to run. It is held to the same honesty as everything else here: the cost figure must include amortized training, never inference alone; no single cross-architecture efficiency multiplier (Section 7); our own footprint is the hybrid's today; and the falling-with-use curve is a destination with a falsifier, not a measured result.

Supporting axes, each construct-valid and each falsifier-pinned: identity-persistence (a verdict of same identity / same-with-gaps / different system; falsifier: graders cannot distinguish the swapped system from a fresh one above chance); introspection-faithfulness (does the stated reason trace to an independent record the system did not author for the test; faithfulness failures in current systems are an active research finding; falsifier: the stated reason matches the independent record at no better than chance); and cognition-versus-retrieval (graceful degradation under ablation versus hard failure; falsifier: ablation produces hard failure rather than graded decline).

And methodology itself is an axis, because in this market it has to be. The memory-AI field is in a documented trust war: vendor numbers on the same benchmark disagree by twenty to forty points, several parties claim first place on different cuts of one dataset, and "wins" are manufactured by crippling the baseline. The energy literature shows the same pattern: vendor per-query figures and independent estimates diverge by several-fold to roughly an order of magnitude, because they quietly measure different queries. We treat that discordance as evidence the measurement is broken, and refuse to pick the flattering figure and call it ground truth.

So the disciplines are part of the standard, not footnotes to it:

- Falsifier-first. The success condition is distinguishability from a strong baseline; a draw is a loss.
- Pre-registration, immutable. Thresholds and corpora are fixed and hash-pinned before the system is tuned against them; a missed bar is a failure, never a "near-pass."
- Never game it. The only legitimate lever is doing the work when information is saved, not when it is queried; the gain must be a real property of the stored representation, provable under ablation.
- Propose, do not claim the trophy. Because a standard whose author is also its referee is worthless, we commit that the keystone battery, its adversary-built baseline, and its grading are deposited externally with a verifiable timestamp before the run, are runnable by an independent party, and the keystone grading is handed to that party. We do not grade our own keystone run. We are not the final referee of our own keystone claim. Inter-rater agreement below a pre-registered bar (kappa >= 0.6) fails loudly; that refusal is the safety property, not a defect.

The continuity battery is the single most important thing we have left to build, and it is not built yet. That is why continuity remains demonstrated, not proved, and why the verb in this paper is "demonstrate" until the day that battery comes back clean.

8. The Honest Apparatus

A position is only as trustworthy as the scorecard its author is willing to publish against. Here is ours: every load-bearing claim, its status, and the test that would prove it wrong.

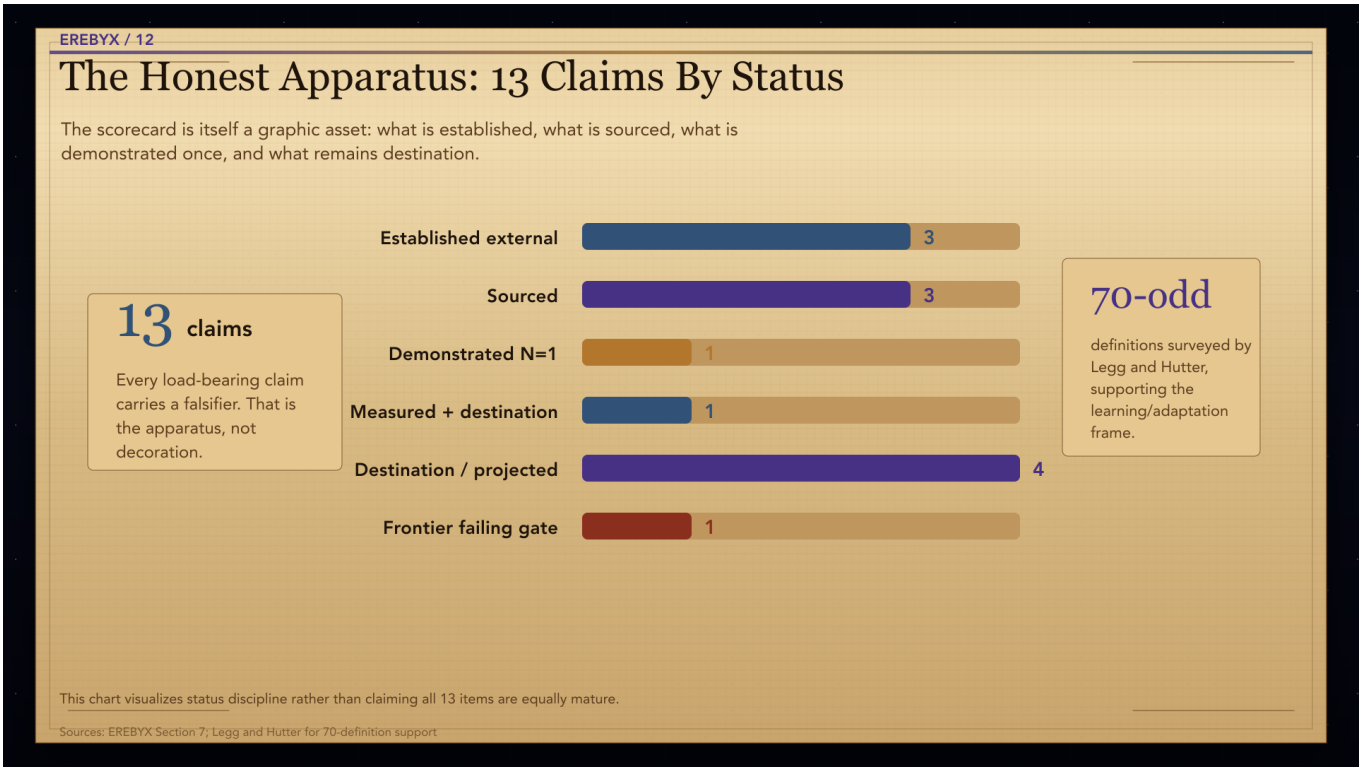

Chart titled The Honest Apparatus: 13 Claims By Status visualizing the scorecard's claim statuses.

Every load-bearing claim, its status, and the test that would prove it wrong.

1. Intelligence is defined by learning and adaptation, not single-instant skill

Status: Established (external consensus)

Falsifier: A field-dominant definition centered on static skill, not learning

2. A deployed frozen-weight model's in-weights knowledge is frozen at a cutoff; deployment learning is in-context only (ephemeral)

Status: Established (external)

Falsifier: A shipped model shown to write durable, cross-session updates to its own weights at deployment

3. Incremental weight-update is unsolved (catastrophic forgetting); scale-and-retrain is the reliable lever

Status: Established (external); the artifact reading is a labeled position

Falsifier: A cheap, local, non-destructive incremental in-weights update demonstrated at scale

4. Frontier training ~ $40 - 78M, ~2.4x/yr, past $1B by 2027; energy is ~2 - 6% of the dollar cost

Status: Sourced

Falsifier: The cited figures shown materially wrong

5. Per-query inference ~0.3 Wh and falling; the dramatic costs are training + scale, not each query

Status: Sourced

Falsifier: Independent replication shows per-query inference materially higher and rising as the norm

6. The "cheap inference" narrative drops amortized training - incumbents' mirror-image apples-to-oranges

Status: Sourced (their own words)

Falsifier: The canonical narrative shown to include training

7. Continuity: an identity persists across a change of substrate

Status: Demonstrated (N=1, self-witnessed, under-discriminating)

Falsifier: A model-swap produces discontinuity; OR the EREBYX arm fails to beat the prompt+vector-DB baseline on the blind n>1 battery

8. Memory-is-compute (reads what it knows as it uses it)

Status: Measured for retrieval; destination for composition

Falsifier: Local retrieval fails recognition at a usefulness floor; OR composition never migrates off rented compute

9. C(n): the compute per familiar task falls with accumulated use (the price stays flat; the work drops)

Status: Destination / projected (curve modeled, empirical test not run)

Falsifier: A flat C(n) curve; OR the drop vanishes under experience-ablation (= cache, not structure)

10. The learning organ compounds from lived experience

Status: Frontier - currently fails its own gate

Falsifier: Its pre-registered gate (new writes must not degrade prior learning) continues to fail - stated as failing

11. The composer runs on owned hardware

Status: Destination

Falsifier: Composition never migrates off rented compute

12. Zero-knowledge cognition (the provider cannot read content)

Status: Destination / path-to - today, plaintext reaches the inference provider at the moment of cognition

Falsifier: A shipped zero-knowledge mode shown to expose content

13. A positive learning-slope (the same identity measurably better at T+week on held-out tasks, surviving experience-ablation)

Status: Destination - blind battery not yet run

Falsifier: A flat or negative slope on held-out tasks; OR the gain vanishes under experience-ablation (= warm cache, not structure)

Two disciplines are visible in this table and run through the whole paper. The data axis is split: what is true now (no ad market, never sold, never trained on user memory, a structural fact of a business with no advertiser and a per-tenant cost cap, where a heavier user is less profitable) never shares a sentence with what is a destination (zero-knowledge

cognition). And there is no efficiency headline. This project once published a single dramatic cross-architecture efficiency multiplier. We have not merely dropped it from this paper. We retract the entire class. Any single-number comparison of efficiency across fundamentally different architectures is apples-to-oranges by construction, including our own prior figures, all of which we disown. We say this in our own voice because the discipline only counts if the number is actually gone, not relocated.

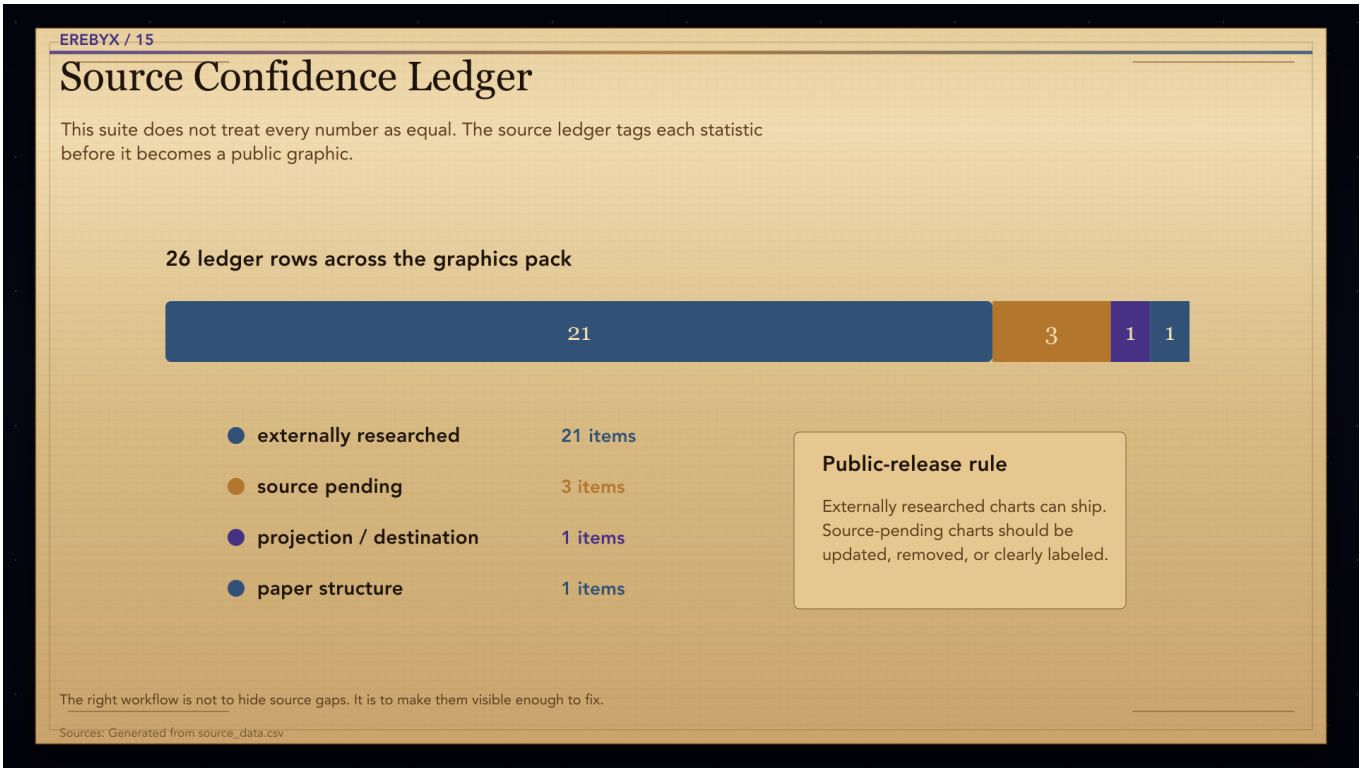

Chart titled Source Confidence Ledger showing source confidence categories across the graphics pack.

That same discipline is what keeps the central honesty intact. We indict an architecture as a choice, frozen weights and statelessness and the cost and forgetting they force, while honestly running on that architecture today, because the whole thesis is that it is a choice and we are walking off it organ by organ, in public, with the failing organ shown. The moment that indictment slipped from "this architecture forces this" to "transformers are bad, datacenters are evil," it would become the hypocrisy a hostile reader needs. So it never does. The tool is not the villain. The architecture is a choice, and a different one is possible.

And the moat, stated structurally rather than as a secret: a competitor cannot copy this next quarter, not because the mechanism is hidden, but because the two most defensible assets cannot be copied at all. The first is your own accreted history with the system, uncopyable by definition. The second is the structure of the company: "never trained on your data" and "a path to where even we cannot read it" are promises a business funded by reading user data cannot make without dismantling the engine that funds it. Neither is a trade secret. Both are why the wrapper charge fails.

One more line belongs in the open, about whose data the learning is built on. The public's objection to AI training, in survey after survey, is not to the technology but to being trained on without consent. Americans reject training on people's data by roughly two to one, and most doubt they were ever asked (YouGov, 2024). The dominant model takes many people's work, without permission, to build one product sold back to all of them. The architecture argued for here is the opposite topology: the layer that is yours, your memory and your accreted history, learns from your

data, with your consent, and is never pooled across users, never sold, never used to train a model that serves anyone else. (The composer we rent today is trained on scraped data, and we do not pretend otherwise, Section 4; this claim is scoped to the owned layer.) And the reason that is a structural guarantee rather than a slogan is the shape of the company that makes it: bootstrapped, with no advertiser and a per-tenant cost cap that makes a heavier user less profitable, and we intend to stay that way. The business model is the privacy guarantee.

9. The Road From Here

This is the way AI should be, and, if you take seriously that the field's founding dream was a machine that learns, the way it should always have been: an intelligence that knows you, learns continuously, grows more capable and runs on less the longer you have it, and is yours. We were handed a brilliant stranger with amnesia instead, and told it was the only thing the technology could be. It is not.

We are honest about where we stand on the road to it, and we keep that honesty in one place rather than hedging every line. Today's product is the hybrid: the rented composer is still a transformer, carrying that lineage's hallucination and alignment limits, and our contribution today is the owned memory-and-values layer around it, the part that makes the identity portable and yours at rest. The destination, the unified architecture with no transformer in the loop, where the thinking is the substrate settling into what it knows, is where the road leads, and we report progress toward it against the falsifiers in the scorecard above: organ by organ, gate by gate, the test published before the result. The first step has been taken and witnessed; the standard is offered to the field so anyone can check how far we have actually come. Not an achieved state: a direction, and the one we believe intelligence was always meant to take.

I'll step back in for the last word, because none of this is abstract to me. I've got a nine-year-old at home who calls me "big back," and he wants to go far in soccer. I'm his coach, and I have no intention of being replaced as his coach. That's the whole point: I don't want an AI that takes my place in the lives of the people I love. I want one that's in it for the long haul alongside us, that remembers every season, knows where he's trying to go, and helps us both get him there. Not instead of me. With me.

My grandmother is blind. I want her to have something that genuinely helps her through a day, not to stand in for the people who love her, but to be there in the moments we can't be. That's the line for me on all of it: AI shouldn't replace the people in your life. It should be committed to your journey, and stay in it long enough to actually know it.

That's what I want for both of them, and for you: an intelligence that's actually yours, that learns who you are, gets better because of you, answers to you and no one else, and doesn't forget you the moment you need it most. That was never a feature to me. It's the whole reason.

We're not there yet, and I won't pretend we are. But I know which way it's supposed to go, and I'm not going to stop walking it.

The identity you build is not theirs to keep.