INSTITUTE FOR PHYSICAL AI @ BMI
The Charlot Lab

Technical Report TR-2026-25
Survey / Review · v0.3
2 August 2026

THE FUTURE OF COMPUTE AS A FUNCTION OF ENERGY

The Energy-First Turn

Post-transformer, learn-while-inferring AI, the substrates built to run it, and the economics of changing what a unit of intelligence costs.

Institute for Physical AI @ Bailey Military Institute

Charlot Lab ecosystem · Energy First Architecture (EFA) research program

Synthesis drafted with AI assistance from public sources. Every quantitative claim carries a verification status in the companion database; figures are labelled measured, reported, claimed or modelled, and a simulated result is never presented as a measured one.

Abstract. Computing's next constraint is not transistors, capital or talent: it is electricity. Data-centre demand is doubling this decade while physics permits computation millions of times cheaper than we practise it, because the bill is paid to move data rather than to compute it. This review maps the escape along two coupled axes — the substrates that execute relaxation natively, and the model classes that learn while inferring — then prices the transition and states what evidence would falsify the thesis.

The Future of Compute

as a Function of Energy

The Energy-First Turn: post-transformer, learn-while-inferring AI, the substrates built to run it, and the economics of changing what a unit of intelligence costs

Working draft v0.3 — August 2, 2026

Prepared in support of the Energy First Architecture (EFA) research program

Institute for Physical AI @ Bailey Military Institute — Charlot Lab ecosystem

Synthesis drafted with AI assistance (Claude, Anthropic) from public sources. Every quantitative claim carries a verification status in the companion workbook (energy_first_global_database.xlsx); an interactive companion (energy_first_explainer.html) animates the core mechanisms and runs the forecast model live; two companion specifications (EFA-RFC-001: OER/1 receipts + DRIFT/1 benchmark; EFA-RFC-002: the RELAX/1 cross-paradigm dialect) formalize this document's central recommendations. Internal review draft — verify flagged figures before external publication.

How to Read This Document

This review is written to be legible at three depths. Choose your path:

- **Ten minutes:** read "The Whole Argument on One Page" (next page) and the Executive Summary, then Section 10 (the forecast) and Section 12 (recommendations). Everything else is evidence.
- **One hour:** add Section 2 (why computing costs what it costs — the physics primer), Section 4 (the map), Section 7 (the software layer), and Section 8 (regions and capital). This is the shape of the field.
- **The full read:** read straight through. Sections 5, 6, 7, 9, and 11 carry the technical detail, the benchmark argument, and the failure-mode analysis. The glossary at the back defines every term of art.

Two conventions hold throughout. First, provenance: figures are labeled measured, reported, claimed, or modeled — a simulation result is never dressed as a measurement, and vendor ratios are never laundered into headlines. Second, plain-terms boxes: each technical section opens with its intuition in one or two sentences, marked in the left margin, so a reader can follow the argument without following every equation.

The Whole Argument on One Page

1. **The bill is real.** Computing's next constraint is not transistors, capital, or talent. It is electricity. Data-center power demand is doubling this decade; AI capacity is already gated by grid connections, not chip supply. The joule has become the denominator of intelligence.
2. **The gap is architectural.** Physics permits computation millions of times cheaper than we practice it. The thermodynamic floor for erasing one bit is about twenty-one orders of magnitude below a watt-second; today's arithmetic sits roughly a hundred million times above that floor, and the memory round trip a further factor of ten to a hundred above the arithmetic. The gap is not law — it is architecture.
3. **The brain is the benchmark.** Nature already filed the existence proof. A human brain runs general intelligence — perception, planning, continual learning — on about twenty watts, spending on the order of femtojoules per synaptic event, because it computes in place, sparsely, and only when something happens.
4. **The lottery, not the optimum.** Today's AI stack was shaped by the hardware that was lying around, not by energy. The GPU rewards dense, synchronous matrix multiplication with frozen weights; model classes that violate that contract never got to compete. Change the fitness function to joules-per-answer and different winners emerge.

5. **Settle, don't shuttle.** Those winners exist and share one move: they treat computation as relaxation. Energy-based models predict by settling into low-energy states; learn-at-inference architectures write memory while running instead of shuttling gradients to a datacenter; active inference updates beliefs in a single pass. The 2024 Physics Nobel to Hopfield and Hinton honored exactly this lineage.
6. **The hardware is landing.** The substrates that execute relaxation natively are arriving on every continent: thermodynamic sampling chips in the US, brain-scale neuromorphic systems and photonic pilot fabs in China, Ising machines sold as services in Japan, processing-in-memory standardized by Korea's rival memory giants, and a public neuromorphic stack in the UK and EU.
7. **The money moved.** Capital noticed in 2025-26: a record $475M seed for Unconventional AI at a $4.5B valuation; $110M and $220M photonic rounds; the US CHIPS office signing two letters of intent in one July week; probabilistic silicon already shipping to Boeing and CERN. Direct paradigm investment now totals several billion dollars.
8. **Software is the tell.** A parallel war one level up prices the thesis: Qualcomm paid $3.9B for Modular's write-once-run-on-any-chip stack; China open-sourced its CUDA rival and now ships frontier models optimized for domestic chips on day one. Universal software does not make hardware irrelevant — it makes hardware comparable, and comparable hardware competes on joules. But every such layer speaks kernels-and-tensors; none can address a chip whose native act is relaxation. The cross-paradigm layer is unbuilt and unclaimed.
9. **One record held, one open.** One capability record already belongs to the new paradigm — microsecond constrained decisions, held by Japanese settling machines in production. The open prize is learning under drift: agents that adapt on-device, scored by regret, latency, and receipted joules. Frozen-weight architectures cannot top that leaderboard without becoming something else.
10. **The economics compound late but large.** Priced — amenable share, adoption friction, efficiency gains set below vendor claims — the paradigm saves roughly 91 TWh per year by 2035 in the base case, about $7B per year in energy and $120B in data-center capital never spent, with the aggressive case roughly double. The forecast's sensitivity is itself the strategy: moving workloads into the amenable column beats another 10x on workloads already there.
11. **The connective tissue decides.** Every prior alternative-computing wave died of unverifiable claims, missing toolchains, or absent model classes. The survival conditions this time are receipts (provenance-tagged joule measurements), a compiler that lowers learned energies onto physical

substrates, and a benchmark the incumbent cannot fake. Building those three is cheaper than any chip — and unlocks all of them.

Contents

Executive Summary

Two constraints are converging on the AI industry. The first is physical: global data-center electricity consumption reached roughly 485 TWh in 2025 (up 17% in a year, with AI-specific facilities growing 50%), is forecast at 565 TWh for 2026, and is on track to roughly double again by 2030 — while power capacity climbs from 132 GW toward 290 GW and the industry's own analysts declare power availability, not chip supply, the binding constraint. The second is architectural: the dominant model class cannot learn after deployment; its weights freeze at training time, adaptation is a data-center event, and inference spends identical computation on trivial and profound predictions alike. Both constraints trace to one root, which Sara Hooker named the hardware lottery: the transformer exists as it does because the GPU rewards dense, synchronous, feed-forward matrix multiplication — and model classes violating that contract never got to compete.

This review maps the escape effort along two coupled axes. Substrates: thermodynamic and probabilistic samplers (Extropic — now holding a $75M CHIPS letter of intent — Normal Computing, Signaloid's commercially shipping uncertainty-native silicon, the UCSB and Tohoku p-bit programs); Ising machines sold today as services (Toshiba, Fujitsu, Hitachi, NEC, NTT, D-Wave); neuromorphic systems at brain scale (Zhejiang's two-billion-neuron Darwin Monkey at two kilowatts, Intel's Hala Point) and at record valuation (Naveen Rao's Unconventional AI: a $475M seed at $4.5B); photonic processors and fabs (Tsinghua's Taichi line, SJTU's LightGen, China's pilot fabs, newly funded Neurophos and OLIX, GlobalFoundries' $300M CHIPS photonics LOI); processing-in-memory entering JEDEC standardization (Samsung and SK hynix, rivals cooperating); and settling fabrics with joule-denominated accounting (Klere). Model classes that learn while inferring: Energy-Based Transformers that think by descending a learned energy; active inference (VERSES' AXIOM); fast-weight and test-time-training architectures (Titans, TTT, DeltaNet); equilibrium propagation for physical networks; ternary LLMs (BitNet); and the Energy First Architecture (EFA), which unifies world model, planner, memory, verifier, and discoverer as readings of one learned scalar energy.

Three findings organize the map. First, one capability record already belongs to the paradigm: microsecond constrained decision-making, held by Japanese relaxation machines in production. Second, the open prize — the prospective AlexNet moment — is learning under drift, scored by regret, decision latency, and receipted joules; no frozen-weight architecture can top it. Third, the connective tissue is missing: no compiler lowers learned energies onto heterogeneous physics, and no standard prices computation in joules. The EFA program, the Ferric compute layer, and Klere-style energy receipts are among the few coherent attempts to build

precisely that missing middle — and this revision formalizes the recommendation into two companion specifications (OER/1 energy receipts with the DRIFT/1 benchmark, and the RELAX/1 cross-paradigm dialect). A fourth finding joins them: the software-portability war one level up — punctuated by Qualcomm's $3.9B acquisition of Modular and China's open-sourcing of its CUDA alternative — makes hardware comparable rather than irrelevant, and comparable hardware competes on exactly the joule metrics this program specifies (Section 7). The companion forecast, under deliberately conservative assumptions, values the paradigm at roughly 91 TWh per year saved by 2035 (base case) — a mid-sized European country's electricity — about $7B per year in energy, $120B in avoided data-center construction, and 32 Mt of annual CO2, with the aggressive case roughly double.

1. The Bill: AI's Energy Demand in Hard Numbers

> **In plain terms:** *Before any argument about how computing should change, here is what staying the course costs — in the incumbent's own accounting.*

The numbers below are the invoice for scaling intelligence on hardware that pays for every bit twice: once to compute it and again — far more — to move it. None is a projection of the paradigm this review maps; all are mainstream forecasts of the current path.

Figure	Value	Source and note
Global data-center electricity, 2024 -> 2025 -> 2026	415 -> ~485 -> 565 TWh	IEA (2024 base; 2025 +17%, AI facilities +50%); Gartner forecast for 2026
Global data-center electricity, 2030 / 2035	~945 / ~1,200 TWh	IEA Energy & AI base case; Deloitte similar (1,065 by 2030)
Data-center power capacity	104 GW (2025) -> 132 GW (2026) -> 290 GW (2030)	Gartner, Jun 2026: 'AI capacity is now constrained by power availability'
US data-center power demand	31 GW (2025) -> 41 (2026) -> 66 GW (2027)	Goldman Sachs Research; US share of peak summer demand 4.1% -> 8.5%
US structural power shortfall	9.3 GW (2026) -> 45 GW (2028)	Goldman Sachs Research; ~34 million households' equivalent
Hyperscaler AI infrastructure capex, 2026	~$725B (+77% on 2025's $410B)	FT-compiled earnings guidance, big five
One US state, already		

	26% of Virginia's electricity (2023)	State data — before the current buildout accelerated
Share of AI energy spent on inference	60-90%	Meta ~70%, Google ~60% of ML energy, AWS 80-90% of compute — the bill is per-answer, forever
Energy per frontier query -> per long reasoning query	0.31 Wh median -> 3.91 Wh (13x)	arXiv 2509.20241; test-time thinking is the right idea on the wrong substrate

The counterweights: per-query efficiency shows 8-20x line-of-sight improvement from serving optimizations, and Jevons' paradox reliably converts savings into more usage rather than less load. Which is precisely why the serious argument is not "use less AI." It is: change what a unit of intelligence physically costs, and meter it so the claim can be audited. Note also what the inference share implies — since 60-90% of the energy is spent answering rather than training, the decisive battleground is the per-answer joule, and the reasoning era is making each answer longer.

2. Why Computing Costs What It Costs: A Ten-Minute Primer

> **In plain terms:** *Physics sets an astonishingly low floor on the energy computing must use. Almost everything we pay above that floor is spent moving data, not transforming it — and a brain, which moves data as little as possible, shows what the other regime looks like.*

2.1 The floor: Landauer's limit

In 1961, Rolf Landauer proved the only step of computation that must dissipate energy is erasing information: destroying one bit costs at least kT ln 2 — at room temperature, about 2.9 zeptojoules, or 0.00000000000000000029 joules. Everything else — copying, transforming, even computing — can in principle approach zero cost if done slowly and reversibly. This is not a curiosity: it fixes the exchange rate between information and energy, and it means the ceiling on efficient computation is astronomically far above where we operate. Reversible-computing efforts (Vaire and academic programs) attack the floor directly by never erasing; every other approach in this review attacks the far larger gap above it.

2.2 The tax: where today's joules actually go

Stack today's costs against that floor and the structure of the problem becomes obvious. The table's numbers are order-of-magnitude figures from the architecture literature; exact values vary by node and design, but the ratios — which are the point — are robust.

Operation

	Approximate energy	Distance above the Landauer floor
Erase one bit (physics floor, 300 K)	~3 zJ	1x — the wall itself
Switch one modern transistor	~1 aJ (order)	~10^2-10^3
One 8-bit multiply-accumulate (the atom of AI)	~0.1-1 pJ	~10^8
Klere ternary accumulate (measured, FPGA stand-in)	0.1596 +/- 0.0006 pJ	~10^8 — an audited datapoint, not a projection
Fetch operands from on-chip SRAM	~1-10 pJ	~10^9
Fetch from off-chip DRAM	~100+ pJ per access	~10^10 — 100-1,000x the arithmetic it feeds
One frontier-model query (serving median)	0.31 Wh = ~1,100 J	~10^23
One biological synaptic event	~10-100 fJ (order)	~10^4-10^5 — general intelligence's operating point

Read the table bottom-up and the diagnosis writes itself. Arithmetic is nearly free; the memory round trip costs one hundred to one thousand times more than the mathematics it feeds; and the von Neumann architecture — separate compute and memory joined by a bus — makes that round trip its central act. Frozen weights make it eternal: the same parameters shuttle across the same bus for every token of every answer, forever. The brain's synapse, by contrast, is memory and processor in one place; it fires sparsely, in the low hertz on average, and only when something happens. Twenty watts. That is not a metaphor for this review's thesis; it is its engineering specification, and it is why every UK public program in Section 7 states the 20-watt figure in its founding documents.

2.3 The consequence: three design rules energy selects for

- **Locality —** keep memory and compute in the same physical place, so weights are used where they live. Processing-in-memory, memristor crossbars, and settling fabrics are this rule as silicon.
- **Sparsity —** compute event-driven and sparsely in time and space, so silence is free. Spiking systems and sparse-activation models are this rule.
- **Adaptive depth —** let difficulty set the budget — relax until settled rather than executing a fixed pass, so easy answers cost little and hard ones cost

> what they must. Energy-based inference is this rule, and thermal noise becomes the sampler instead of the enemy.

Every substrate and model class in this review is a bet on some combination of these three rules. The transformer-on-GPU stack violates all of them by construction — which is not a moral failing but a fitness-function artifact, as the next section explains.

3. The Hardware Lottery Meets the Energy Wall

> **In plain terms:** *Today's AI looks the way it does because of which hardware happened to be cheap in 2012 — not because it is the optimum. Change the scoring function from throughput to joules-per-answer, and the leaderboard reshuffles.*

Deep learning waited twenty years for gaming hardware to make it viable. Attention displaced recurrence largely because it parallelizes. Unstructured sparsity failed commercially because GPUs penalize irregularity. Even the leading post-transformer candidates had to pass the GPU gate before competing: recurrence re-entered the field only when reformulated as a parallel scan. The lesson is not that the transformer is wrong; it is that the transformer is the fixed point of a particular fitness function — floating-point operations at peak utilization on a von Neumann cluster — and a different fitness function selects different winners. The joule is that different fitness function, and Section 1 shows it is no longer optional: utilities, not foundries, now pace frontier expansion. A joules-per-answer metric selects exactly the three rules of Section 2.3 — locality, sparsity, adaptive depth — which are precisely the properties of energy-based and learn-at-inference models. The physics community spent the same decade proving computation's substrate need not simulate energy minimization; it can be energy minimization. Hopfield and Hinton's 2024 Nobel Prize in Physics honored the founding insight; the 2025-26 hardware generation is that insight returning as silicon, optics, and spin.

4. The Map: Substrates x Model Classes

> **In plain terms:** *Everything in this review fits on one grid: what physical thing runs the computation (rows), and what mathematical object the computation is (columns). The action concentrates where a model's native operation is a substrate's native physics.*

The companion database places 60+ organizations on this grid, with a flag for whether learning happens at inference time. Seven substrate families: thermodynamic/probabilistic devices, whose useful output is samples and whose engine is thermal noise; Ising machines and annealers, which solve constrained optimization by physical relaxation; neuromorphic spiking systems, event-driven silicon neurons; photonic processors, computing with light's propagation and interference; processing-in-memory and analog compute-in-memory, arithmetic performed inside the memory array; settling fabrics with energy accounting,

attractor networks plus joule metering; and the incumbent dense-digital GPU as baseline. Seven model classes: energy-based transformers and EBMs generally; Hopfield and associative memories; active inference and predictive coding; equilibrium propagation; fast-weight, test-time-training, and surprise-gated memories; ternary LLMs; and unified architectures (EFA). Two diagonal observations carry the strategy. First, learn-at-inference mechanisms concentrate in the rows that map onto physics: a Hebbian outer-product write is literally a memristor crossbar's native update; equilibrium propagation's settle-nudge-settle is literally what a relaxation fabric executes; sampling is literally what thermal noise provides free. Second, the GPU column is not empty — Titans, TTT, and DeltaNet are the incumbent conceding that models must learn at inference while contesting the substrate — so the race is not over whether runtime learning happens, but over which physical execution of the update wins on joules.

5. Model Classes That Learn While Inferring

5.1 Energy-Based Transformers: thinking as descent

> **In plain terms:** *Instead of computing the answer in one fixed pass, the model scores candidate answers with a learned energy and improves them by sliding downhill — so extra thinking is extra descent, and difficulty sets the budget.*

The 2025 Energy-Based Transformers work (Gladstone, Du, et al.) trains transformers to assign a scalar energy to every input-prediction pair and to predict by gradient descent on that energy to convergence. Reported: pretraining scaling up to 35% faster than the tuned Transformer++ recipe across data, parameters, FLOPs, and depth; inference-time thinking gains 29% larger than standard transformers; image denoising surpassing Diffusion Transformers with 99% fewer forward passes; largest gains out-of-distribution. The structural significance: the post-transformer argument became a scaling-law claim — and EBT inference is iterative settling, which costs a full forward pass per step on a GPU but is simply elapsed time on a substrate whose physics embodies the energy.

5.2 Active inference: AXIOM

> **In plain terms:** *Rather than training a network by a million gradient steps, the agent maintains beliefs and updates them once per observation — learning happens in the act of perceiving.*

VERSES' AXIOM (2025) abandons neural networks and backpropagation for growing, pruning Bayesian mixture models that internalize each observation in a single update, planning by minimizing expected free energy. On the company-designed, third-party-validated Gameworld 10K benchmark it outperformed DeepMind's DreamerV3 by ~60% (77 vs 48 normalized) while learning 7.6x more sample-efficiently, running 39x faster, at 400x smaller size (0.95M vs 420M parameters). Caveats are real — arcade-scale domains, object-centric priors doing

work — but the result's shape matters: online, gradient-free learning beating deep RL on capability is the drift benchmark's native tenant announcing itself.

5.3 Fast weights and test-time training: the incumbent concedes the premise

> **In plain terms:** *The frontier labs now agree models must keep learning while running — they just execute the update on GPUs. That agreement is the paradigm's strongest external validation.*

Titans (Google, 2025) writes a neural memory at inference gated by surprise; TTT layers (Stanford/Meta, 2024) make the hidden state itself a set of weights updated by gradient during inference; Gated DeltaNet gives the delta-rule write a hardware-efficient form. Modern Hopfield theory (Ramsauer et al.; Krotov) closes the intellectual loop: attention is a Hopfield update — the transformer was a covert energy-based model all along, executed on hardware that hides the fact.

5.4 Equilibrium propagation: learning as two relaxations

> **In plain terms:** *Let a physical system settle; nudge its output toward the target; let it settle again. The difference between the two settlings, read locally at each connection, is the gradient — no backpropagation machinery, so physical circuits can teach themselves.*

EqProp (Scellier and Bengio, 2017 onward) computes backprop-equivalent gradients from two relaxations using only local information — the learning algorithm settling hardware was waiting for. It has escaped simulation: UPenn's self-learning resistive networks physically learn tasks with no digital processor in the loop, and Cornell's physics-aware training extends the recipe to arbitrary physical systems.

5.5 Ternary models: the shared alphabet

> **In plain terms:** *If weights are only -1, 0, or +1, multiplication collapses to add/subtract/skip — an operation cheap for digital chips and native to analog arrays, optical schemes, and settling fabrics alike.*

Microsoft Research Asia's BitNet b1.58 demonstrated that LLMs trained natively with ternary weights retain quality at scale. This matters beyond compression: ternary is the lingua franca of compute-in-memory, settling fabrics, and several photonic schemes. The model class that alternative substrates execute natively was invented, notably, in Beijing.

5.6 The Energy First Architecture: one energy, five jobs

> **In plain terms:** *EFA's bet is unification: a single learned energy function over one sparse, readable internal state simultaneously predicts (by descending), plans (by descending toward goals), remembers (by writing fast weights during*

> *inference, no gradient), verifies (low energy = valid), and discovers scientific laws (the law is the sparse energy that fits the data).*

EFA (Charlot Lab, Institute for Physical AI @ BMI, 2026) is the most explicit unification attempt in the database, resting on a keystone identity: one sparse-positive latent that is at once the readable feature space, the associative-memory space, and the world-model prediction space. Its distinguishing feature is validation epistemology. Measured results include a descent-trained EBT whose accuracy climbs monotonically with thinking (22% to 100% across K=1 to 6); planning gains of 39% to 69% goal-reach on an unchanged value network purely from test-time search; a generate-then-check split where a flow-matching policy actuates at 100% while a contrastively trained energy verifies at 99.6%; a certified contraction region of 30.4% with 100% empirical convergence from it; and recovery of Lotka-Volterra dynamics from the real 1900-1920 Hudson Bay lynx-hare records. Equally load-bearing are the published negatives: a pre-registered claim tested and falsified; an early feedforward-is-0% claim caught and corrected; iterative energy-descent actuation identified as the known-failing Implicit Behavior Cloning recipe (0% on a 2-DOF arm) and replaced with flow matching (100%), the hybrid potential carrying a measured 65% of the control field. Everything is labeled nano-to-small; no frontier-scale capability is claimed. What the program earns is the map of exactly where an energy-native architecture wins and what each next rung costs — and its stack runs in a browser tab on Ferric, the Institute's pure-Rust cross-fabric compute layer: the substrate inversion made demonstrable.

6. The Substrates

6.1 Thermodynamic and probabilistic computing

> **In plain terms:** *These chips stop fighting thermal noise and hire it: the random jitter of electrons becomes the sampler that generative AI otherwise simulates expensively in software.*

Extropic's thermodynamic sampling units use the natural fluctuations of standard CMOS to sample directly from programmable distributions; the Z1 chip (hundreds of thousands of probabilistic circuits) is slated for 2026 early access, and the July 2026 $75M CHIPS letter of intent adds a US-fabricated Z1.5 as the project capstone — the first US industrial-policy commitment to thermodynamic AI hardware. The accompanying Denoising Thermodynamic Model reframes hardware EBMs as diffusion-style denoising on sparse Boltzmann machines, with simulations projecting on the order of 10,000x less energy per generated sample than GPU baselines — a projection, flagged as such. Normal Computing taped out the CN101 thermodynamic ASIC (2025) with a linear-algebra and SDE orientation, and published a Nature Communications prototype demonstrating matrix inversion and Gaussian sampling from noise. The 2026 sweep adds commercial and entrant depth: Signaloid (UK) ships probability-native computing today — an AWS cloud with Boeing and CERN as customers, an edge module, and a sub-10 W ASIC taped

out at TSMC in May 2026 — proving uncertainty-native compute has a market before the thermodynamic generation lands; Ludwig Computing (founded by p-bit co-originator Behtash Behin-Aein) joins the entrant field. Academically, magnetic-tunnel-junction probabilistic Ising machines reached 250-device scale with cluster updates ~10x over serial Gibbs (Nature Communications, 2026), and oscillator Ising machines were shown to operate as samplers (Communications Physics, 2026) — formerly separate paradigms collapsing into one sampling story. Quanta Magazine's July 2026 feature marks the moment the idea went mainstream.

6.2 Ising machines and annealers: relaxation as a commercial service

> **In plain terms:** *Japan's conglomerates already sell physical relaxation: pose a constrained decision as an energy landscape, and the machine falls into a good answer in microseconds.*

Toshiba's Simulated Bifurcation Machine demonstrated microsecond-latency detection and execution of optimal currency arbitrage across eight-currency combinations with >90% probability of finding the most profitable opportunity — a latency class nothing else had reached — and the company's 2026 positioning frames Ising machines as the white-box decision layer for autonomous systems (explicit argmin rather than black-box inference), deployed in AGV warehouse routing. Fujitsu's Digital Annealer, Hitachi's STATICA, NEC's vector annealing service, and NTT's 2,000-spin optical Coherent Ising Machine complete a national portfolio; D-Wave's Advantage2 continues the quantum branch. The strategic lesson: the paradigm's first capability record — time-to-good-decision — was set by conglomerates selling it as a service.

6.3 Neuromorphic systems: brain scale, and record capital

> **In plain terms:** *Silicon neurons that fire only when something happens — silence is free — now exist at monkey-brain scale on two kilowatts, and in 2025 attracted the largest seed round in semiconductor history.*

Zhejiang University and Zhejiang Lab's Darwin Monkey (“Wukong,” August 2025) is the first neuromorphic computer past two billion spiking neurons: 960 Darwin 3 chips, 100+ billion synapses — macaque scale — at roughly 2,000 watts, with an online-learning instruction set and a brain-inspired OS, running a spikified DeepSeek-derived model. Intel's Hala Point (1.15B Loihi 2 neurons) and IBM's NorthPole mark the US corporate research position; SpiNNcloud commercializes Dresden's SpiNNaker2 (a 650-million-neuron deployment at Leipzig, 2025); BrainChip's Akida and SynSense's Speck ship on-chip-learning edge silicon today; Cortical Labs (Australia) sells the biological limit case — living neurons on silicon. The July 2026 Peking University-CAS memristor chip (real-time brain-surface modeling, up to 478x an A100 on that domain-specific workload) signals China's in-memory neuromorphic ambitions. And the capital event of the field: Unconventional AI — founded by Naveen Rao (Nervana, MosaicML/Databricks)

explicitly to reach the brain's ~20 W envelope with brain-inspired analog compute — raised a $475M seed at a $4.5B valuation (a16z and Lightspeed co-leading, Lux, DCVC, Jeff Bezos participating) roughly two months after founding. It is pre-product and its efficiency claims are aspirational; as a price signal on the paradigm, it is unambiguous.

6.4 Photonic computing

> **In plain terms:** *Light computes as it travels — a lens performs a mathematical transform at literally zero marginal energy — so the bet is moving AI's linear algebra into optics and paying only for the conversions at the edges.*

Tsinghua's Taichi chiplet (Science 2024; 160 TOPS/W) was followed by Taichi-II, reported to train models optically with media-claimed efficiency near 1,000x an H100 — a ratio to treat cautiously. SJTU and Tsinghua's LightGen (Science, December 2025) integrates two million photonic neurons for generative workloads with order-100x task-specific claims, explicitly not general-purpose. The industrial-policy layer matters equally: China's first photonic pilot fabs (Wuxi six-inch series production from June 2025; a ~12,000-wafer/yr thin-film lithium-niobate line) are framed domestically as the strategic answer to export controls — the game restated as how many joules useful work costs. The 2026 Western counter-wave arrived with capital: Neurophos (Austin) raised an oversubscribed $110M Series A led by Gates Frontier with Microsoft's M12, Bosch, and Aramco, for metasurface-modulator optical inference — over a million optical processing elements per chip, with up-to-100x vendor claims; OLIX (UK) raised $220M at a >$1B valuation for an optical-memory-and-interconnect inference architecture; and GlobalFoundries signed a $300M CHIPS letter of intent for silicon photonics and co-packaged optics — the same July week as Extropic's, meaning the US CHIPS office now backs two non-incumbent compute directions simultaneously. Western capital still concentrates on interconnect (Lightmatter ~$850M raised, Celestial AI) — attacking the energy of data movement — alongside Q.ANT's photonic processors and Microsoft's Analog Iterative Machine for optical optimization.

6.5 Processing-in-memory: the pragmatic middle path

> **In plain terms:** *If fetching data costs 100x the math, do the math inside the memory chip. This needs no new model class — which is why it will pay first, and why it is a waypoint rather than the destination.*

Korea's memory duopoly is dissolving the von Neumann bottleneck from the memory side. Samsung and SK hynix — direct rivals — are jointly standardizing LPDDR6-PIM through JEDEC for a 2026 window, with claimed ~2x performance and ~70% power reduction on target operations; SK hynix's CES 2026 lineup spans PIM, the AiMX LLM-offload card, compute-using-DRAM, and custom HBM. The analog compute-in-memory startups (EnCharge, Mythic, Axelera, Rain, TetraMem; Witmem and Houmo in China) supply the edge lane. PIM will likely deliver the

paradigm's first broad commercial joule savings precisely because it demands no model-class change — and for the same reason it cheapens the incumbent's data movement without enabling learning-in-place.

6.6 Settling fabrics and the accounting layer

> **In plain terms:** *One program treats the missing instrument — a trustworthy joule meter with receipts — as the product. Whatever wins the hardware race will need exactly this to prove it.*

Klere occupies a niche no other entry does: metrology and economics. Its FPGA settling fabric (a ternary Hopfield engine on AWS F2 silicon) demonstrated 100% pattern recall by iterative settling with 68% of weight memory stuck at zero — against 84% for a one-shot read of the same defective weights — making attractor dynamics a yield argument: energy descent as free error correction on cheap, sparse, faulty hardware. Around it sits a joule-typed stack: a language whose functions carry compiler-checked energy budgets, a dataflow VM that routes each operation to the cheapest substrate or refuses, an OS with energy as the scheduling resource, and — most transferable — two-axis provenance tags (how a figure was obtained x what device it describes) so a prediction can never impersonate a measurement. Its headline figure, 0.1596 +/- 0.0006 pJ per accumulate, is measured by frequency-sweep slope and tagged as an FPGA stand-in for an unfabbed ASIC. All artifacts are CC0 and self-published; no independent replication exists yet — flagged accordingly — but the receipts discipline is exactly the instrument layer the field lacks.

7. The Software Layer: Making Hardware Comparable

> **In plain terms:** *A parallel war is being fought one level above silicon: universal software stacks that let one program run on any chip. They do not make hardware irrelevant — they make it comparable. And comparable hardware competes on joules.*

The 2025-26 evidence is decisive that this layer now has market-clearing prices and state strategies. In the West: Qualcomm completed a roughly $3.9B all-stock acquisition of Modular (July 2026) — the Mojo language and MAX serving stack that retarget one codebase across NVIDIA, AMD, and Apple GPUs plus CPUs, built by Chris Lattner, whose lineage (LLVM, Clang, Swift, MLIR) makes him the purest carrier of compiler-layer insight in the industry; Modular's own framing of the shift — from selling GPU-hours to selling tokens — is the joule economy stated in commercial terms. OpenAI's Triton has become PyTorch's de facto portable kernel layer; Google's MLIR/StableHLO/IREE form the compiler commons; and community consensus around ZLUDA's sixth release is that matured Vulkan and native paths already erode the CUDA moat for inference. In the East, portability is industrial policy: Huawei open-sourced CANN (August 2025) as an explicit CUDA counter, with torch_npu removing the PyTorch switching cost; DeepSeek now ships frontier models optimized for Ascend, Cambricon, and Hygon on day one; Alibaba

open-sourced its own stack (July 2026); and reporting describes movement toward a national unified programming model so one codebase runs on any domestic accelerator. One caution travels with the enthusiasm: NVIDIA's 2024 acquisition of OctoAI — the commercial home of the pioneering TVM compiler — shows incumbents buy portability layers as containment, which is why the specifications this review contributes are CC0 and consortium-governed by design.

The strategic reading requires one distinction the coverage misses: these are three different games wearing one name. Within-paradigm portability commoditizes the GPU abstraction itself — kernels, tensors, clocks. Sovereign portability runs the same abstraction under state alignment. Neither can address the substrates of Sections 5-6: no Triton kernel, Mojo function, or CANN operator can be lowered onto a device whose native act is relax-until-stable. The one shipping exception is the template: NIR, the Neuromorphic Intermediate Representation, runs one spiking model across Loihi 2, SpiNNaker2, Xylo, and Speck precisely because its primitives match the physics. The relaxation analogue — primitives of settle, sample, anneal, local write, verify, certify, with typed energy budgets and native receipts — did not exist when this review began; it is now specified as EFA-RFC-002 (RELAX/1), designed to ride the MLIR commons and bridge NIR, with a mandatory GPU-simulation lowering so adoption requires no exotic hardware. The convergence closes the argument of Sections 9-10: once software makes substrates interchangeable, procurement optimizes cost- and joules-per-answer, the energy receipt becomes the purchasing document, and the portability layers themselves become the natural mass issuers of OER receipts. China, whose substrate portfolio spans photonic, memristor, and neuromorphic devices no kernel dialect can address, is the actor that needs the cross-paradigm layer most — a reason to publish the open version first.

Game	What it commoditizes	Who is playing
Within-paradigm portability	The GPU abstraction (kernels, tensors, clocks)	Mojo/MAX (Qualcomm, $3.9B), Triton, MLIR/ StableHLO/IREE, SYCL/ UXL, Vulkan compute, ggml/llama.cpp, ZLUDA, tinygrad
Sovereign portability	The vendor, within the same abstraction	Huawei CANN + torch_npu, Alibaba's open stack, Moore Threads MUSA, a reported national unified programming model
Cross-paradigm portability	The paradigm itself (settle / sample / write primitives)	NIR (shipping, spiking); RELAX/1 (specified, companion RFC); Klere

			flowg and Ferric as nearest existing artifacts

8. Regions and Capital

> **In plain terms:** *Five regions are running five different strategies that happen to interlock: American startups and first policy money, Chinese breadth under export pressure, Japanese services, Korean memory, and European institutions.*

The United States leads in thermodynamic startups, energy-based theory, and — via two CHIPS letters of intent in a single July week (Extropic $75M, GlobalFoundries $300M) — has made its first industrial-policy commitments to the paradigm, atop the DARPA/NSF/ONR pipeline and now the field's record private round (Unconventional AI). China runs the most vectors simultaneously: neuromorphic scale (Darwin Monkey), photonics plus pilot fabs, memristor in-memory systems, and the ternary model class itself (BitNet, from Microsoft Research Asia in Beijing) — with export controls as the forcing function and the March 2026 Fifteenth Five-Year Plan naming brain-inspired technology strategic. Japan monetizes relaxation today through corporate Ising services and owns the photonics-electronics convergence roadmap (NTT's IOWN). Korea industrializes the memory-side attack through JEDEC standardization. Europe and the UK hold the institutional base: the ~EUR 607M Human Brain Project's EBRAINS legacy, SpiNNcloud, Innatera, Q.ANT, Axelera — now joined by a deliberate UK public stack: the UCL-led Neuroware Innovation and Knowledge Centre (GBP 12.8M, nine institutions, commercialization mandate), the EPSRC national neuromorphic centre (GBP 4.48M, its founding documents citing the 20-watt brain), and ARIA's ~GBP 42M scaling-compute push (flagged for verification) targeting order-1,000x reductions — the West's most explicitly energy-first public programs, plus new UK entrants Signaloid and OLIX.

Capital, tiered. Direct paradigm investment — startup rounds plus targeted public programs — now totals on the order of $2.5-4B cumulative disclosed, of which roughly $1.3B arrived in the eight months to August 2026 (Unconventional AI $475M, OLIX $220M, Neurophos $110M, plus $375M of CHIPS letters of intent — non-binding, flagged as such). Enabling capital the paradigm rides is far larger (China's ~$47.5B Big Fund III, memory-maker capex, EU Chips JU). Adjacent demand-side programs (Korea's $390M AI champions) sit apart. Against the incumbent's ~$725B single-year capex the direct figure is rounding error — which is the asymmetry, not the refutation: the paradigm's entry points are the tiers where the incumbent's capex buys the least.

Region	Signature efforts	Public capital signal	Strategic logic
USA	Extropic Z1/Z1.5; Unconventional AI;	$75M + $300M CHIPS LOIs (Jul	Startup-led; first industrial-policy

	Neurophos; Normal; Signaloid-adjacent labs; EBT theory; EFA/Ferric/Klere stack	2026); $475M record seed; federal grants	entry; record private pricing of the paradigm
China	Darwin Monkey; Taichi/LightGen + pilot fabs; PKU-CAS memristor; BitNet	15th FYP designation; Big Fund III enabling (~$47.5B); state labs	Export controls force substrate diversification
Japan	Toshiba SBM; Fujitsu DA; Hitachi; NEC; NTT CIM + IOWN; Sakana	Corporate R&D; NEDO; IOWN consortium	Relaxation as commercial service, today
Korea	LPDDR6-PIM JEDEC standard (Samsung + SK hynix); AiMX; CuD	Industry capex; national incentives	Dissolve the bottleneck from the memory side
EU / UK	EBRAINS legacy; SpiNNcloud; Innatera; Q.ANT; Neuroware IKC; EPSRC centre; ARIA; OLIX; Signaloid; Cortical Labs (AU)	EUR 607M HBP (2013-23); GBP 12.8M + 4.48M centres; ~GBP 42M ARIA [verify]	Institutional base; explicit 20 W / 1,000x public framing

9. The Benchmark Question: What Would Count as the AlexNet Moment

> **In plain terms:** *New paradigms win by naming a scoreboard the incumbent cannot top. One such record is already held (microsecond decisions); the open one is adapting on the fly — scored in regret, latency, and audited joules.*

One record stands: time-to-good-decision on constrained problems, held by physics-based settling since Toshiba's microsecond arbitrage — a capability metric, not an efficiency one. The open prize is learning under drift. Proposed scoreboard: regret under distribution shift x decision latency x receipted joules, on embodied tasks whose dynamics change after deployment. The incumbent cannot top it without abandoning frozen-weight deployment: adaptation is a data-center event, and autoregressive inference cannot reach the latency floor. The native tenants are Section 5's mechanisms — Hebbian writes, dendritic gating, single-pass Bayesian updates, equilibrium learning — and the EFA nano suite already demonstrates the

components (memory-load-bearing navigation, continual learning across three physics worlds without forgetting, thinking curves that fall as the landscape recodes). What remains is composition at benchmark scale, with joule receipts. A second requirement rides along: certification. An agent that rewrites itself at runtime needs an intrinsic safety object, and the energy paradigm has one — the descended energy as a control-Lyapunov certificate, upgraded in EFA's 2026 revision to verified contraction regions. Receipts for joules and certificates for stability are two faces of one accounting layer; this is where Hinton's mortal-computation economics turns practical, since a computer whose weights cannot be copied can still be certified by its behavior and its bills.

10. The Economic Forecast: Capital and Energy Savings

> **In plain terms:** *Price the transition with every assumption on the table and deliberately below the vendor claims. The result: modest to 2030, then compounding — about 91 TWh and $120B of unbuilt data centers per year by 2035 in the base case.*

The companion workbook implements the model; the interactive explainer runs it live with sliders; the figures below are the formulas' output. Method: AI data-center electricity starts at 190 TWh in 2026 (anchored to the IEA trajectory at a 35-45% AI share of total data-center demand) and grows 22% annually to 2030, then 12% to 2035. A scenario-dependent share is amenable to relaxation-native compute — sampling, optimization, verification, edge-adaptive work (10% / 20% / 30% across Conservative / Base / Aggressive). Efficiency gains on amenable workloads are 10x / 100x / 1,000x — set below the most-cited vendor claims; the 10,000x DTM projection is excluded from headlines. Adoption follows an S-curve reaching 62% of the amenable share by 2035 in the base case, capped at 80%. Conversions: $0.08/kWh; $10B per GW of all-in AI data-center build; 0.35 kg CO2/kWh.

Headline (formulas in workbook)	**Conservative**	**Base**	**Aggressive**
Energy saved in 2030 (TWh/yr)	1.9	8.3	18.9
Energy saved in 2035 (TWh/yr)	20.7	91.1	177.9
Energy cost saved in 2035 ($B/yr)	$1.7	$7.3	$14.2
Cumulative energy saved 2026-2035 (TWh)	60.3	265.3	573.4
Cumulative energy cost saved 2026-2035 ($B)	$4.8	$21.2	$45.9
	$27.8	$122.3	$238.9

Capex avoided by 2035 ($B)			
CO2 avoided in 2035 (Mt/yr)	7.2	31.9	62.2

Interpretation. Through 2030 savings are modest — single-digit TWh in the base case — because adoption is early and the amenable share bounded. The economics arrive 2031-2035, when a maturing adoption curve compounds against an AI load approaching 750 TWh per year: base-case savings of ~91 TWh/yr by 2035 equal a mid-sized European country's electricity, and the ~$120B of avoided build is the figure for capital allocators, accruing as campuses not constructed. What would move these numbers most: not the efficiency multiple but the amenable-share and adoption parameters — a research-agenda statement, since every benchmark win that migrates a workload class into the amenable column outweighs another order of magnitude on one already there. Three caveats bound the exercise: Jevons effects may convert savings into demand (the capex-avoided line is partially robust, measuring capability per facility); grid decarbonization shrinks the CO2 line independently; and the model assumes the compiler-and-receipts layer gets built — without it, amenable workloads cannot migrate at all.

11. Failure Modes: Why Previous Waves Died, and What Differs Now

> **In plain terms:** *Alternative computing has a graveyard, and every stone names a current risk. Four things are genuinely different this time — but only if the field polices its own claims.*

Analog computing lost to digital in the 1960s-70s on precision, composability, and programmability. Lisp machines lost to commodity scaling. The 2010s memristor wave over-promised device uniformity and under-delivered toolchains. Neuromorphic silicon existed fifteen years without a paradigm shift because spiking networks never received a competitive learning algorithm — substrate alone does not summon a model class. Each failure names a live risk: analog variability and calibration drift; the absent lowering compiler from learned energies to device physics; benchmark contestation (vendor-designed tasks, simulation-based projections, task-specific ratios quoted as general — several appear in this very review, flagged); and the incumbent's compounding advantage in capital and software, now $725B per year.

Four things differ. The energy constraint now binds the incumbent — the first time the alternative's fitness function is also the industry's bottleneck. The algorithms arrived: EqProp, test-time training, active inference, and EBT scaling give substrates a model class worth running, where the neuromorphic wave had none. The alphabet converged: ternary works at LLM scale, and ternary is what crossbars, fabrics, and several photonic schemes natively speak. And the accounting layer is being built: provenance-tagged joule receipts make claims checkable — the precondition for surviving the hype cycle now visibly underway (a record seed

round for a two-month-old company is a price signal and a warning in one). The entry path is correspondingly asymmetric: not frontier chat, where the capex compounds, but the tiers where GPUs are worst — microsecond decisions, always-on edge perception, and above all continual adaptation.

12. Recommendations

- **For the EFA program —** Stand up the drift benchmark: publish the regret x latency x receipted-joules scoreboard on embodied tasks with post-deployment dynamics shifts, a frozen-weights strong baseline (TD-MPC2 class), and mandatory energy receipts. One artifact gives the whole field its shared target.
- **For the EFA program —** Run the flagged annealing experiment on the coupled-chain ceiling: the whitepaper's own open confound is also the precise seam where thermodynamic substrates enter. A negative is publishable; a positive is a bridge.
- **For the EFA program —** Lower one mechanism to physical hardware: the Hebbian outer-product write is the native operation of a memristor crossbar and of a settling fabric's coupling update. EFA's memory on either, with receipts, is the first physical EFA organ.
- **For the EFA program —** Integrate a meter: joules-per-task is the declared currency, but WebGPU cannot read power. Klere-style frequency-sweep receipts or wall-plug MLPerf-Power methodology for Ferric turns the currency real.
- **For funders —** Fund the missing middle, not another substrate: the compiler that lowers learned energies onto heterogeneous physics, and the metrology standard for energy receipts, apply to every substrate at once. Both now exist as CC0 draft specifications (EFA-RFC-001: OER/1 + DRIFT/1; EFA-RFC-002: RELAX/1, with a reference-runtime roadmap for Ferric); funding their reference implementations costs a fraction of one 2026 chip round.
- **For funders and reviewers —** Treat vendor efficiency ratios as unpriced options until independently replicated on standard workloads with wall-plug measurement; require provenance tags (measured vs modeled, target device vs stand-in) as a condition of citation.
- **For standards bodies —** Begin energy-receipt standardization: the JEDEC LPDDR6-PIM precedent proves rivals standardize when economics demand it. A common receipt format — operation, substrate, joules, provenance — is smaller than a memory standard and more catalytic.
- **For policymakers —** Count avoided gigawatts as policy output: a workload migrated to relaxation-native compute is grid capacity that never needs

permitting. The CHIPS letters of intent priced this logic; the 45 GW shortfall gives it urgency.

13. Honest Limits of This Review

This is a snapshot compiled August 2, 2026 from public sources; the field's velocity guarantees omissions, and coverage skews toward English-language and Chinese-state-media-covered efforts. Funding figures marked [verify] in the companion database were not independently confirmed. Several load-bearing performance claims are vendor-reported, simulation-based, or task-specific, and are flagged rather than laundered into headlines. Two entries central to this document's framing — Klere and the EFA program — are self-published primary sources without independent replication; their length here reflects unusual reproducibility discipline (open artifacts, validation ledgers, published negatives), not third-party confirmation, and the review's association with the EFA program is disclosed on the title page. The physics primer uses order-of-magnitude figures whose ratios, not decimals, carry the argument. The forecast is a transparent model, not a prediction: its purpose is to make the arithmetic inspectable, and every parameter is exposed — in the workbook's cells and the explainer's sliders — for the reader to move.

Glossary

Term	Meaning in this review
Joule (J) / watt (W)	The unit of energy / energy per second. One watt-hour = 3,600 J. A TWh is a billion kWh; a GW is a billion watts of capacity.
pJ / fJ / aJ / zJ	Pico (10^-12), femto (10^-15), atto (10^-18), zepto (10^-21) joules — the scales of single operations.
Landauer limit	The minimum energy to erase one bit: kT ln 2, ~3 zJ at room temperature. The physics floor of irreversible computing.
Von Neumann bottleneck	The energy and time cost of shuttling data between separate compute and memory — the dominant tax in today's AI.
Energy-based model (EBM)	A model that scores configurations with a learned energy and answers by finding low-energy states — prediction as descent.
Energy descent / settling / relaxation	Sliding downhill on an energy landscape until stable. On the right hardware, this is literal physics, and its cost is elapsed time.
Test-time thinking	Spending variable computation at answer time (more descent steps, more samples) so difficulty sets the budget.

Learn at inference	Updating a model's memory or weights while it runs, on-device, without a training pipeline — fast weights, TTT, Titans, Hebbian writes.
Hebbian write / fast weights	A local memory update from co-activity (an outer product), needing no gradient — and natively the write operation of a crossbar.
Active inference	Agents that maintain beliefs and act to minimize expected surprise, updating per observation in a single pass (VERSES/Friston).
Equilibrium propagation (EqProp)	Learning from two physical relaxations (free and nudged); gradients read locally — how physical circuits teach themselves.
Ising machine / annealer	Hardware that encodes a problem as an energy landscape of coupled spins and relaxes to low-energy answers.
p-bit / TSU	Probabilistic bit — a device that fluctuates by design; thermodynamic sampling units are chips of them, using noise as the sampler.
Neuromorphic / SNN	Brain-inspired event-driven hardware; spiking neural networks fire only on events, so silence costs nothing.
PIM / CIM / crossbar / memristor	Processing-in-memory and compute-in-memory: arithmetic inside the memory array; a memristor crossbar does matrix math in analog physics.
Ternary (BitNet)	Weights restricted to -1/0/+1 — multiplication becomes add/subtract/skip; the shared alphabet of efficient substrates.
Photonic computing	Computing with light's propagation and interference; linear transforms at near-zero marginal energy, paid at the electro-optical edges.
Jevons paradox	Efficiency gains lowering cost so much that total consumption rises. Why savings must be argued in capability-per-joule, not demand reduction.
Provenance / receipts	Tags on every figure: measured vs modeled, target device vs stand-in — plus audited joules per task. The anti-hype instrument.
Mortal computation	Hinton's term for compute whose learned weights are inseparable from its physical substrate — certified by behavior and bills, not copies.

Regret / drift	Regret: performance lost versus the best policy in hindsight. Drift: the world changing after deployment. Together: the open benchmark.
Portability layer / IR	Software letting one program run on many chips (Mojo/MAX, Triton, MLIR, SYCL, CANN). Today's layers assume kernels-on-clocks; NIR is the spiking exception; RELAX/1 (companion RFC-002) is the proposed relaxation analogue.
OER/1, DRIFT/1, RELAX/1	This program's three companion specifications: the Open Energy Receipt schema, the learning-under-drift benchmark, and the cross-paradigm relaxation dialect (EFA-RFC-001/-002, CC0).
Control-Lyapunov / contraction certificate	Mathematical proof that a controller converges — the energy paradigm's built-in safety object for self-modifying agents.

Appendix: Timeline, 2019-2026

Year	What moved
2019	Tianjic hybrid neuromorphic chip on the cover of Nature; Loihi and SpiNNaker mature quietly.
2020-22	Toshiba SBM demonstrates microsecond arbitrage; EqProp escapes simulation (UPenn circuits physically learn); Hinton names mortal computation.
2024	Hopfield & Hinton win the Physics Nobel. Taichi reaches 160 TOPS/W. Hala Point reaches 1.15B neurons. BitNet proves ternary at LLM scale. TTT layers arrive.
2025	Darwin Monkey passes 2B spiking neurons at ~2 kW. Normal tapes out CN101. Extropic ships XTR-0 + the DTM paper. EBT posts scaling wins; AXIOM beats DreamerV3; Titans writes memory at inference. UK launches Neuroware (GBP 12.8M). Unconventional AI raises a $475M seed at $4.5B.
2026	Industrial policy arrives: Extropic $75M and GlobalFoundries $300M CHIPS LOIs in one week. JEDEC LPDDR6-PIM window. Z1 early access; Signaloid's sub-10 W ASIC at TSMC ships to a market including Boeing and CERN. Neurophos ($110M) and OLIX ($220M) fund photonic inference. PKU-CAS memristor chip in Science. OIMs shown to be samplers. Quanta takes thermodynamic computing mainstream. The software layer

	prices the thesis: Qualcomm acquires Modular for ~$3.9B; Alibaba open-sources its stack against CUDA; DeepSeek ships day-one domestic-chip support. EFA publishes its measured whitepaper, validation ledger, and the OER/DRIFT/RELAX specifications.
Open	The AlexNet slot: learning under drift — on-device adaptation scored by regret x latency x receipted joules.

References (Selected)

Landauer, R. Irreversibility and Heat Generation in the Computing Process. IBM J. R&D (1961).

Hooker, S. The Hardware Lottery. CACM (2021). arXiv:2009.06489.

IEA. Energy and AI (Apr 2025); Key Questions on Energy and AI (Apr 2026).

Gartner. Data Center Electricity Consumption to Grow 26% in 2026 (press release, Jun 10, 2026).

Goldman Sachs Research. US Data Center Power Demand Projected to Double by 2027 (May 2026).

FT-compiled hyperscaler capex guidance, via Axis Intelligence (Jun 2026). Brookings, Global Energy Demands within the AI Regulatory Landscape (2026).

Energy Use of AI Inference, Efficiency Pathways, and Test-Time Scaling. arXiv:2509.20241.

Gladstone, A., Du, Y., et al. Energy-Based Transformers are Scalable Learners and Thinkers. arXiv:2507.02092 (2025).

VERSES AI. AXIOM + Gameworld 10K (validated by Soothsayer Analytics), 2025.

Behrouz, A., et al. Titans (arXiv:2501.00663). Sun, Y., et al. TTT (2024). Yang, S., et al. Gated DeltaNet (arXiv:2412.06464).

Ramsauer, H., et al. Hopfield Networks is All You Need (2021). Krotov, D. Energy Transformer (2023).

Scellier, B., Bengio, Y. Equilibrium Propagation (2017-). Dillavou, S., et al. (UPenn) physical learning networks. Wright, L., McMahon, P., et al. Deep physical neural networks (2022).

Microsoft Research. BitNet b1.58 (2024-25). Kosowski, A., et al. The Dragon Hatchling. arXiv:2509.26507.

Extropic. Thermodynamic Computing: From Zero to One (Oct 2025); $75M Commerce LOI (Jul 2026); DTM, npj Unconventional Computing (2026). Quanta Magazine, Thermodynamic Computers Go With the (Energy) Flow (Jul 2026).

Normal Computing. CN101; Nature Communications thermodynamic prototype. EE Times, Probabilistic Computing Is Already Here (Jul 2026) — Signaloid UxHw (Boeing, CERN; TSMC ASIC May 2026); Ludwig Computing.

DCD / Bloomberg. Unconventional AI raises $475M seed (a16z, Lightspeed, Bezos; $4.5B).

PRNewswire / Photonics Spectra. Neurophos $110M Series A (Gates Frontier, M12; Jan 2026). SiliconANGLE / FT. OLIX $220M (Feb 2026). GlobalFoundries $300M CHIPS LOI (Jul 2026).

Nature Communications (2026): 250-MTJ probabilistic Ising machine. Communications Physics (Feb 2026): oscillator Ising machines as samplers.

Global Times / SCMP / People's Daily. Darwin Monkey (Aug 2025). Science: Taichi (2024); LightGen (Dec 2025); PKU-CAS memristor chip (Jul 2026). Photonic pilot fab coverage (2025-26).

Toshiba SBM materials (IEEE Spectrum 2020; IISM May 2026). Fujitsu, Hitachi, NEC, NTT, D-Wave public materials.

Samsung / SK hynix LPDDR6-PIM JEDEC coverage (2024-26); SK hynix CES 2026.

UCL / UKRI. Neuroware IKC (GBP 12.8M, Oct 2025); UK Multidisciplinary Centre for Neuromorphic Computing (GBP 4.48M).

Klere. EPU, settling fabric, joule-typed stack. klere.ai (self-published, CC0, 2026).

Charlot, D. J. Energy First Architecture (EFA) whitepaper v1 + validation ledger. Charlot Lab, IPAI @ BMI (2026). efa.physicalai-bmi.org; github.com/dcharlot-physicalai-bmi/efa.

Qualcomm-Modular acquisition (~$3.9B, Jun-Jul 2026): EE Times / NAND Research / company statements. Modular platform materials (Mojo 1.0 beta; MAX).

Garcia-Herrero, A., Martens, B. Stack battles: the US-China AI rivalry beyond chips. Bruegel (Jun 2026) — CANN open-sourcing, torch_npu, adoption strategy.

DeepSeek day-one domestic-chip support: Tom's Hardware (Sep 2025; Jan 2026). Alibaba open-source stack: ChinaTechNews (Jul 2026). ZLUDA v6 updates: Phoronix / project blog (2026).

Pedersen, J., et al. Neuromorphic Intermediate Representation (NIR). Nature Communications (2024).

Hinton, G. The Forward-Forward Algorithm; mortal computation (2022-23). Hopfield, J., Hinton, G. Nobel Prize in Physics (2024).