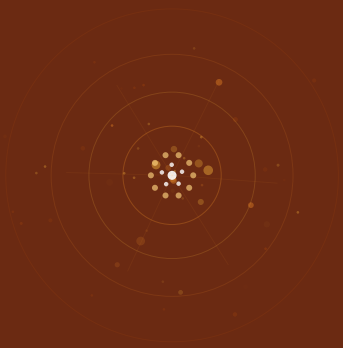


PHYSICS OF INTELLIGENCE

THE ENTROPY TAX

Why Intelligence Costs Energy, and Whether That Changes Everything

NABIL EL MAHYAOUI

The Entropy Tax

The Entropy Tax

Why Intelligence Costs Energy, and Whether That Changes Everything

Nabil El Mahyaoui

An essay on the thermodynamics of computation and intelligence

Copyright © 2026 Nabil El Mahyaoui.
All rights reserved.

Contents

1 The Cost of Knowing **1**

1.1 Landauer's Whisper 3

1.2 The Irreversibility Trap 6

1.3 Why It Matters Now 9

2 The Physics of Erasure **13**

2.1 Maxwell's Demon Gets an Invoice 14

2.2 Reversible Gates and the Bennett Escape. 17

2.3 The Quantum Card. 19

2.4 The Thermodynamic Ledger 21

3 The Burning of Intelligence 25

3.1 The Training Furnace 26

3.2 Inference at Scale 28

3.3 The Efficiency Counterargument 30

3.4 The Biological Benchmark 32

3.5 Is Intelligence Necessarily Expensive? . 34

4 The Architecture of Cool 38

4.1 Neuromorphic Dissent 39

4.2 Thermodynamic Computing 42

4.3 The Reversibility Renaissance 45

4.4 The Quantum Horizon 48

4.5 Convergence 50

5 Epilogue: A Memo to the Engineers 54

5.1 If You Are Building AI 55

5.2 If You Are Funding AI 58

5.3 If You Are Thinking About AI 60

5.4 The Watt and the Thought. 62

References **66**

1

The Cost of Knowing

Before there were transistors, there were women with pencils. In the early 1940s, the Ballistic Research Laboratory at the Aberdeen Proving Ground employed over a hundred human "computers"—mostly women with mathematics degrees—whose job was to calculate artillery firing tables by hand. Each table required roughly forty hours of sustained arithmetic: multiplication, interpolation, the numerical integration of differen-

tial equations governing projectile trajectories through variable atmosphere. The work was gruelling, exact, and fuelled, in the most literal sense, by the metabolic oxidation of glucose. A human computer performing sustained mental labour burns approximately 1.5 calories per minute above the basal rate. Forty hours of computation therefore cost roughly 3,600 dietary kilocalories—about two days' worth of food.

The point is not the quaintness of the arithmetic. The point is that computation has *always* had a thermodynamic cost. Every bit ever processed, every logical operation ever performed, has required energy to be drawn from somewhere and dissipated as heat into the environment. The notion that information is "weightless" or "virtual"—a comfortable metaphor of the digital age—is, and has

always been, a polite fiction. Even the abacus requires a hand to move the beads, and that hand requires a body, and that body requires fuel.

At Los Alamos, Richard Feynman organised teams of human computers into parallel processing chains, passing intermediate results on index cards from desk to desk. It was, in retrospect, a biological multicore architecture—one that consumed sandwiches and coffee instead of electricity, but obeyed the same fundamental constraints. You cannot compute without consuming. You cannot know without paying.

1.1 Landauer's Whisper

In 1961, a quiet, meticulous physicist named Rolf Landauer published a paper at IBM's

Thomas J. Watson Research Center that would take decades to be fully appreciated. Its title was unassuming: "Irreversibility and Heat Generation in the Computing Process." Its conclusion was devastating.

Landauer demonstrated that every *irreversible* logical operation—every time a bit of information is erased, every time a computational state is destroyed rather than preserved—must dissipate a minimum amount of energy as heat. That minimum is $kT \ln 2$: Boltzmann's constant multiplied by the absolute temperature, multiplied by the natural logarithm of two. At room temperature, this works out to approximately 2.85×10^{-21} joules.

The number is almost comically small. It is roughly a trillionth of a trillionth of the energy required to lift a grain of sand by one millimetre. And yet it is *not* zero.

This is the crucial distinction. Landauer's limit is not an engineering constraint. It is not a limitation of our materials, our manufacturing processes, or our cleverness. It is a consequence of the second law of thermodynamics—the same law that dictates that ice melts in warm water, that perfume diffuses through a room, that the universe trends inexorably towards disorder. To erase a bit is to reduce the entropy of a computational system, and the second law demands that this reduction be compensated by an equal or greater increase in entropy elsewhere. That compensation takes the form of heat.

"Information is physical."
—Rolf Landauer

Three words, and they rearranged the relationship between physics and computation perma-

nently. If information is physical, then processing it is a thermodynamic act. If processing it is a thermodynamic act, then every computation has an entropy tax—a minimum tithe paid to the universe for the privilege of knowing something that was not known before.

1.2 The Irreversibility Trap

Landauer's limit is beautiful in theory. In practice, it is a cruel joke.

A modern logic transistor in a state-of-the-art processor—say, a 3-nanometre node chip fabricated by TSMC in 2025—dissipates roughly 10^{-17} joules per switching operation. That is approximately 10,000 times the Landauer minimum. Older architectures, larger nodes, and less optimised designs routinely waste 100,000 to 1,000,000 times the theoretical floor.

The gap between where we are and where physics says we *could* be spans six orders of magnitude. It is, to put it in industrial terms, as if the internal combustion engine were operating at 0.01% thermal efficiency instead of the 30–40% we have actually achieved. We are, thermodynamically speaking, still rubbing sticks together.

Why? The reasons are layered and instructive.

First, there is the tyranny of speed. Switching a transistor faster requires driving it harder, which means higher voltages and steeper current gradients, both of which dissipate energy quadratically. The Landauer limit assumes quasi-static operation—infinitely slow, infinitely careful manipulation of bits. Real chips operate at billions of cycles per second, and haste has a thermodynamic premium.

Second, there is the legacy of architecture. Modern processors descend from designs conceived in the 1970s and 1980s, when energy was cheap and transistors were expensive. The design philosophy was to minimise transistor count, not energy dissipation. That heritage persists in every instruction set, every cache hierarchy, every branch predictor. We have been optimising for the wrong variable for half a century.

Third, and most fundamentally, there is the irreversibility trap itself. Conventional Boolean logic gates—AND, OR, NOT—are inherently irreversible: information flows in and is destroyed. A two-input AND gate produces one output; the input states cannot be recovered from the output alone. Every such gate erases information, and every erasure incurs the Landauer tax. Reversible computing architectures

exist in theory, but building them at scale remains one of the great unsolved problems of computer engineering.

1.3 Why It Matters Now

For fifty years, none of this mattered.

Or rather: it mattered in the same way that the heat death of the universe matters—as an abstract concern, safely deferred to a future so distant it could be ignored by anyone with quarterly earnings to report. Moore's Law, that astonishing empirical regularity observed by Gordon Moore in 1965, ensured that the number of transistors on a chip doubled approximately every two years. More transistors per chip meant more computation per watt, which meant the entropy tax—while real—was shrinking as a fraction of useful work. The

thermodynamic bill was growing, but the computational output was growing faster.

That free ride is over.

Dennard scaling—the companion law that allowed voltage and current to shrink in proportion to transistor size, holding power density constant—collapsed in the mid-2000s. Since then, making transistors smaller no longer makes them proportionally more efficient. The industry responded with multi-core architectures, with specialised accelerators, with increasingly baroque tricks to squeeze more useful work from each watt. But the fundamental trend is clear: the exponential improvement in energy efficiency per operation that sustained computing for five decades has decelerated to a crawl.

Meanwhile, demand is exploding.

Training GPT-4 in 2023 consumed an estimated 50 gigawatt-hours of electricity—roughly the annual consumption of 4,000 American homes. Each successive generation of frontier AI models requires approximately four times the compute of its predecessor. If that scaling continues, and every major laboratory insists it will, a model trained in 2028 will demand on the order of 3,000 gigawatt-hours: the annual output of a small power station, devoted entirely to teaching a neural network to predict the next token.

Two curves are converging. The curve of energy efficiency improvement is flattening. The curve of computational demand is steepening. Their intersection defines a wall—not a metaphorical wall, not a strategic inconvenience, but a *thermodynamic* wall, as real and as implacable as the speed of light.

The question confronting the computing industry is no longer whether the entropy tax will come due. It is whether we can restructure the architecture of intelligence itself before the bill arrives.

2

The Physics of Erasure

In 1867, James Clerk Maxwell—the same man who unified electricity and magnetism into a single elegant theory—proposed a thought experiment so subversive that it haunted physics for over a century. Imagine a tiny, intelligent being stationed at a small door between two chambers of gas at equal temperature. The being observes individual molecules approaching the door and selectively opens it: fast molecules are allowed

through in one direction, slow molecules in the other. Over time, one chamber grows hot and the other cold—without any work being performed. The second law of thermodynamics, that iron commandment forbidding the spontaneous decrease of entropy, appears to be violated by nothing more than careful observation.

Maxwell called his creation a "finite being." Lord Kelvin, with characteristic flair, called it a "demon." The name stuck, and so did the problem.

2.1 Maxwell's Demon Gets an Invoice

For sixty years, physicists argued about the demon without resolving it. The breakthrough came in 1929, when Leo Szilard—a Hungarian physicist with an uncanny talent for being

first—published a deceptively simple analysis. Szilard reduced Maxwell's demon to its minimal case: a single molecule in a box, a partition, and a being that observes which side the molecule occupies. He showed that the act of acquiring one bit of information about the molecule's position must, by thermodynamic necessity, generate at least $kT \ln 2$ of entropy somewhere in the universe.

The demon, Szilard argued, does not get to observe for free. Information has a thermodynamic cost. This was a radical claim in 1929—decades before Shannon formalized information theory, decades before anyone built a digital computer. Szilard was asserting that information is not abstract, not merely mathematical, but *physical*. A bit is a real thing, as real as a joule or a kilogram, and it obeys thermodynamic law.

The argument was suggestive but incomplete. It took another half-century for Charles Bennett, working at IBM Research in 1982, to deliver the definitive resolution. Bennett's insight was precise and devastating: the demon's thermodynamic cost does not arise from *acquiring* information, as Szilard had supposed, but from *erasing* it. The demon must keep records of its observations—which molecule went where, which side of the partition was occupied. As its memory fills, the demon must eventually erase old records to make room for new ones. And erasure, as Rolf Landauer had proved in 1961, is the only logically irreversible operation in computation—the one act that *must* dissipate energy.

The demon, it turns out, does not violate the second law. It merely defers the entropy in-

crease until the moment it cleans its notebook. The ledger always balances.

2.2 Reversible Gates and the Bennett Escape

Bennett's resolution of the demon paradox had a remarkable corollary. If the thermodynamic cost of computation arises specifically from erasure, then computation performed *without* erasure should, in principle, cost nothing. Bennett proved exactly this.

The technique is elegant in theory and punishing in practice. First, perform the desired computation, but save every intermediate result—erase nothing. Second, copy the final output to a separate register. Third, run the entire computation *backward*, step by step, restoring every intermediate value to its original state. The

intermediates are "uncomputed" rather than erased, and the only new information that persists is the copied output.

This procedure—sometimes called the Bennett trick—demonstrates that any computation can be made thermodynamically reversible. The Landauer limit is not a speed limit on thought itself. It is a tax levied specifically on forgetting.

The catch, of course, is that reversible computation is grotesquely expensive in other resources. Running a computation forward, saving all intermediates, and then running it backward requires memory proportional to the number of steps—an overhead that renders the approach impractical for any real-world problem of interesting size. You can escape the entropy tax, but only by paying an equally ruinous bill in memory and time. Thermody-

namics, like any competent tax authority, does not permit loopholes without compensating assessments.

2.3 The Quantum Card

Quantum mechanics offers what appears, at first glance, to be a more promising escape route. Quantum gates are unitary transformations—mathematical operations that are, by definition, reversible. Every quantum operation has an inverse. No information is destroyed during the evolution of a quantum state; the wave function evolves deterministically under the Schrödinger equation, preserving all information until the moment of measurement.

This makes quantum computing thermodynamically seductive. In principle, a quan-

tum circuit could perform a calculation that would require billions of irreversible classical operations, yet dissipate energy only at the final measurement step. The computation itself—the vast unitary evolution through Hilbert space—would be thermodynamically free.

The catch, again, is measurement. The moment you extract a classical result from a quantum computation, you collapse the wave function, irreversibly projecting the system into a definite state. Information is destroyed. Landauer's limit reasserts itself.

And then there is error correction. Current quantum hardware is noisy—qubits decohere, gates introduce errors, and maintaining a useful quantum state requires continuous active intervention. Quantum error correction schemes demand enormous overhead: thou-

sands of physical qubits to simulate a single logical qubit, with each correction cycle involving measurements and conditional resets that are thoroughly, unavoidably irreversible. The dream of thermodynamically free quantum computation remains precisely that—a dream, constrained not by any fundamental impossibility but by the staggering practical gap between theory and engineering.

2.4 The Thermodynamic Ledger

Let us put numbers to the gap.

The Landauer limit—the absolute minimum energy required to erase one bit of information at room temperature—is $kT \ln 2$, approximately 2.85×10^{-21} joules. This is not a theoretical aspiration. It is a hard floor, derived from the second law of thermodynamics, con-

firmed experimentally in 2012 by a team at the École Normale Supérieure in Paris who measured the heat dissipated by erasing a single bit stored in the position of a colloidal particle.

Now consider a modern NVIDIA H100 GPU, the workhorse of contemporary AI training. It consumes roughly 700 watts while performing approximately 10^{15} floating-point operations per second. That works out to about 7×10^{-13} joules per operation—call it 10^{-12} joules in round numbers, accounting for memory access and overhead.

The ratio between what we *spend* and what physics *requires* is approximately one million. Six orders of magnitude of thermodynamic waste, hiding in plain sight inside every GPU cluster on the planet.

Where does the waste live? Some of it is fundamental: operating at high speed requires driving transistors far from equilibrium, and the excess dissipation grows with clock frequency. Some of it is architectural: the von Neumann bottleneck, the ceaseless shuttling of data between memory and processor, burns energy on transportation rather than transformation. Some of it is material: resistive losses in copper interconnects, leakage currents in transistors that never fully switch off. And some of it is simply historical accident—design choices made decades ago when energy was cheap and performance was the only metric that mattered.

That millionfold gap is not a single problem. It is a layered archaeology of engineering compromises, each stratum representing a different era's priorities and constraints. Under-

standing where the entropy tax is collected—and distinguishing the fundamental from the contingent—is the first step toward building machines that think without burning.

3

The Burning of Intelligence

Training a frontier AI model is an act of industrial energy consumption, and we should be honest about what that means. GPT-3, released in 2020, consumed an estimated 1,287 MWh during training—roughly the annual electricity budget of 120 American homes. That number seemed extraordinary at the time. It was, in retrospect, a rounding error.

GPT-4, by credible estimates from Epoch AI and semi-conductor analysts, consumed between 50 and 100 GWh during its training runs—enough to power 5,000 to 10,000 American homes for a full year. That is not a typo. The jump from GPT-3 to GPT-4 was not a doubling or a tripling. It was an increase of roughly 50 to 80 times.

And the next generation will require four times more again.

3.1 The Training Furnace

We are in an exponential regime, and the exponent shows no sign of bending. The training of a single model now rivals the annual electricity consumption of small nations—Tonga, Nauru, entire Pacific island states whose total grid capacity is dwarfed by a cluster of

NVIDIA H100s running for three months in an Iowa data centre. The energy cost of *knowing*, it turns out, is staggering.

Each training run requires tens of thousands of GPUs operating in concert for weeks or months, connected by specialised networking fabric that itself consumes megawatts. The water used for cooling these clusters—an often-overlooked externality—runs to millions of litres per training run. Microsoft's water consumption rose 34% in a single year, a spike its own environmental report attributed largely to AI workloads. This is not a cottage industry experimenting at the margins. This is heavy industry, with all the material consequences that term implies.

3.2 Inference at Scale

Training is a one-time cost. You pay it once, amortise it across millions of users, and move on. Inference is different. Inference is perpetual.

ChatGPT alone is estimated to consume approximately 500,000 kWh per day—roughly 180 GWh per year. To put that in perspective, it is more electricity than some mid-sized European cities use annually. And ChatGPT is one product, from one company, serving one interface. Behind it lie thousands of API integrations, enterprise deployments, and fine-tuned variants, each drawing their own power.

The International Energy Agency projects that AI data centres could consume 1–2% of global electricity by 2027, up from approximately 1%

in 2023. Goldman Sachs estimates that AI will drive a 160% increase in data centre power demand by 2030. These are not fringe projections from environmental activists. They are consensus forecasts from institutions whose business is getting numbers right.

Every query burns real energy. Every token generated dissipates real heat. Every agent invocation, every chain-of-thought reasoning trace, every retrieval-augmented generation pipeline transforms ordered electrical potential into disordered thermal waste—entropy, flowing inexorably outward, exactly as the second law demands.

The arithmetic is relentless. If AI usage grows at even half the rate that current adoption curves suggest, global AI-related electricity consumption will exceed that of the entire United Kingdom by the end of the decade.

3.3 The Efficiency Counterargument

Optimists—and there are many, particularly in the venture capital community—point to genuine efficiency gains. They are not wrong about the gains. They are wrong about what the gains imply.

Techniques like quantization, knowledge distillation, mixture-of-experts architectures, flash attention, and successive generations of purpose-built silicon have made inference roughly 100 times cheaper per token since GPT-3. That is a real and remarkable achievement. A query that cost $0.06 in 2020 now costs a fraction of a cent. Hardware utilisation has improved. Cooling systems have grown more sophisticated. The cost curve, measured per unit of output, is falling beautifully.

But total energy consumption is rising.

This is Jevons paradox, first articulated by William Stanley Jevons in 1865 when he observed that James Watt's more efficient steam engine did not reduce coal consumption—it *increased* it, dramatically, by making steam power economical for applications that had previously been unthinkable. The more efficient the engine, the more engines were built. The more engines were built, the more coal was burned. Efficiency did not conserve the resource. It *consumed* it.

The same logic applies with full force to AI inference. Cheaper inference means more inference. More inference means more applications—code generation, image synthesis, conversational agents, autonomous research, real-time translation, drug discovery, personalised education. More applications mean more users. More users mean more to-

kens. And more tokens, even at a hundredth of the unit cost, mean more total energy.

The efficiency gains are real. The net effect on consumption is the opposite of what efficiency would naively suggest. History has no examples of a general-purpose efficiency improvement that led to reduced total resource use. Not one.

3.4 The Biological Benchmark

The human brain offers a humbling benchmark for anyone who believes that intelligence must cost what our current machines charge.

It runs on approximately 20 watts—the power of a dim light bulb. It performs an estimated 10^{15} synaptic operations per second, processing language, vision, planning, emotion, and

motor control simultaneously, in a package that weighs 1.4 kilograms and fits inside a skull. The energy cost per operation is roughly 10^{-17} joules—a number so small that it borders on the thermodynamically absurd.

That figure is approximately 100 times more energy-efficient than Landauer's limit would require for equivalent *irreversible* operations at body temperature. This is a profound clue. It suggests that the brain exploits some form of near-reversible, thermodynamically optimised computation—molecular processes that dissipate far less heat per logical operation than any digital circuit we have ever built.

We do not yet understand how it does this. We do not know whether the mechanism is quantum, biochemical, structural, or some combination of all three. What we do know is that evolution, operating under four billion years

of energy scarcity, has produced a cognitive engine that makes our best silicon look like a bonfire compared to a candle.

A single GPT-4 training run consumes more energy than three million human brains would use in the same period. The gap between biological and artificial intelligence, measured in joules per useful cognitive operation, is not a small engineering margin. It is a chasm.

3.5 Is Intelligence Necessarily Expensive?

This brings us to the deepest open question of this essay, and perhaps of computation itself: is the energy cost of machine intelligence a temporary engineering problem, or a fundamental property of intelligence?

If it is engineering—bad architectures, wasteful hardware, brute-force training on undifferentiated data, gradient descent as a blunt instrument—then dramatic improvements are possible. We might build neuromorphic chips that mimic the brain's efficiency. We might discover training algorithms that converge in a thousandth of the compute. We might find that most of what current models learn is redundant, and that a lean, targeted architecture could achieve comparable intelligence at a fraction of the thermodynamic cost. This is the optimistic scenario, and it is not unreasonable.

But there is a darker possibility.

If intelligence *inherently* requires a certain thermodynamic investment—a minimum rate of entropy production to maintain the far-from-equilibrium state that constitutes "knowing"—then there is a floor we cannot go below. The

brain may already be near that floor, and our machines may be approaching it from above, asymptotically, never quite arriving.

Consider what intelligence *is*, thermodynamically. It is a system that maintains internal models of external reality—models that must be continuously updated, error-corrected, and refined against incoming data. This is, by definition, an entropy-reducing process within the system, which means it *must* produce entropy in the environment. The second law is not a suggestion. It is a constraint.

The question is not whether intelligence has an entropy tax. It does. The question is how large that tax is, and whether the receipt is negotiable.

If the minimum thermodynamic cost of human-level intelligence is close to what the

brain already pays—20 watts, give or take—then we may eventually build machines that match that benchmark. But "eventually" may be decades or centuries away, and in the interim, we will be running civilisation-scale intelligence on hardware that is five or six orders of magnitude less efficient than the biological proof of concept.

The implications are profound. If intelligence is expensive by nature, then the dream of ubiquitous, always-on, infinitely scalable AI is not merely an engineering challenge. It is a *thermodynamic* challenge—one that confronts the same unyielding physics that governs every other energy-intensive activity of human civilisation.

We do not get to opt out of the entropy tax. We can only decide how much intelligence we are willing to burn for.

4

The Architecture of Cool

The preceding chapters have described a problem that admits no negotiation: computation dissipates heat, intelligence demands computation, and the exponential appetite of modern AI is steering civilisation toward a thermodynamic wall. The rational response is despair, or at least resignation. The actual response, emerging from a handful of laboratories and semiconductor startups, is

something far more interesting: a wholesale rethinking of what it means to compute.

The conventional processor—the von Neumann machine that has dominated computing since the 1940s—fights physics at every clock cycle. It imposes deterministic order on inherently probabilistic matter, synchronises billions of transistors to a single metronome, and dissipates the resulting entropy as waste heat. This is not the only way to build a thinking machine. It is merely the way we inherited.

4.1 Neuromorphic Dissent

In 2021, Intel released Loihi 2, a neuromorphic research chip containing roughly one million artificial neurons. In 2023, IBM unveiled NorthPole, a neural inference processor that integrates computation and memory on a sin-

gle die, eliminating the energy-hungry data shuttling that dominates conventional architectures. Together, they represent a radical dissent from the orthodoxy of the clock cycle.

The idea is deceptively simple: mimic the brain. Biological neurons do not compute continuously. They fire only when incoming signals exceed a threshold—spikes, not cycles. A neuron that has nothing to say consumes almost nothing. A neuromorphic chip replicates this event-driven architecture in silicon, activating transistors only when data changes, leaving the rest of the circuit quiescent.

The thermodynamic payoff is enormous. IBM's NorthPole achieves approximately 12 trillion operations per second per watt—roughly ten times the efficiency of a state-of-the-art GPU for comparable inference tasks. For workloads that are sparse and event-

driven—sensor fusion, anomaly detection, real-time pattern recognition—the advantage stretches to 100 or even 1,000 times. Intel's Loihi 2 has demonstrated similar gains on neuromorphic benchmarks, processing temporal data streams at a fraction of the energy budget required by conventional hardware.

The trade-offs are real and unromantic. Neuromorphic chips are specialised: they excel at inference but cannot currently perform the massive matrix multiplications that training a large language model demands. They are harder to program, requiring entirely new software frameworks—Intel's Lava, IBM's proprietary toolchains—that bear no resemblance to the CUDA ecosystem that GPU programmers have spent a decade mastering. And they remain, for now, research platforms rather than production infrastructure.

But the thermodynamic argument is unanswerable. A chip that fires only when it has something to say will always, in principle, dissipate less heat than a chip that screams continuously at five billion cycles per second.

4.2 Thermodynamic Computing

If neuromorphic processors represent a reform of conventional computing, thermodynamic computing represents a revolution. The premise is almost perverse: instead of fighting thermal noise—spending enormous energy to maintain deterministic bit states against the ceaseless jostling of molecules—*use* the noise.

A startup called Extropic, founded in 2022 by former quantum computing researchers, is building processors that exploit the natural fluctuations of matter at the nanoscale to per-

form probabilistic computation. At the Santa Fe Institute, researchers have been developing the theoretical foundations of "thermodynamic AI"—systems where the Boltzmann distribution of particle energies directly encodes probability distributions over model parameters.

This is not digital computing. There are no deterministic bits, no logic gates, no clock. The substrate is analog, stochastic, and operates at the boundary between solid-state physics and statistical mechanics. Computation emerges from the system's tendency to settle into low-energy configurations, much as a ball rolling on a landscape finds the nearest valley.

The potential efficiency gains are staggering. Sampling from probability distributions—the core operation of diffusion models, Bayesian inference, and large portions of deep

learning—is thermodynamically expensive on digital hardware because it requires generating pseudo-randomness from deterministic circuits. A thermodynamic computer performs sampling *natively*, using the actual thermal fluctuations of its substrate. The entropy that digital processors spend energy suppressing becomes, in this paradigm, the computational resource itself.

The engineering challenges are commensurate with the ambition. Fabrication tolerances, noise characterisation, and interfacing analog stochastic outputs with digital systems remain unsolved at production scale. But the theoretical floor is breathtakingly low: thermodynamic computation can, in principle, approach the Landauer limit not as an asymptote but as an operating regime.

4.3 The Reversibility Renaissance

Recall Landauer's principle: only *irreversible* operations—those that erase information—incur the minimum thermodynamic tax. The logical corollary is that *reversible* operations, which preserve all information, can in principle dissipate no heat at all.

For decades, reversible computing was a theoretical curiosity, a topic for doctoral dissertations and footnotes in physics textbooks. No longer.

Adiabatic CMOS circuits—designs that recycle the electrical charge used in each switching operation rather than dumping it to ground—have demonstrated 10 to 100 times lower energy dissipation than conventional logic at low clock speeds. The technique is called "charge recovery": instead of discharging a capacitor

through a resistor (the thermodynamic sin of conventional switching), adiabatic circuits use resonant oscillators to gently rock charge back and forth, recovering most of the energy that would otherwise become heat.

DARPA's IRIS program (Innovative Reversible Integrated Systems) has funded multiple university and industry teams to develop partially reversible processor architectures. Superconducting logic circuits, operating at cryogenic temperatures where resistance drops to zero, offer another pathway: the Josephson junction, a quantum device exploiting superconducting tunnelling, can switch with dissipation approaching the Landauer floor.

The engineering tax on reversibility is significant. Reversible circuits require more transistors—roughly three to five times as many for equivalent logical functionality—

because every operation must preserve enough state information to be undone. They are inherently slower at maximum clock speed, because charge recovery requires gentle, quasi-static transitions that resist being hurried. And the design tools, compilers, and programming models for reversible architectures remain embryonic.

But the thermodynamic arithmetic is compelling. If AI workloads grow by four orders of magnitude over the next decade, and conventional efficiency improvements yield at most one order of magnitude, the remaining three orders must come from architectural innovation. Reversible and partially reversible computing is one of the very few paths that can bridge that gap.

4.4 The Quantum Horizon

Quantum computing offers the most dramatic thermodynamic promise—and the most formidable engineering obstacles.

Quantum gates are inherently reversible: every unitary transformation can be undone. In principle, a quantum computer can perform complex computations while dissipating energy only at the input (state preparation) and output (measurement) stages, not during the computation itself. Grover's search algorithm provides a quadratic speedup for unstructured search problems. Quantum Boltzmann machines, proposed by researchers at Google and the University of Toronto, could perform the sampling operations central to generative AI at exponentially lower cost than classical hardware.

The reality, as of 2025, is humbling. Current quantum processors—IBM's Heron, Google's Willow, Quantinuum's H2—operate at temperatures of 10 to 15 millikelvin, maintained by dilution refrigerators that themselves consume tens of kilowatts of electrical power. The qubits are noisy, error-prone, and decohere on timescales of microseconds to milliseconds. Error correction, which is necessary for any useful computation, requires hundreds or thousands of physical qubits per logical qubit, inflating the hardware overhead enormously.

The thermodynamic promise is therefore real but temporally remote. A fault-tolerant quantum computer performing useful AI workloads at lower total energy cost than a classical alternative is, by the most optimistic credible estimates, one to two decades away. By the

most pessimistic, it is a problem for the next century.

What quantum computing contributes to the present conversation is not a solution but a proof of concept: the laws of physics *permit* computation at energy costs far below what we currently accept. The gap between the quantum thermodynamic floor and our current silicon ceiling is not a wall. It is an invitation.

4.5 Convergence

The future of efficient intelligence is not one architecture. It is many.

GPUs will persist for dense linear algebra—the massive matrix multiplications of training runs—where their brute-force parallelism

remains unmatched. Neuromorphic chips will handle sparse, event-driven inference at the edge: autonomous vehicles, industrial sensors, always-on perception systems where every milliwatt matters. Thermodynamic processors will perform the sampling and optimisation that generative models demand, exploiting noise rather than suppressing it. Quantum circuits, when they mature, will attack specific algorithmic bottlenecks—combinatorial optimisation, quantum chemistry simulation, certain classes of machine learning—where their exponential or quadratic advantages justify the cryogenic overhead.

The engineering challenge of the next decade is not selecting the "best" architecture. It is building the orchestration layer—the heterogeneous runtime, the thermodynamically

aware scheduler—that routes each computation to its optimal substrate. A training run that begins on GPUs, delegates sampling to a thermodynamic coprocessor, offloads sparse inference to a neuromorphic accelerator, and dispatches specific subroutines to a quantum backend is not science fiction. It is the logical endpoint of taking the entropy tax seriously.

The entropy tax is not abolished. Landauer's limit does not yield to cleverness, funding, or good intentions. But it can be *approached*—operation by operation, architecture by architecture, each computation matched to the substrate that pays the smallest possible tithe to the second law.

The architecture of cool is not a single blueprint. It is a portfolio of thermodynamic strategies, assembled with the same care a financial engineer applies to managing risk: di-

versified, adaptive, and relentlessly optimised against a cost that the universe will never stop collecting.

5

Epilogue: A Memo to the Engineers

This essay began with physics and ends with choices. The physics is not negotiable. Every computation dissipates heat. Every bit erased pays a tax to the second law. Every intelligence—carbon or silicon—runs on energy drawn from somewhere and dumped as waste heat into the universe. Landauer's limit is not a suggestion. It is a floor, confirmed

experimentally, derived from the same thermodynamic principles that govern the burning of stars and the cooling of corpses. We cannot think for free. We never could.

But what we do about it is entirely up to us.

The entropy tax is real. The question is whether we treat it as an inconvenient externality or as the central design constraint of the century. What follows is addressed, in turn, to the people whose decisions will determine the answer.

5.1 If You Are Building AI

Report energy alongside accuracy.

This is not a polite request. It is a professional obligation that the field has neglected for too long. Every paper submitted to NeurIPS, ev-

ery model card published on Hugging Face, every benchmark result announced in a corporate blog post should include the energy consumed during training and the energy consumed per unit of inference. Not as a footnote. Not as an appendix. As a first-class metric, displayed with the same prominence as perplexity, accuracy, or F1 score.

The machine learning community has optimised relentlessly for performance—for state-of-the-art results on benchmarks that measure capability while ignoring cost. We celebrate models that achieve a half-point improvement on MMLU while consuming twice the energy of their predecessors. We publish scaling laws that chart the relationship between compute and capability with exquisite precision, then treat the electricity bill as someone else's prob-

lem. This is intellectually dishonest and practically unsustainable.

The next great breakthrough in artificial intelligence may not be a better architecture. It may be a more efficient one. Energy per token, energy per training run, energy per useful output—these are the metrics that will matter when the thermodynamic wall arrives, and it is arriving faster than most researchers care to admit.

Efficiency is not a concession to environmentalism. It is not a marketing exercise. It is the next frontier of engineering excellence, and it deserves the same prestige, the same conference tracks, the same tenure cases as raw capability. The researcher who halves the energy cost of inference at constant quality has done more for the future of intelligence than the one who doubles the parameter count.

5.2 If You Are Funding AI

Demand energy audits from every company you finance.

You would not invest in a manufacturing firm without understanding its unit economics. You would not fund a logistics company without examining its fuel costs. Yet venture capital and sovereign wealth funds pour billions into AI companies whose energy consumption is treated as a rounding error—an operational detail, a line item buried in a data-centre lease, someone else's problem until the next funding round.

This is negligence dressed as optimism.

Understand the energy cost of the models you are financing. Ask how many megawatt-hours each training run consumes. Ask what the in-

ference cost per query is, and how it scales with demand. Ask whether the company has a roadmap for reducing energy per unit of output, or whether it plans to scale by brute force alone.

And fund fundamental research. Not just the next foundation model, but the substrate it runs on. Thermodynamic computing, neuromorphic hardware, reversible circuit architectures, optical interconnects, in-memory processing—these are not speculative curiosities. They are the technologies that will determine whether artificial intelligence remains the province of a handful of corporations with access to gigawatt-scale power, or becomes infrastructure available to the world.

The bottleneck is shifting. For two decades, the constraint on AI was intelligence: we did not know how to build systems that could rea-

son, generate, and learn. That constraint is loosening. The new constraint is energy. Your capital allocation should shift accordingly, or you will find yourself holding equity in systems that are brilliant but too expensive to run.

5.3 If You Are Thinking About AI

Consider what happens when the cost of intelligence determines its distribution.

If intelligence becomes cheap—if efficiency gains compress the energy cost of inference by orders of magnitude—then intelligence becomes ubiquitous. Available to every person, every business, every school, every clinic, every government, every nation. A world in which a subsistence farmer in Bihar has access to the same analytical power as a hedge fund

in Greenwich is a world transformed in ways we have barely begun to imagine.

If intelligence remains expensive—if the entropy tax stays high and the efficiency gains stall—then intelligence concentrates. It accrues to the organisations that can afford the energy bill: the hyperscalers, the sovereign AI projects, the well-capitalised few. The rest of the world becomes a consumer of intelligence produced elsewhere, on terms set by others, subject to dependencies as profound as any oil import.

This is not a technical abstraction. It is a question of political economy as fundamental as the distribution of electricity was in the twentieth century. The nations that electrified broadly became democracies with broad-based prosperity. The nations where electricity remained a luxury became something else.

The entropy tax will shape the twenty-first century in precisely the same way.

Policymakers should understand Landauer's limit the way they understand interest rates—as a fundamental constraint that governs the cost of a critical resource. Citizens should understand that the energy debate and the AI debate are not separate conversations. They are the same conversation, conducted in different rooms by people who rarely talk to each other. It is time to open the door between them.

5.4 The Watt and the Thought

Let us return, one final time, to where we began.

A human computer at the Aberdeen Proving Ground, calculating artillery tables by

hand, burned roughly 1.5 calories per minute above her basal rate. A modern GPU cluster training a frontier language model consumes megawatts—the output of a small power station, sustained for weeks. The scale has changed by a factor of billions. The principle has not changed at all.

Every thought requires energy. Every bit erased leaves a thermodynamic trace. Every inference computed, every gradient descended, every token predicted exacts a cost that is ultimately paid in heat dumped into the atmosphere of a planet that is already warming.

Rolf Landauer's quiet paper in 1961 told us something we were not ready to hear: that information is physical, that computation is thermodynamic, that the act of knowing is inseparable from the act of dissipating. Charles Ben-

nett showed us that the tax falls specifically on forgetting—on the irreversible erasure of information that conventional architectures perform billions of times per second without a thought.

We cannot repeal the second law. We cannot negotiate with entropy. But we can close the millionfold gap between what our machines dissipate and what physics requires. We can build architectures that approach the Landauer limit rather than insulting it. We can fund the research, demand the metrics, and make the policy choices that determine whether intelligence becomes humanity's great equaliser or its most expensive luxury.

The universe charges rent for intelligence. The lease is non-negotiable. But how we furnish the apartment—how wisely we spend our entropy budget, how carefully we choose which

computations to perform, how equitably we distribute the intelligence we purchase—is entirely up to us.

References & Further Reading

Landauer, R. (1961). "Irreversibility and Heat Generation in the Computing Process." *IBM Journal of Research and Development*, 5(3), 183–191.

The foundational paper establishing that erasure of information has an irreducible thermodynamic cost: $kT \ln 2$ per bit.

Bennett, C. H. (1973). "Logical Reversibility of Computation." *IBM Journal of Research and Development*, 17(6), 525–532.

Proved that computation can in principle be per-

formed reversibly, sidestepping Landauer's limit at the cost of memory and time.

Bérut, A. et al. (2012). "Experimental Verification of Landauer's Principle Linking Information and Thermodynamics." *Nature*, 483, 187–189.
The first direct experimental confirmation of Landauer's bound, using a colloidal particle in a double-well potential.

Patterson, D. et al. (2022). "The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink." *IEEE Computer*, 55(7), 18–28.
An optimistic analysis of efficiency trends in ML training, arguing hardware and algorithmic improvements will bend the energy curve.

Strubell, E., Ganesh, A., and McCallum, A. (2019). "Energy and Policy Considerations

for Deep Learning in NLP." *Proceedings of the 57th Annual Meeting of the ACL.*
The paper that first brought widespread attention to the carbon cost of training large NLP models.

Davies, M. et al. (2021). "Advancing Neuromorphic Computing with Loihi." *Proceedings of the IEEE*, 109(5), 911–934.
A comprehensive overview of Intel's neuromorphic computing platform and its energy-efficiency advantages for inference workloads.

Hylton, T. (2020). "Thermodynamic Computing." Research report, IARPA / Santa Fe Institute.
Lays out the theoretical framework for computation that harnesses thermal noise rather than fighting it.

Feynman, R. P. (1982). "Simulating Physics with Computers." *International Journal of Theoretical Physics*, 21(6/7), 467–488.

Feynman's visionary proposal for quantum simulation, anticipating the thermodynamic advantages of quantum computation.

About the Author

Nabil El Mahyaoui

https://nabil-el-mahyaoui.com/

https://www.linkedin.com/in/elmahyaoui