Capacity-Optimal Codes Are Not Energy-Optimal: The Synaptic Efficiency Coefficient as an Architectural Prior for Language Models

Paulo Câmara

LUA Vision Tecnologia — São Paulo, Brazil

paulo@lua.vision

> *Author note. The author holds a doctorate and postdoctoral training in neuroscience, with research on attention and synaptic plasticity. The models, the training methodology, and the results reported here were developed at LUA Vision in São Paulo, Brazil.*

Version 1.1, 21 September 2026. Corrects the energy-per-bit arithmetic in Section 3.2 and Figure 1: the synaptic cost reported by Laughlin et al. (1998) corresponds to about 3.6×10^5 times the Landauer limit, and graded signals or spike coding to 3.6×10^7 to 3.6×10^8. The argument and the conclusions are unchanged.

Abstract

Scaling parameters improves accuracy per parameter, and that metric has governed a decade of language model architecture. Levy and Baxter (1996) proved that coding schemes with maximal representational capacity are not optimal once energy expenditure enters the objective function, a result the field never absorbed. This work recovers that result and proposes the Synaptic Efficiency Coefficient (SEC), an inductive bias derived from quantitative data on atypical cortical development. The SEC is the product of three terms: pruning rate (α), potentiation selectivity (β), and consolidation efficiency (γ). The multiplicative form follows from a three-stage cascade over nested synaptic populations, and we show that additive composition violates the absorption condition the phenomenon requires. Each term is anchored in published measurement: dendritic spine density in human post-mortem tissue (n=20, two-way ANOVA, interaction $F(1,16)=28.35$, $p<0.001$), long-term potentiation with NR2A/NR2B overexpression in an animal model, and fear-memory persistence resistant to extinction in the same model. We derive a correspondence between each term and a verifiable architectural decision, including an analytically derived sparse routing threshold that is not obtained by hyperparameter search. Internal results indicate three to four orders of magnitude fewer operations per inference than open models of comparable size, with abstention concentrated outside the trained domain. We propose accuracy per joule as an evaluation metric and state the conditions that would refute the coefficient.

Keywords: inductive bias, energy-efficient computation, synaptic pruning, sparse activation, neurodivergent cognition, language model architecture

1. Introduction

A human brain runs on roughly twenty watts. On that budget it sustains attention, resolves ambiguity, and retrieves a face last seen two decades ago. A single accelerator board in a data centre draws more than a thousand watts.

The asymmetry is usually treated as a curiosity. We treat it as a specification.

This work was carried out in Brazil, and the constraint is not hypothetical here. Research groups outside the small set of institutions with unbounded compute budgets do not get to postpone the energy question until after training. That position is usually described as a disadvantage. We argue it is an epistemic advantage: a binding constraint forces the correct objective function into view, and the objective function is what this paper is about.

The dominant research programme of the past decade optimises accuracy per parameter. Sutton (2019) articulated the underlying position: general methods that leverage computation eventually dominate methods that embed human knowledge. The historical record he cites is real. Hand-coded structure in chess, in speech recognition, and in vision was overtaken by systems that simply scaled.

That argument carries a premise it does not state. Leveraging computation presupposes computation available to leverage. The premise held while compute grew faster than demand for it. Energy cost, data-centre siting, and sovereignty requirements are ending that condition, and when the binding constraint changes the optimisation target must change with it.

This paper makes four contributions.

First, we recover a result from Levy and Baxter (1996) showing that capacity-maximal codes are not energy-optimal, and we argue that the field's objective function has been misspecified since before the architectures in question existed.

Second, we define the Synaptic Efficiency Coefficient, SEC = $\alpha \cdot \beta \cdot \gamma$, and derive its multiplicative form from a cascade of nested synaptic populations rather than asserting it.

Third, we anchor each term in published quantitative measurement, drawn from atypical cortical development rather than from the typical case, and we justify that choice on efficiency grounds rather than on grounds of representation.

Fourth, we map each term onto a verifiable architectural decision and propose accuracy per joule as the evaluation metric that follows.

We state throughout what would refute the proposal.

2. Related Work

Scaling and inductive bias. Sutton (2019) is the reference statement of the scaling position. Recent evidence supports it in specific settings: non-equivariant networks reach competitive performance on geometric tasks where symmetry priors appeared mandatory. We take this seriously and delimit it. The result holds in the regime where the required compute is available, and says nothing about the cost of arriving there.

No free lunch, and its misuse. Wolpert (1996) established that no learning algorithm outperforms another when averaged over all possible target functions. The result is frequently stretched past what it supports. Goldblum et al. (2024) show that uniformly sampled datasets have high complexity while real problems generate low-complexity data, and that neural networks already prefer low-complexity outputs; a general simplicity prior therefore covers a wide class of problems. We accept this. Our use of Wolpert is narrower: it establishes only that some bias is unavoidable. The question that follows is which bias, and at what cost to install, and that question sits on the efficiency axis rather than the generalisation axis Goldblum et al. address.

Plasticity in artificial networks. Miconi et al. (2018) introduced differentiable plasticity, in which a connection carries a baseline weight and a Hebbian component whose contribution is learned. Miconi et al.

(2019) extended this to neuromodulated plasticity. Meta-learning of biologically plausible plasticity rules has been demonstrated with random feedback pathways (Shervani-Tabar and Rosenbaum, 2023). This line establishes that plasticity is trainable. It has not addressed the regime argument: why plasticity moves from a research curiosity to a design constraint once the energy budget binds.

Efficient coding. Barlow (1961) proposed that sensory systems recode to reduce redundancy. Olshausen and Field (1996) showed that imposing sparsity on a natural image model causes V1-like receptive fields to emerge without supervision, demonstrating that an appropriate prior substitutes for data.

Energy as a design variable. Mead (1990) argued that subthreshold analog VLSI, where drain current depends exponentially on gate voltage as ionic flux depends on membrane potential, permits computation at currents orders of magnitude below digital logic. The principle preceded the tooling by three decades. Izhikevich (2004) introduced the relevant comparison for neuron models: floating-point operations per simulated millisecond against the repertoire of behaviours reproduced. Hodgkin-Huxley requires roughly 1,200 operations, integrate-and-fire roughly 5, and his own model 13 with nearly the full repertoire. Behaviour per operation has not been carried up a level of abstraction.

3. The Energy-Capacity Tradeoff

3.1 The result the field did not absorb

Levy and Baxter (1996) analysed neural codes under a metabolic constraint. They derived an optimal firing probability for a given mean rate, and established the following:

> *Coding schemes with the largest representational capacity are not, in general, optimal when energy expenditures are taken into account.*

The statement is stronger than it appears. It does not say that efficient codes are a reasonable compromise. It says that maximising capacity and maximising efficiency are distinct objectives with distinct optima, and that optimising the first does not approach the second.

Contemporary practice optimises the first and treats the second as a deployment concern, addressed after training through quantisation and distillation. Under the Levy-Baxter result, this ordering cannot recover the efficiency optimum, because the architecture was selected under the wrong objective before compression was applied.

3.2 How far from the floor

Landauer (1961) established the thermodynamic minimum for erasing one bit:

$$E \geq k \cdot T \cdot \ln 2 \tag{1}$$

At 310 K this yields 2.97×10^{-21} J.

Laughlin et al. (1998) measured the metabolic cost of information transmission in the blowfly visual system. A chemical synapse costs approximately 10^4 ATP molecules per bit; graded signals in interneurons and photoreceptors, and spike coding, cost 10^6 to 10^7 ATP per bit. With ATP hydrolysis releasing about 25 kT (≈

1.07×10^{-19} J at 310 K), a bit crossing a chemical synapse costs roughly 1.1×10^{-15} J, and a bit carried by graded signals or spikes between 10^{-13} and 10^{-12} J.

The ratios are instructive:

$$1.1 \times 10^{-15} / 2.97 \times 10^{-21} \approx 3.6 \times 10^{5} \text{ (chemical synapse)} \tag{2}$$

$$10^{-13} \text{ to } 10^{-12} / 2.97 \times 10^{-21} \approx 3.6 \times 10^{7} \text{ to } 3.6 \times 10^{8} \text{ (graded signals, spikes)}$$

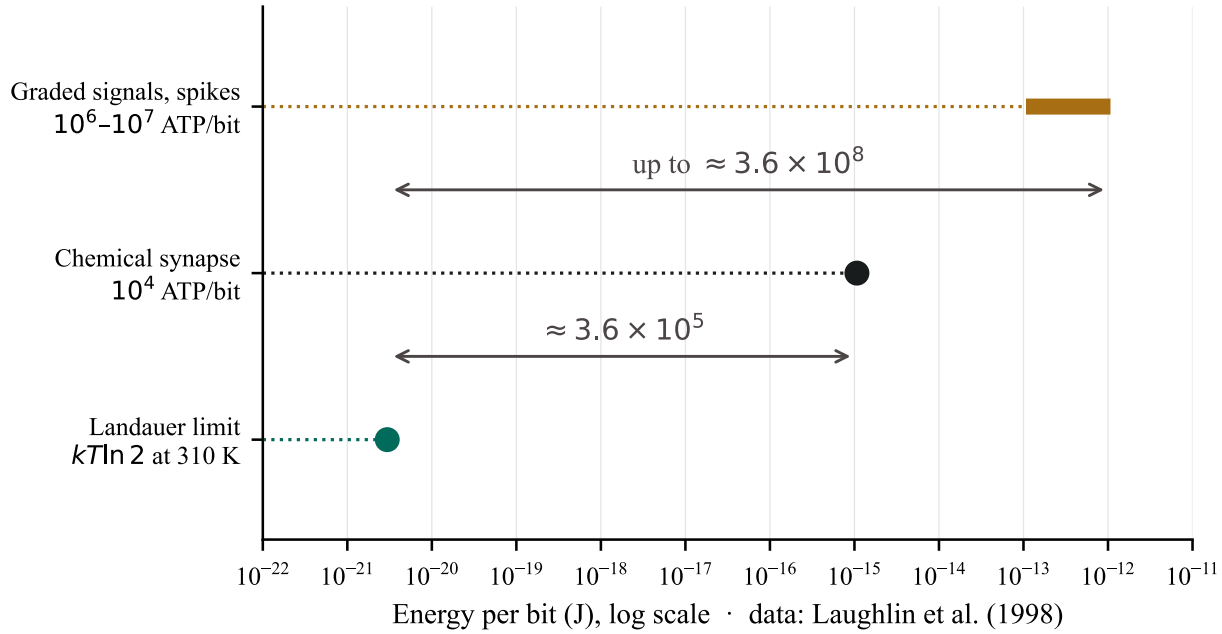


Figure 1. Energy per bit on a logarithmic scale. The Landauer limit at 310 K against measured transmission costs. The interval spans five to eight orders of magnitude, depending on the stage.

Biological information transmission operates five to eight orders of magnitude above the physical floor, depending on the stage; Laughlin and colleagues describe it as several orders of magnitude above the thermodynamic minimum. The most efficient information-processing substrate known is nowhere near the limit its own physics imposes. Any argument that the design space is exhausted, and that scale is the only remaining axis, is inconsistent with this measurement.

3.3 The cost that did not fall

Attention compares each token against every other token, at $O(n^2 \cdot d)$ operations. Doubling context quadruples the bill, paid in full at every token.

FlashAttention (Dao et al., 2022) is widely described as having solved this. It did not, and the title states as much: exact attention with IO-awareness. Tiling and recomputation avoid materialising the n×n matrix, reducing memory cost from $O(n^2)$ to $O(n)$. The FLOP count remains quadratic, and the algorithm performs more arithmetic than the naive formulation while running faster, because the binding constraint on the device is data movement between high-bandwidth memory and on-chip SRAM rather than arithmetic throughput.

This is excellent engineering against the wrong constraint for our purposes. It made the quadratic cost tolerable. It did not make it smaller.

4. The Synaptic Efficiency Coefficient

4.1 Definition

Let S_0 denote the initial synaptic repertoire, overproduced during development. Define functional efficiency as retained functional capacity per unit metabolic cost.

The coefficient decomposes into three sequential stages.

Structural selection. A fraction $\alpha \in (0,1]$ survives elimination, where α measures the proportion of retained synapses carrying functional value.

Efficacy gain. Each surviving synapse acquires relative efficacy $\beta \geq 1$ over baseline.

Temporal retention. A fraction $\gamma \in (0,1]$ of acquired efficacy persists under interference.

Effective functional capacity at the end of the cascade:

$$C = S_0 \cdot \alpha \cdot \beta \cdot \gamma \tag{3}$$

$$SEC \equiv C / S_0 = \alpha \cdot \beta \cdot \gamma \tag{4}$$

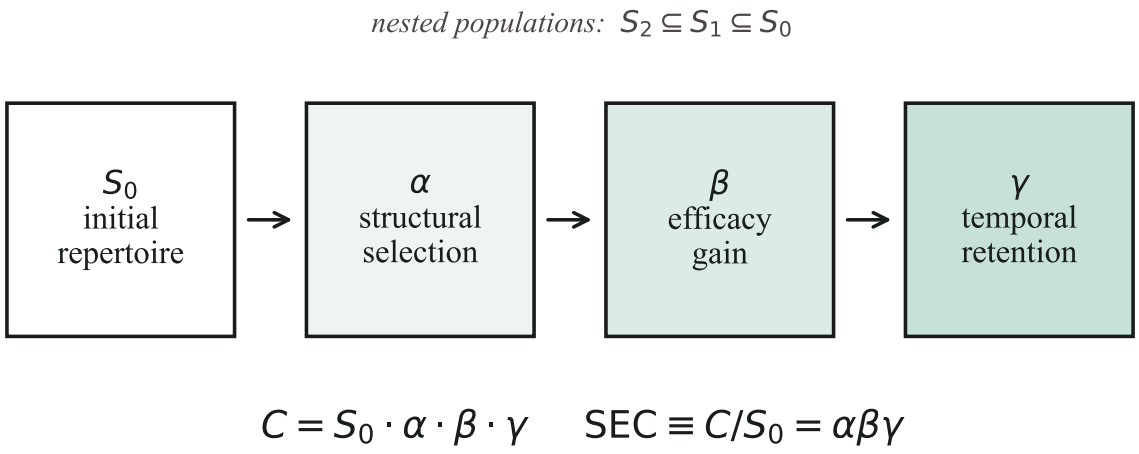


Figure 2. The three-stage cascade over nested synaptic populations. Each stage acts on the output of the previous one.

4.2 Why the product, and not the sum

Absorption. If any stage vanishes, the outcome vanishes. With $\alpha = 0$ no synapse is retained, and no value of β or γ recovers capacity. Additive composition violates this: $\alpha + \beta + \gamma$ with $\alpha = 0$ remains positive. Only the product respects the structure of the process.

Nesting. The three stages operate over nested populations, $S_2 \subseteq S_1 \subseteq S_0$, in distinct temporal windows. Each acts on the output of the previous one rather than in parallel. Sequential transformation of a single quantity composes multiplicatively by construction, as yield does in a cascade.

Logarithmic additivity. Taking logarithms,

$$\log SEC = \log \alpha + \log \beta + \log \gamma \tag{5}$$

Gains add on a logarithmic scale, which is the formal content of exponential composition. A twenty percent gain in each term yields 1.728, not 1.2. This accounts for why modest advantages at each stage produce large differences in outcome.

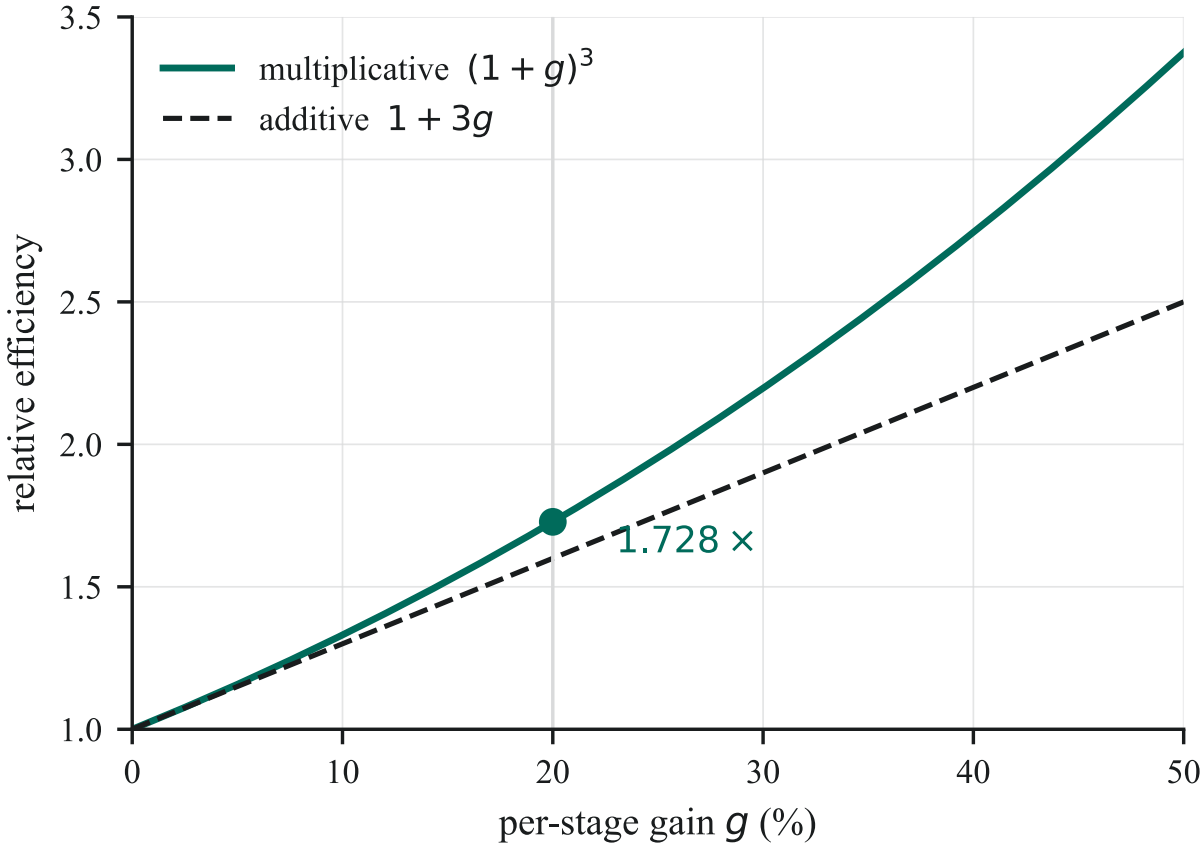
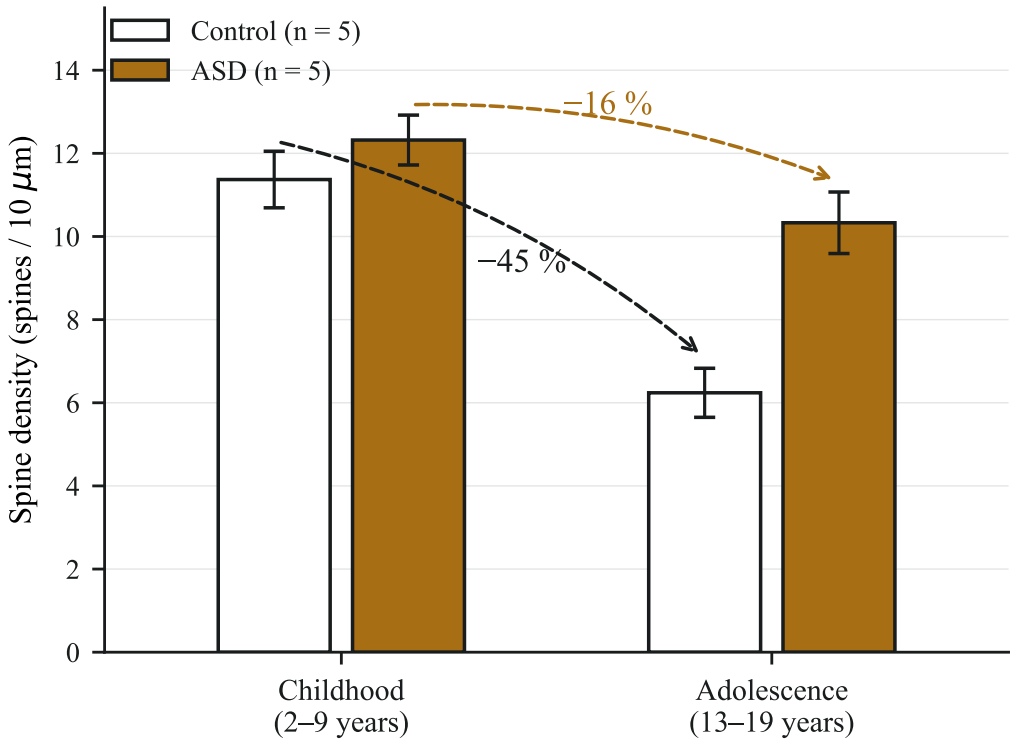


Figure 3. Multiplicative against additive composition of per-stage gains. At a twenty percent gain per stage the two forms differ by a factor of 1.08; at forty percent the gap widens to 1.25.

4.3 Empirical anchoring

We anchor the coefficient in atypical development rather than in the typical case. The justification is not clinical. It is that atypical development dissociates the three stages, exposing terms that covary in the typical case.



Two-way ANOVA: diagnosis $F(1, 16) = 73.11$; age $F(1, 16) = 145.7$; interaction $F(1, 16) = 28.35$; all $p < 0.001$.

Figure 4. Dendritic spine density in layer V pyramidal neurons, BA21 basal dendrites, human post-mortem tissue. Mean ± SEM, n = 5 per cell. Data from Tang et al. (2014).

α, from Tang et al. (2014). Dendritic spine density in layer V pyramidal neurons of BA21, basal dendrites, human post-mortem tissue. Controls in childhood: 11.37 ± 0.68 per 10 µm (n=5); in adolescence 6.24 ± 0.59 (n=5). ASD in childhood: 12.32 ± 0.60 (n=5); in adolescence 10.33 ± 0.74 (n=5). Net reduction of 45% against 16%. Regression slopes: controls −0.40, R^2=0.93; ASD −0.19, R^2=0.55. Two-way ANOVA: diagnosis $F(1,16)=73.11$, $p<0.001$; age $F(1,16)=145.7$, $p<0.001$; interaction $F(1,16)=28.35$, $p<0.001$. The mechanism reported is mTOR hyperactivation with consequent autophagy inhibition, and roughly 70% of postnatal spine elimination is autophagy-dependent.

β, from Markram and Markram (2010). Valproate model, 500 mg/kg at embryonic day 12.5. Hyper-connectivity of +50% within a radius below 50 µm. Long-term potentiation doubled across somatosensory cortex, prefrontal cortex, and amygdala. NR2A and NR2B overexpressed more than twofold with AMPA and NR1 unchanged. Network response twice as strong despite individual hypo-excitability.

γ, from the same work. Conditioned fear memory amplified and persisting for up to three months, generalising to a previously non-aversive tone, and more resistant to extinction than in controls. Resistance to extinction is, operationally, consolidation efficiency, and persistence under interference is its measure.

The computational analogue is elastic weight consolidation (Kirkpatrick et al., 2017), which weights parameters by the Fisher information matrix. We cite it as an analogy rather than as a foundation, and note that the penalty was contested by Huszár with a published reply from the original authors (2018).

5. From Coefficient to Architecture

Each term maps onto a design decision. We state the principle and withhold the parameterisation.

Hyper-connectivity within a local radius maps to recurrent state carried across layers rather than recomputed, and to a layer pattern alternating recurrent and attention blocks at a fixed ratio.

Hypo-excitability with strong network response maps to sparse expert routing through a sigmoid gate whose threshold is derived analytically from an irrational constant. The threshold is not obtained by hyperparameter search. This distinction matters: a searched value is a local optimum over the training distribution, while a derived value is a commitment that the architecture either satisfies or fails.

Dual NMDA pathway with differential subunit expression maps to differential attention with separate key and value projections.

Expansion capacity under stimulation maps to elastic parameter activation, in which the active parameter count varies with input difficulty rather than remaining fixed.

Pruning as a developmental stage rather than a deployment step maps to directed sparsification during training. This inverts standard practice, where pruning is applied after training to fit hardware. Under the cascade in Section 4, post-hoc pruning cannot recover α, because selection was never part of the objective.

Metacognitive probes at intermediate depth provide a confidence signal used for abstention, with a refusal threshold derived from a self-referential fixed point rather than calibrated on a validation set.

The dense-network expansion ratio is a rational approximation of a transcendental constant, chosen on the same grounds.

We do not publish layer counts, hidden dimensions, head configurations, expert counts, probe indices, threshold values, or elastic activation ranges. The principles above are sufficient to evaluate the argument and insufficient to reconstruct the system.

6. Empirical Indications

We maintain an internal research line, without transformer attention, oriented to energy efficiency. Its architecture is not the subject of this paper. Results are sealed with per-run cryptographic hashes and available to reviewers under agreement.

Against open models, where operation counts are exact and any reader can derive them from published architecture, we measure **4,169× fewer operations per inference than Phi-4 3.8B** and **8,750× fewer than Llama-3 8B**. Taking the public estimate of approximately 90 T FLOPs per inference for current frontier systems, the ratio falls in the range of 10^5 to 10^6; we label this an estimate, since operation counts for closed systems cannot be verified.

On a reasoning set of 42 cases spanning fifteen cognitive domains, the system reached **73.8% accuracy with a 16.7% abstention rate**. Abstention was not uniform. In causal reasoning, five of eight cases produced abstention rather than an answer. This is the behaviour the coefficient predicts: a system operating outside its trained domain declines rather than confabulates.

These results do not establish the coefficient. They indicate that a path exists outside the dominant design, and that it warrants measurement on the appropriate axis.

7. Limitations and Falsifiability

The empirical section reports internal results. Operation counts against open models are independently derivable; the remaining figures are not reproducible by third parties without access under agreement, and should be weighted accordingly.

Energy figures in the current implementation are derived from device power measurement rather than from per-operation instrumentation, and projections to neuromorphic substrates are projections. We report them separately from measurements and label them as such.

The coefficient is refuted under either of two conditions.

Statistical dependence. The factorisation presumes separability of the three stages. If α, β, and γ show significant statistical dependence when measured in the same tissue within the same developmental window, the multiplicative form does not hold.

Ordering failure. If an architecture with lower SEC consistently outperforms one with higher SEC under matched energy budget and matched evaluation set, the coefficient does not predict what it claims to predict.

We regard the second as the decisive test and consider it tractable with current instrumentation.

8. On where this was built

A note on provenance, because it bears on the argument rather than on credit.

Brazil is an emerging market with a large domestic corpus, a data protection statute that restricts cross-border transfer, and electricity prices that make inference cost a line item rather than a rounding error. Every one of those conditions pushes toward the objective function proposed here, and away from the one the field has been optimising.

A model trained elsewhere and translated does not solve the problem. It inherits a tokenisation tuned to another language, an alignment procedure shaped by another legal regime, and a corpus in which the long tail of this market does not appear. These are not deployment details. They are properties of the weights.

The work reported here was done in São Paulo, by a Brazilian group, under those constraints. We regard that as a methodological fact worth stating: the constraint produced the question.

9. Discussion

The disagreement with the scaling position is not historical. It concerns which regime the field is in.

Sutton's account describes what happens when computation is abundant and cheap. It describes an era, and eras close. When computation becomes expensive again, the question of which rule governs weight change stops being a matter for seminars and becomes a budget decision.

Our claim is narrower than it may appear. We do not argue that scale fails. We argue that the objective function under which scale was selected is misspecified for the constraint the field is entering, that a correctly specified objective was published in 1996, and that the biological record contains a tested solution to the same optimisation problem.

Much of the answer was written between 1949 and 1996. The field was well funded enough not to need it.

Declarations

Competing interests. The author is the founder and chief technology officer of LUA Vision Tecnologia, which develops language models, and holds equity in the company. The internal results mentioned in Section 6 were obtained on LUA Vision models.

Funding. No external funding was received for this work.

Data and code availability. This is a theoretical study; no new experimental data were generated. All quantitative values used in the argument are taken from the cited publications and reported in the text. Code that reproduces Figures 1 to 4 is provided as Source code 1 to 4 (Python, numpy and matplotlib) and deposited with the preprint at https://doi.org/10.5281/zenodo.22868263. The internal benchmark results mentioned in Section 6 are proprietary to LUA Vision and are not required for the theoretical argument; they can be made available to editors and reviewers under a confidentiality agreement.

Figures

References

Attwell, D., Laughlin, S. B. (2001). An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow and Metabolism* 21(10), 1133–1145. DOI: 10.1097/00004647-200110000-00001

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In *Sensory Communication*, MIT Press, 217–234.

Bienenstock, E. L., Cooper, L. N., Munro, P. W. (1982). Theory for the development of neuron selectivity. *Journal of Neuroscience* 2(1), 32–48. DOI: 10.1523/JNEUROSCI.02-01-00032.1982

Dao, T., Fu, D. Y., Ermon, S., Rudra, A., Ré, C. (2022). FlashAttention: fast and memory-efficient exact attention with IO-awareness. *NeurIPS 35*. arXiv:2205.14135

Giurfa, M., Zhang, S., Jenett, A., Menzel, R., Srinivasan, M. V. (2001). The concepts of 'sameness' and 'difference' in an insect. *Nature* 410, 930–933. DOI: 10.1038/35073582

Goldblum, M., Finzi, M., Rowan, K., Wilson, A. G. (2024). The no free lunch theorem, Kolmogorov complexity, and the role of inductive biases in machine learning. arXiv:2304.05366

Hebb, D. O. (1949). *The Organization of Behavior*. Wiley.

Huszár, F. (2018). Note on the quadratic penalties in elastic weight consolidation. *PNAS* 115(11), E2496–E2497. DOI: 10.1073/pnas.1717042115

Izhikevich, E. M. (2004). Which model to use for cortical spiking neurons? *IEEE Transactions on Neural Networks* 15(5), 1063–1070. DOI: 10.1109/TNN.2004.832719

Kirkpatrick, J. et al. (2017). Overcoming catastrophic forgetting in neural networks. *PNAS* 114(13), 3521–3526. DOI: 10.1073/pnas.1611835114

Kirkpatrick, J. et al. (2018). Reply to Huszár: the elastic weight consolidation penalty is empirically valid. *PNAS* 115(11), E2498. DOI: 10.1073/pnas.1800157115

Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development* 5(3), 183–191. DOI: 10.1147/rd.53.0183

Laughlin, S. B., de Ruyter van Steveninck, R. R., Anderson, J. C. (1998). The metabolic cost of neural information. *Nature Neuroscience* 1(1), 36–41. DOI: 10.1038/236

Levy, W. B., Baxter, R. A. (1996). Energy efficient neural codes. *Neural Computation* 8(3), 531–543. DOI: 10.1162/neco.1996.8.3.531

Markram, K., Markram, H. (2010). The intense world theory: a unifying theory of the neurobiology of autism. *Frontiers in Human Neuroscience* 4, 224. DOI: 10.3389/fnhum.2010.00224

Mead, C. (1990). Neuromorphic electronic systems. *Proceedings of the IEEE* 78(10), 1629–1636. DOI: 10.1109/5.58356

Miconi, T., Stanley, K. O., Clune, J. (2018). Differentiable plasticity: training plastic neural networks with backpropagation. *ICML 35*, 3559–3568.

Miconi, T., Rawal, A., Clune, J., Stanley, K. O. (2019). Backpropamine: training self-modifying neural networks with differentiable neuromodulated plasticity. *ICLR*.

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology* 15(3), 267–273. DOI: 10.1007/BF00275687

Olshausen, B. A., Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609. DOI: 10.1038/381607a0

Shaw, P. et al. (2006). Intellectual ability and cortical development in children and adolescents. *Nature* 440, 676–679. DOI: 10.1038/nature04513

Shervani-Tabar, N., Rosenbaum, R. (2023). Meta-learning biologically plausible plasticity rules with random feedback pathways. *Nature Communications* 14, 1805. DOI: 10.1038/s41467-023-37562-1

Sutton, R. (2019). The bitter lesson. Incomplete Ideas, 13 March 2019.

Tang, G. et al. (2014). Loss of mTOR-dependent macroautophagy causes autistic-like synaptic pruning deficits. *Neuron* 83(5), 1131–1143. DOI: 10.1016/j.neuron.2014.07.040

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation* 8(7), 1341–1390. DOI: 10.1162/neco.1996.8.7.1341