



## Research Article

# The Thermodynamic Cost of Neural Computation: From Landauer's Bound to the Energetic Gap of Artificial Intelligence

Edwar D Montenegro<sup>1,2\*</sup>

<sup>1</sup>Postgraduate Program in Materials Science and Engineering, Federal University of Piauí, Teresina, Piauí, Brazil

<sup>2</sup>Researcher, Graviton Scientific Society, Teresina, Piauí, Brazil

Submitted : 07 September, 2026

Accepted : 15 September, 2026

Published : 16 September, 2026

\*Corresponding author: Edwar D Montenegro, Postgraduate Program in Materials Science and Engineering, Federal University of Piauí, Teresina, Piauí, Brazil, E-mail: [e.montenegro@ufpi.edu.br](mailto:e.montenegro@ufpi.edu.br)

**Keywords:** Landauer principle; Artificial intelligence; Neural computation; Energy efficiency; Thermodynamics of computation; Large language models

**Copyright License:** © 2026 Montenegro ED. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

<https://www.engineergroup.us>



## Abstract

Landauer's principle establishes a thermodynamic lower bound for physical processes that reduce the logical entropy of information-bearing degrees of freedom, but it does not assign a universal energy cost to every floating-point operation, multiply-accumulate instruction, or transistor transition. This review develops a hierarchical framework for locating energy costs in neural computation while distinguishing three quantities that are frequently conflated: process-specific thermodynamic lower bounds, Landauer-normalized reference ratios, and measured or modelled operational energy. The analysis synthesizes results from stochastic thermodynamics, finite-time and quantum Landauer theory, thermodynamic control, mismatch cost, device physics, computer architecture, and full-stack AI serving. The resulting hierarchy separates logical entropy reduction, finite-time corrections, quantum and strong-coupling effects, control costs, distribution-dependent mismatch, architectural and data-movement costs, and facility overhead. The literature supports rigorous bounds for specified physical processes, but not a universal additive decomposition of the energy consumed by a large-language-model query. An operational accounting analysis shows how accelerator throughput, utilization, host power, sequence length, memory traffic, and facility overhead determine query-level energy. Published full-stack studies place representative text-prompt energies in the sub-watt-hour range while revealing substantial variation across models and serving conditions. Contemporary dissipation is therefore dominated by implementation and systems choices long before fundamental erasure bounds become active constraints. The framework provides a consistent basis for evaluating emerging computing paradigms and designing energy-aware AI benchmarks.

## Abbreviations

AI: Artificial Intelligence; CMOS: Complementary Metal–Oxide–Semiconductor; FLOP: Floating–Point Operation; LLM: Large Language Model; MAC: Multiply–Accumulate; PUE: Power–Usage Effectiveness.

## Introduction

### Energy as a first-class property of artificial intelligence

The energy required to train and serve artificial-intelligence systems has become a systems-level research problem rather than a secondary operational detail. Data-centre electricity demand is rising rapidly, and accelerated computing is an

important component of that growth [1]. Within AI research, this development has motivated calls to report computational efficiency alongside predictive performance [2,3], measurements of the power consumed by deployed models [4,5], and full-stack methods for estimating inference energy across model sizes and serving regimes [6,7].

The numerical value assigned to a “query” is not universal. It depends on the model and the number of active parameters, input and output lengths, batching, numerical precision, accelerator utilization, memory and communication traffic, host-system power, idle capacity, and the energy overhead of the facility. Recent full-stack studies illustrate both the usefulness and the limits of a single representative number. A

production measurement of the Gemini serving stack reported a median energy of approximately 0.24 per text prompt, including accelerator, host, idle-capacity, and data-centre overhead [4]. A broader model-based study reported a sub-watt-hour median for large-model inference together with a wide interquartile range across configurations [6]. Such values are informative descriptions of particular workload populations; they are not physical constants.

A different scale enters through Landauer's principle. For an isothermal physical process that reduces the Shannon entropy of information-bearing degrees of freedom by  $\Delta H_{\text{erase}} \frac{1}{2}$  bits, the heat transferred to a bath at temperature  $T$  obeys

$$Q_{\text{bath}} \geq E_L k_B T \ln 2 \Delta H_{\text{erase}}. \quad (1)$$

For the ideal reset of one unbiased bit,  $\Delta H_{\text{erase}} = 1$  and at  $T = 300$ . The principle constrains logical entropy reduction, not an arbitrary count of floating-point operations, multiply-accumulates, instructions, or transistor transitions [8–11]. A computational operation may erase one bit, less than one bit, more than one bit, or no information at all, depending on its logical map, input distribution, physical representation, and implementation.

This qualification is essential. Dividing an observed energy by  $k_B T \ln 2$  gives a useful dimensionless reference ratio,

$$R_L(E; T) = \frac{E}{k_B T \ln 2}, \quad (2)$$

but is not, by itself, the number of erasures performed and does not measure the distance from a realizable lower bound for the workload. In this article, we therefore use the phrase *Landauer-normalized reference ratio* whenever energy is normalized without a process-specific entropy audit. A genuine Landauer lower bound is reserved for cases in which the information-bearing degrees of freedom are defined.

## The conceptual problem

The literature needed to analyse neural-computation energy is distributed across several communities. Stochastic thermodynamics characterizes irreversible transformations, fluctuation relations, and finite-time dissipation [12–14]. Geometric thermodynamics identifies optimal protocols and thermodynamic length [15–17]. Finite-time Landauer results quantify the additional cost of rapidly erasing information under specified dynamics [18,19]. Quantum and strong-coupling extensions determine how coherence, correlations, and system-bath interactions modify those results [20–22]. Thermodynamic-computation theory studies the dependence of dissipation on input distributions, priors, and circuit realization [23–25]. Device and computer-architecture research, by contrast, measures energy in switching events, arithmetic operations, memory transfers, and complete workloads [26,27]. Finally, data-centre studies place chip energy inside the larger serving system [1,4].

These bodies of work use different system boundaries and units. A finite-time erasure inequality may apply to a controlled two-state memory. A device paper may report joules per MAC at a stated precision. An inference study may report joules per token or watt-hours per prompt. A facility study may report power-usage effectiveness. Combining such quantities without first matching the physical process, normalization, and accounting boundary can produce apparently precise but physically invalid decompositions.

The central thesis of this review is therefore narrower and more defensible than the claim that every observed joule can be uniquely assigned to an independent thermodynamic level:

A hierarchy is useful only when it records the assumptions and system boundary of each contribution. Fundamental bounds, implementation losses, and facility multipliers must be compared within a common accounting unit, but they are not generally additive.

## Contributions

The article makes four contributions.

First, it gives a self-contained review of Landauer's principle and its finite-time, quantum, control-theoretic, and distribution-dependent extensions. Throughout, we distinguish a proved inequality from a device-level analogy or order-of-magnitude estimate.

Second, it introduces a seven-level framework, with the Landauer term labelled Level 0 and six subsequent levels:

[leftmargin=2.4em,itemsep=0.25em]

1. logical entropy reduction and the process-specific Landauer bound;
2. finite-time corrections under an explicit stochastic dynamics;
3. quantum, many-body, and strong-coupling corrections;
4. energetic costs of control, measurement, and feedback;
5. distribution-dependent mismatch between a physical implementation and its operating distribution;
6. architectural realization, including arithmetic, memory, communication, conversion, and dataflow;
7. facility-scale power delivery and heat rejection.

Levels 0–4 can yield lower bounds when a physical process and its assumptions are specified. Level 5 is primarily a property of a realized architecture and workload. Level 6 changes the accounting boundary from chip to facility. The aggregation rule formalized in sec:synthesis:formal states the mathematically warranted relationship among compatible lower bounds; it does not assume that the levels sum to the operational energy.

Third, we connect the framework to AI practice through transparent operational accounting. Rather than deriving a

query energy from a universal “bit-equivalent operation,” we express it in terms of measurable quantities such as active FLOPs or MACs, device energy per operation, utilization, memory traffic, output length, host power, and facility overhead. The resulting equations show exactly where uncertainty enters and allow comparison with full-stack measurements [4,6].

Fourth, we use the framework to assess emerging computing paradigms. Reversible, neuromorphic, analogue in-memory, photonic, and thermodynamic computing attack different parts of the stack and therefore require different benchmarks. The comparison identifies a set of open problems that are falsifiable and operationally relevant to AI researchers (Figure 1).

The seven levels distinguish physical constraints and accounting boundaries relevant to computational energy. Level 0 represents logical entropy reduction and its process-specific Landauer bound. Levels 1–4 describe finite-time, quantum and strong-coupling, control, and distribution-dependent contributions that may provide process-specific constraints when their respective physical assumptions are defined. Level 5 represents realized architectural energy associated with arithmetic, memory, communication, conversion, and data movement, whereas Level 6 extends the accounting boundary to facility-scale power delivery and heat rejection. The diagram is conceptual and intentionally unscaled. The levels do not constitute a universal additive decomposition of computational energy, and quantities belonging to different physical processes or system boundaries should not be compared numerically without explicit co-normalization.

## Scope and terminology

The review concerns operational energy during computation. Embodied emissions, water use, manufacturing, and electronic waste are important but outside its principal thermodynamic

boundary. We discuss training only where it clarifies the distinction between logical irreversibility and implementation cost; the quantitative example in sec:overhead:reconstruction concerns inference.

We use *FLOP* for one floating-point Arithmetic operation under the convention adopted by the cited hardware specification. A fused multiply-add is commonly counted as two FLOPs. We use *MAC* only when the source explicitly reports multiply-accumulates. We do not convert either quantity into a number of erased bits without a model of the logical map and physical implementation.

The article is organized as follows. sec:landauer fixes the thermodynamic notation. sec:finite-time, sec: quantum, sec: control review process-specific corrections and control costs. sec: architecture moves to realized computing architectures. sec: overhead establishes a full-stack accounting boundary. sec: synthesis formalizes the hierarchy and develops the open problems, and sec: conclusions summarizes implications for AI research.

## Foundations: Landauer’s principle in modern stochastic thermodynamics

The purpose of this section is to fix notation and to re-derive Landauer’s bound within the framework of modern stochastic thermodynamics. We emphasize at the outset that the modern derivation differs from the heuristic argument originally given by Landauer [10] in a significant respect: Landauer’s bound is not, on this view, an independent principle but a particular asymptote of a much more general inequality that constrains dissipation in arbitrary non-equilibrium transformations. This reframing, established in the works of Esposito and Van den Broeck [28] and developed throughout the literature on stochastic thermodynamics [12–14], provides the foundation for Levels 0 and 1 of Fig. 1. Higher levels introduce additional

### Hierarchical framework for thermodynamic and operational energy accounting

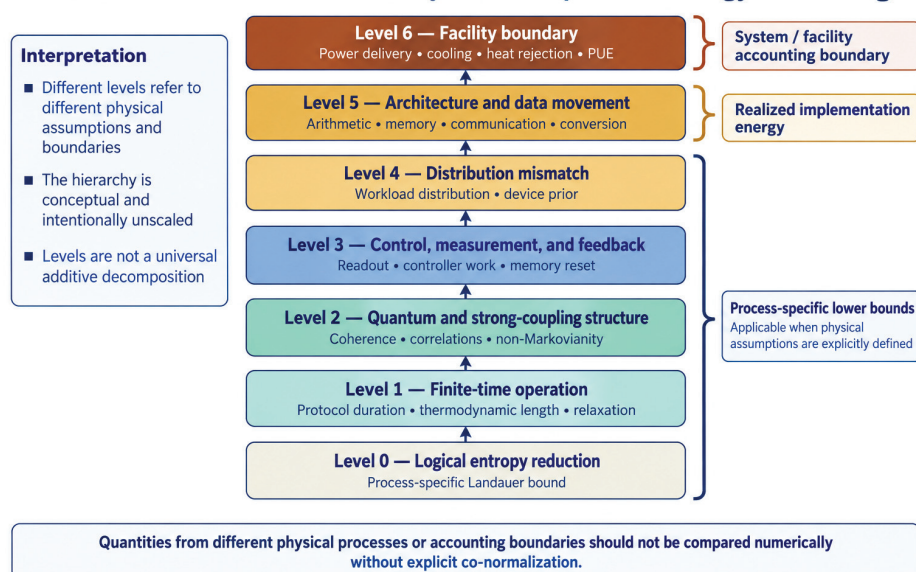


Figure 1: Hierarchical framework for thermodynamic and operational energy accounting.

system boundaries and physical structures rather than arising from one single sequence of relaxed assumptions.

### Information-theoretic preliminaries

We work throughout with classical random variables; the quantum extension is deferred to sec. quantum. Let  $x$  be a random variable taking values in a discrete alphabet with probability distribution  $p(x)$ . Two information measures will recur.

**Definition 2.1 (Shannon entropy)** The Shannon entropy of  $X$ , in units of bits, is

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log_2 p(x). \quad (3)$$

The same quantity expressed in nats is  $H^{\text{nat}}(X) = H(X) \ln 2 = - \sum_{x \in \mathcal{X}} p(x) \ln p(x)$ .

**Units and conversion convention:** Throughout this review, Shannon entropy, mutual information, and relative entropy are dimensionless information-theoretic quantities. When natural logarithms are used, these quantities are expressed in nats; when base-2 logarithms are used, Shannon entropy is expressed in bits. The corresponding physical entropy is obtained by multiplication by Boltzmann's constant,

$$S_{\text{phys}} = k_B H^{\text{nat}} = k_B \ln 2 H, \quad (4)$$

and therefore has units of  $\text{JK}^{-1}$ . Multiplication of a dimensionless information quantity in nats by produces an energy scale in joules. Unless an explicit factor of  $k_B$  is shown, information-theoretic entropies and divergences in this article are dimensionless.

**Definition 2.2 (Relative entropy)** For two distributions  $p, q$  on  $\mathcal{X}$  with  $\text{supp}(p) \subseteq \text{supp}(q)$ , the Kullback-Leibler divergence (relative entropy) is

$$D_{\text{KL}}(p \| q) = \sum_{x \in \mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} \geq 0, \quad (5)$$

with equality if and only if almost everywhere.

The non-negativity of relative entropy is central to stochastic thermodynamics. For a discrete information-bearing variable  $X$ , the entropy associated with uncertainty over its logical states is

$$S_{\text{info}}(X) = k_B H^{\text{nat}}(X) = k_B \ln 2 H(X). \quad (6)$$

This is the informational contribution associated with the chosen coarse-graining, not necessarily the complete thermodynamic entropy of all microscopic degrees of freedom.

**Remark 2.3** The logical variable, its physical representation, and the coarse-graining must be stated. Two descriptions of the same device can assign different Shannon entropies while agreeing on the full thermodynamic state. We use only explicitly identified information-bearing degrees of freedom and treat the remaining microscopic variables through the stochastic dynamics and reservoir [8,9,24].

### Stochastic thermodynamics framework

We consider a classical system whose state evolves in a phase space under stochastic dynamics in contact with a single thermal reservoir at temperature  $T$ . The dynamics are controlled by an external parameter  $\lambda(t)$  (the "protocol"), through which the system Hamiltonian depends explicitly on time. Two classes of dynamics will be relevant in this review.

**Overdamped Langevin dynamics:** For a one-dimensional system with control potential  $V(x; \lambda)$  and mobility  $\mu$ , the overdamped Langevin equation

$$dx = -\mu V_x(x; \lambda(t)) dt + \sqrt{2D} dW_t, \quad D = \mu k_B T, \quad (7)$$

where  $W_t$  is a standard Wiener process and the diffusion coefficient  $D$  obeys the Einstein relation, generates the dynamics of the probability density  $\rho(x, t)$  via the Fokker-Planck equation

$$\rho_t = x[\mu V_x \rho] + D[2\rho x - jx], \quad (8)$$

with probability current  $j(x, t) = -\mu(\partial_x V)\rho - D\partial_x \rho$ . eq:langevin, eq:fokker-planck are the standard setting in which Landauer's bound has been verified experimentally [29,30] and theoretically refined [17,18,31].

**Markovian master equations:** For a system with discrete state space  $\{1, \dots, N\}$  and time-dependent transition rates  $W_{ij}(\lambda(t))$  from state  $j$  to state  $i$ , the master equation

$$\frac{dp_i}{dt} = \sum_{j \neq i} [W_{ij}(\lambda(t)) p_j(t) - W_{ji}(\lambda(t)) p_i(t)] \quad (9)$$

governs the evolution of the occupation probabilities  $p_i(t)$ . Detailed balance with respect to the instantaneous Boltzmann distribution  $p_i^{\text{eq}}(\lambda) \propto \exp[-\beta E_i(\lambda)]$ , where  $\beta = 1/(k_B T)$ , requires

$$\frac{W_{ij}(\lambda)}{W_{ji}(\lambda)} = \exp\{-\beta[E_i(\lambda) - E_j(\lambda)]\}. \quad (10)$$

**Trajestochastic thermodynamics:** A central insight of stochastic thermodynamics [13,14] is that work, heat, and entropy can be defined consistently along individual stochastic trajectories  $\mathbf{x}_{[0,\tau]} \{ \mathbf{x}(t) \}_{t \in [0,\tau]}$ , not merely as ensemble averages. For a trajectory of duration  $\tau$ , the trajectory work is

$$W[\mathbf{x}_{[0,\tau]}] = \int_0^\tau \dot{\lambda}(t) H(\mathbf{x}(t); \lambda(t)) dt, \quad (11)$$

the change in internal energy is  $\Delta U = H(\mathbf{x}(\tau); \lambda(\tau)) - H(\mathbf{x}(0); \lambda(0))$ , and the heat absorbed from the bath is fixed by the first law:

$$Q[\mathbf{x}_{[0,\tau]}] = \Delta U - W[\mathbf{x}_{[0,\tau]}]. \quad (12)$$

The total entropy production along the trajectory combines

the system entropy change with the bath entropy change  $\Delta s_{\text{bath}} = -Q/T$ . Averaging over an ensemble of trajectories generated under a fixed protocol yields the central inequality

$$\langle \Sigma \rangle \geq 0, \quad (13)$$

which is the second law of thermodynamics in its stochastic form. We shall not redevelop here the trajectory-level fluctuation theorems (Crooks, Jarzynski, Seifert) underlying eq. (13); the reader is referred to the canonical reviews of Seifert [13] and Jarzynski [12], and to the pedagogical synthesis of Van den Broeck and Esposito [14].

### Landauer's bound as quasi-static asymptote

We now derive Landauer's bound from the framework just established. The derivation we present follows the strategy of Esposito and Van den Broeck [28], in which the Landauer bound emerges as a special case of a more general inequality relating dissipated work to relative entropy. This derivation is conceptually superior to the heuristic original [10] because it makes explicit the two conditions under which the bound is saturated, both of which will be relaxed in subsequent sections.

**Setup:** Consider a system with a control parameter that takes the value at and at  $t = \tau$ . Let denote the actual non-equilibrium probability density at time  $t$ , and let denote the instantaneous equilibrium distribution for control value  $\lambda$ . We do not assume that or coincide with their respective equilibrium distributions.

**Non-equilibrium free energy:** Define the non-equilibrium free energy associated with the actual distribution  $\rho(\mathbf{x}, t)$  at protocol value  $\lambda(t)$  as

$$\mathcal{F}[\rho; \lambda] \langle H(\mathbf{x}; \lambda) \rangle_{\rho} - k_{\text{B}} T S^{\text{Sh}}[\rho], \quad (14)$$

where is the differential Shannon entropy of  $\rho$  (in nats). This functional has the following key property, valid for any normalized  $\rho$ :

$$\mathcal{F}[\rho; \lambda] - F^{\text{eq}}(\lambda) = k_{\text{B}} T D_{\text{KL}}(\rho \| \rho^{\text{eq}}(\cdot; \lambda)) \geq 0, \quad (15)$$

where is the equilibrium free energy. The decomposition (15) expresses the non-equilibrium free energy as the equilibrium free energy plus a strictly non-negative "information cost" that quantifies the deviation from equilibrium.

**Esposito–Van den Broeck inequality:** For the Markovian dynamics of eq:langevin, eq: master, one can show [28] that the average work performed on the system over the protocol obeys

$$\langle W \rangle \geq \mathcal{F}[\rho_f; \lambda_f] - \mathcal{F}[\rho_i; \lambda_i], \quad (16)$$

where and are the actual initial and final distributions. (Eq. (2) of Ref. [28], in our sign convention) eq:esposito-vdbroeck is more general than the standard second law in two respects: first, it does not assume that or are equilibrium distributions;

second, the right-hand side is the difference of *non-equilibrium* free energies. Equality holds in the quasi-static limit when the system tracks instantaneous equilibrium throughout

$$\rho(\mathbf{x}, t) \rightarrow \rho^{\text{eq}}(\mathbf{x}; \lambda(t)).$$

**Heat convention and the general inequality:** We use one sign convention throughout:  $Q_{\text{sys}} > 0$  denotes heat absorbed by the system and denotes heat delivered to the bath. The first law is therefore

$$\Delta U = W + Q_{\text{sys}}, \quad Q_{\text{bath}} = W - \Delta U. \quad (17)$$

**Work, dissipated work, excess work, and heat:** These quantities should not be identified without specifying the thermodynamic conditions. We define the dissipated work relative to the appropriate non-equilibrium free-energy difference as

$$W_{\text{diss}} = \langle W \rangle - \Delta \mathcal{F},$$

whereas, with the heat convention adopted above,

$$\langle Q_{\text{bath}} \rangle = \langle W \rangle - \Delta U.$$

Therefore, dissipated work and heat delivered to the bath are not generally identical. Their identification requires additional endpoint or cyclic conditions. Likewise, "excess work" in a slow-driving expansion denotes the work above the quasistatic reference process; its identification with dissipated work is approximation-dependent and does not imply a general identification with heat.

Combining eq:esposito-vdbroeck with the definition of non-equilibrium free energy gives

$$\begin{aligned} \langle Q_{\text{bath}} \rangle &= \langle W \rangle - \Delta \langle U \rangle \\ &\geq \Delta \mathcal{F} - \Delta \langle U \rangle = -k_{\text{B}} T \Delta S^{\text{Sh}}, \end{aligned} \quad (18)$$

where is measured in nats. No additional relative-entropy term should be added to eq:heat-bound-general: such terms are already contained in eq:noneq-free-energy-decomp. Expanding the endpoint free energies into equilibrium and relative-entropy components is an alternative representation of the same inequality, not an additional correction.

If the information-bearing degrees of freedom lose  $\Delta H_{\text{erase}} = H_i - H_f \Delta 0$  bits, then

$$\langle Q_{\text{bath}} \rangle \geq k_{\text{B}} T \ln 2 \Delta H_{\text{erase}}. \quad (19)$$

For the reset of one unbiased bit,  $\Delta H_{\text{erase}} = 1$  and

$$\langle Q_{\text{bath}} \rangle \geq k_{\text{B}} T \ln 2. \quad (20)$$

**Theorem 2.4 (Landauer bound for a specified entropy reduction)** Consider information-bearing degrees of freedom

coupled to a heat bath at temperature  $T$ . Assume that the dynamics admit the non-equilibrium free-energy inequality (16), that the informational entropy is the entropy term entering the chosen coarse-grained free energy, and that all other entropy changes are either fixed or included explicitly. If the Shannon entropy decreases from  $H_i$  to  $H_f$  bits, then the average heat delivered to the bath obeys

$$\langle Q_{\text{bath}} \rangle \geq k_B T \ln 2 (H_i - H_f). \quad (21)$$

A reversible isothermal protocol provides the saturation limit under the standard equilibrium assumptions. Finite-time operation, restricted control, strong coupling, correlations, or additional controller degrees of freedom can prevent saturation and introduce further costs.

**Remark 2.5 (What the theorem does not say)** *thm:landauer* does not assign to each instruction, FLOP, MAC, gate evaluation, or transistor transition. Its right-hand side is determined by the entropy reduction of the specified information-bearing degrees of freedom. Applying the theorem to a complete AI workload therefore requires a logical and physical accounting of erasure, not only an operation count.

**Remark 2.6 (Trajectory-level fluctuations)** *The result concerns ensemble averages. Individual stochastic trajectories may dissipate less than the average bound, or may transiently absorb heat from the bath, without violating the fluctuation theorems [29,12,13,29].*

## Information erasure and the bistable potential

The canonical physical realization of single-bit erasure [9,10,29] is the bistable potential, in which a single particle (colloidal, atomic, magnetic) occupies one of two metastable wells separated by an energy barrier. The two wells encode the logical states  $^1$ ; the bit is erased by a protocol that lowers the barrier and tilts the potential toward the target well, returning the system to a configuration where the particle has a high probability of being in the target well regardless of its initial location.

A useful explicit form for theoretical analysis is

$$V(x; \lambda) = V_0 \left[ \left( \frac{x}{x_0} \right)^2 - 1 \right]^2 + f(\lambda) \frac{x}{x_0}, \quad (22)$$

where  $V_0$  is the barrier height in the symmetric configuration,  $x_0$  sets the well separation, and  $f(\lambda)$  is the time-dependent tilting force that drives erasure. The protocol typically combines a phase of barrier lowering  $V_0 \rightarrow V_0 - \Delta V(\lambda)$  —with a tilt that biases the dynamics toward the target well. For sufficiently slow protocols and sufficiently low residual error (probability of remaining in the source well), the dissipated heat saturates in the limit  $\tau \rightarrow \infty$  [29,30] (Figure 2).

Three schematic snapshots of the canonical erasure

protocol implementing the mapping  $H = 1\text{bit} \rightarrow H \approx 0\text{bit}$ . (a) Initial state  $t = 0$ : a symmetric bistable potential with two metastable wells representing logical states “0” and “1”. The corresponding distribution  $\rho(x, 0)$  (orange, offset below) has approximately equal weight in the two wells, giving a bit. (b) Mid-protocol  $t = \tau / 2$ : the barrier has been lowered and a tilt applied toward the target state. The probability distribution progressively concentrates in the target well, so that the logical entropy lies in the intermediate range bit. No particular numerical value of the intermediate entropy is implied by the schematic. (c) Final state  $t = \tau$ : the distribution is strongly localized in the target well, corresponding ideally to the bit. In the quasistatic, low-error limit, the mean heat dissipated to the bath approaches the Landauer value for complete one-bit erasure (thm:landauer). The potential profiles and probability distributions are illustrative and are not numerical solutions of a specified protocol.

fig: bistable illustrates the canonical erasure protocol in three representative snapshots: the initial symmetric configuration, an intermediate state in which both the barrier and a tilt force are applied, and the final single-well configuration in which the bit has been reset.

The bistable potential will recur throughout this review as the canonical theoretical and experimental testbed: in sec:finite-time for the analysis of finite-time corrections, in sec:control:cd for the analysis of counterdiabatic protocols, and in sec:architecture:cmos as the conceptual model underlying CMOS gate dynamics.

## Generalizations: Feedback, multiple reservoirs, partial erasure

For completeness, we record three generalizations of thm:landauer that will be invoked at later points in the review.

**Feedback control:** When the protocol depends on a measurement outcome performed on the system at some intermediate time, the bound (21) is modified by a term proportional to the mutual information acquired by the measurement [32,33]. The generalized inequality

$$\langle W \rangle \geq \Delta F - k_B T I(X; M) \quad (23)$$

permits work extraction below the Landauer cost when feedback control provides effective “negative entropy” from measurement, but only at the price of accounting for the thermodynamic cost of the measurement itself. We return to this point in sec:control:feedback.

**Partial erasure:** eq:landauer-thm continues to apply for protocols that reduce by less than one bit; the right-hand side is simply scaled by  $|\Delta H|$  [34]. This generalization is relevant for neural-network operations such as soft thresholding and probabilistic readout, which we discuss in sec:architecture:neural.

**Multiple reservoirs:** For systems coupled to multiple thermal reservoirs at different temperatures, the bound (21)

## Information erasure protocol on a bistable potential

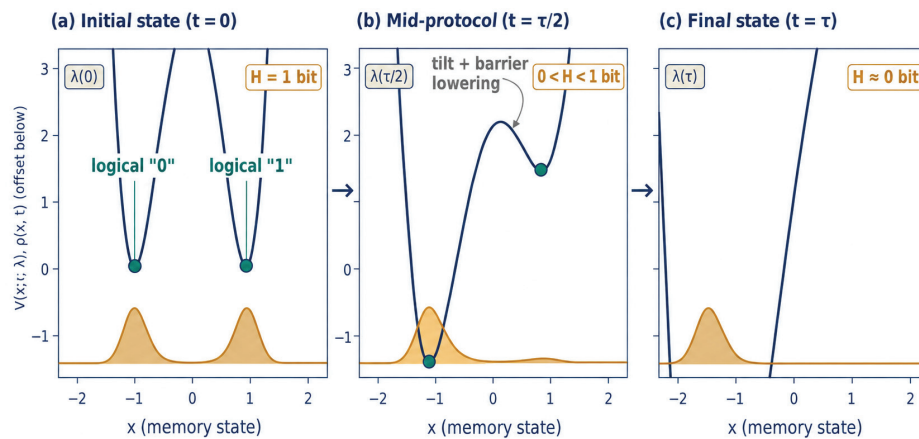


Figure 2: Information erasure protocol on a bistable potential (schematic representation).

generalizes to a sum of reservoir-specific contributions weighted by the corresponding heat flows [13,28]. We do not develop this case in detail, but it is conceptually relevant to the analysis of heat rejection in Sec. Overhead.

### Experimental verifications: an interpretive synthesis

The experimental verification of Landauer's bound has progressed in roughly three waves over the past decade and a half. We summarize them here not in chronological exhaustiveness—for which the reader is referred to the surveys of Bormashenko [11,12]—but in a way that anticipates the theoretical structure of the hierarchy to be developed in subsequent sections.

**First wave: classical single-particle confirmations:** The first direct experimental confirmation of Landauer's bound was reported by Bérut et al. [29], who realized a single-bit memory using a colloidal particle ( $\sim 2$  silica beads) trapped in a double-well optical potential and immersed in water at room temperature. The dissipated heat in the long-protocol limit was observed to saturate at the predicted value  $k_B T \ln 2$ , with finite-time deviations consistent with the scaling predicted by stochastic thermodynamics. The companion experiment of Jun, Gavrilov, and Bechhoefer [35][30] achieved substantially higher precision using virtual double-well potentials implemented through feedback traps, and provided the first quantitative test of the finite-time correction.

The interpretive significance of these experiments goes beyond the confirmation of the asymptotic bound. Both experiments simultaneously verified the form of the leading finite-time correction predicted by the optimal-protocol analysis of Aurell et al. [15,16], foreshadowing the “tight” bound of sec:finite-time [18]. In our hierarchical framework, the first wave of experiments thus simultaneously probed levels (i) and (ii) of the hierarchy.

**Second wave: solid-state and atomic platforms:** The colloidal experiments were followed by tests on more device-

relevant platforms. Hong et al. [31] measured the dissipation in single-bit operations on nanomagnetic memory bits, verifying the Landauer bound to within experimental uncertainty in a context where the information-bearing degree of freedom is a collective magnetization rather than a single particle position. Single-electron [32] and single-atom [33] realizations have similarly extended the empirical scope of thm:landauer into regimes of practical relevance for solid-state memories. The convergence of these results across radically different physical substrates—colloidal, magnetic, electronic, atomic—is itself an experimental confirmation of Landauer's claim that the bound is independent of the device implementation [34].

**Third wave: quantum many-body regime:** The most recent and conceptually distinctive experimental result is the work of Aïme, Tajik et al. [35], which probes Landauer's principle in a genuinely many-body quantum system: a quasi-one-dimensional Bose gas undergoing a global mass quench from a massive to a massless Klein–Gordon model.

The methodological advance is not a calorimetric inference of entropy from dissipated heat. Instead, the experiment employs dynamical tomography of measured field correlations to reconstruct the covariance matrix of the quantum field. From these reconstructed covariance matrices, the relevant information-theoretic quantities, including system entropy, mutual information, and relative-entropy contributions, are evaluated.

This setting also goes beyond the special case of an initially uncorrelated system and thermal environment. For the experimentally considered system–environment partitions, initial correlations and deviations of the environment from the reference thermal state must be included explicitly in the generalized entropy-production balance. We return to this generalized formulation and its experimental interpretation in Sec. 4.4 (Figure 3).

Selected experimental milestones are organized chronologically and by physical platform, from colloidal

### Experimental platforms investigating Landauer's principle (2012–2025)

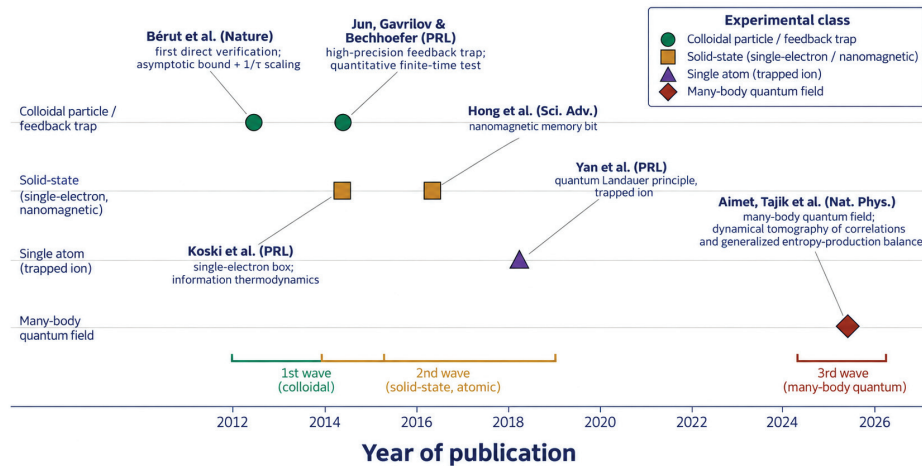


Figure 3: Experimental platforms investigating Landauer's principle (2012–2025).

and solid-state realizations to single-atom and many-body quantum systems. The three brackets indicate broad experimental waves and are intended only as a conceptual classification. Vertical placement denotes experimental class rather than the magnitude of dissipation. The experiments employ different physical systems, observables, system boundaries, and formulations of Landauer-type thermodynamic relations and therefore should not be interpreted as a quantitative comparison on a common energy scale. In particular, the many-body experiment of Aïmet, Tajik et al. [35] evaluates a generalized entropy-production balance involving energetic, mutual-information, and relative-entropy contributions reconstructed from dynamical tomography of field correlations.

**Synthesis:** Fig. experiments summarizes the experimental landscape developed above. Across these platforms, the reported results are broadly consistent with the thermodynamic predictions applicable to their respective protocols and operating regimes. Quasistatic erasure experiments approach the Landauer scale, whereas finite-time, feedback-controlled, quantum, and many-body implementations probe different corrections and generalizations of the principle. Taken together, the experimental literature supports the central conceptual claim of this review: Landauer's bound is a process-specific asymptotic limit embedded within a broader hierarchy of thermodynamic constraints, rather than a universal energy cost assigned to every computational operation.

#### Notation summary and accounting conventions

The following distinctions are used throughout:

[leftmargin=2.2em,itemsep  $Q_{\text{sys}}$ : heat absorbed by the system;  $Q_{\text{bath}} = -Q_{\text{sys}}$ : heat delivered to the bath.

$H(X)$ : Shannon entropy in bits;  $S^{\text{Sh}} = H \ln 2$  is the corresponding dimensionless entropy in nats;  $S_{\text{phys}} = k_{\text{B}} S^{\text{Sh}}$  is the corresponding physical entropy in  $\text{JK}^{-1}$ .

$\Delta H_{\text{erase}} = H_i - H_f \geq 0$ : logical entropy reduction for a specified physical process.

$E_{\text{L}} = k_{\text{B}} T \ln 2 \Delta H_{\text{erase}}$ : the process-specific Landauer lower bound.

$R_{\text{L}}(E; T) = E / (k_{\text{B}} T \ln 2)$ : a dimensionless reference normalization. Unless has been identified,  $R_{\text{L}}$  is not a count of erasures and is not a workload-specific lower bound.

$W_{\text{diss}} = \langle W \rangle - \Delta \mathcal{F} \geq 0$ : dissipated work relative to the non-equilibrium free-energy difference.

$\tau$ : protocol duration;  $\lambda(t)$ : externally controlled protocol.

For a one-bit erasure, it is useful to define the finite-time excess

$$Q_{\text{excess}}(\tau) \langle Q_{\text{bath}} \rangle_{\tau} - k_{\text{B}} T \ln 2, \quad (24)$$

but this definition is not transferred automatically to arbitrary computer operations. In Sec. Architecture and sec: overhead we therefore switch to operational metrics such as joules per FLOP, joules per token, and watt-hours per prompt.

#### Finite-Time corrections in classical stochastic thermodynamics

For the one-bit erasure process of thm:landauer, finite-duration operation generally produces a positive excess as defined in eq:finite-time-excess. This section reviews two complementary descriptions: the slow-driving geometry of thermodynamic length [15–17] and the activity-based tight bound of Lee, Lee, Kwon, and Park [18]. Both results apply to specified stochastic dynamics. They should not be converted into a numerical correction for a processor without a device model.

## The leading-order correction: $1/\tau$ scaling and thermodynamic geometry

The earliest characterization of at finite operation time emerged not from the analysis of erasure protocols specifically, but from the more general question of how to minimize dissipation along a prescribed transformation between two control values  $\lambda_f$ . The answer, developed independently by Schmiedl and Seifert [37], by Aurell and collaborators [15,16], and by Sivak and Crooks [17], takes the form of a geometric structure on the space of control parameters, in which optimal protocols are geodesics and the minimum dissipation is determined by the geodesic length.

**Excess work and the friction tensor:** Following the linear-response analysis of Sivak and Crooks [17], consider a system whose Hamiltonian depends on a vector of control parameters  $\lambda = (\lambda^1, \dots, \lambda^K)$ . Within the slow-driving, near-equilibrium linear-response approximation, in which the control parameters vary slowly compared with the relevant relaxation times, the leading excess power above the quasistatic reference process takes the quadratic form

$$P_{\text{ex}}(t) = \dot{\lambda}^i(t) \zeta_{ij}(\lambda(t)) \dot{\lambda}^j(t), \quad (25)$$

where Einstein's summation convention applies, and the friction tensor  $\zeta_{ij}(\lambda)$  is given by the time integral of the equilibrium covariance of the conjugate forces,

$$\zeta_{ij}(\lambda) = \beta \int_0^\infty dt \langle \delta f_i(0) \delta f_j(t) \rangle_\lambda, \quad f_i = H \lambda^i, \quad (26)$$

with and the average taken in the equilibrium ensemble at fixed  $\lambda$ . This is Eqs. (7)–(8) of Ref. [17], in our notation.

The friction tensor is positive semidefinite by construction and induces a Riemannian metric on the manifold of control parameters. Integrating eq. excess-power over the protocol duration yields the leading excess work. Within the slow-driving linear-response approximation, this quantity represents the leading dissipative correction relative to the quasistatic reference process: (Figure 4).

$$\langle W_{\text{diss}} \rangle = \int_0^\tau P_{\text{ex}}(t) dt = \int_0^\tau \dot{\lambda}^i \zeta_{ij} \dot{\lambda}^j dt. \quad (27)$$

**(a)** Conceptual illustration of two protocols connecting the same initial and final control points on a parameter manifold with spatially varying thermodynamic friction. A direct path may traverse a region of higher friction, whereas the minimum-dissipation geodesic associated with a specified friction tensor can follow a different trajectory. **(b)** Schematic cumulative excess-work profiles illustrating that protocols traversed at different metric speeds can accumulate different amounts of dissipation. The background friction landscape, trajectories, and cumulative-work curves shown here are illustrative and are not obtained from a specified physical friction tensor or from a numerical optimization. No quantitative dissipation ratio or device-specific prediction should be inferred from the

figure. The rigorous result discussed in the text is the slow-driving bound under the assumptions stated in Sec. 3.1.

**Optimal protocols are geodesics:** The minimization of eq.  $\langle W_{\text{diss}} \rangle$ , integrated over all protocols connecting and in time, is, by the Cauchy–Schwarz inequality applied to the metric defined by  $\zeta_{ij}$ , a geodesic problem on the Riemannian manifold  $(\mathcal{M}, \zeta)$  [17]. The minimum dissipated work is bounded below by

$$\langle W_{\text{diss}} \rangle \tau \geq \mathcal{L}^2, \quad \mathcal{L} = \int_{\lambda_i}^{\lambda_f} \sqrt{\zeta_{ij} d\lambda^i d\lambda^j}, \quad (28)$$

where is the *thermodynamic length* of the protocol. Equality is achieved when the protocol traverses the geodesic at constant rate, accumulating excess work at a constant power [17]. Fig. 1 provides a schematic illustration of this geometric structure. For a specified friction tensor, the minimum-dissipation protocol is the corresponding geodesic traversed at constant metric speed. A direct path through control space need not minimize thermodynamic length, because different regions of the control manifold may carry different friction. The figure is intentionally qualitative and does not represent a numerical optimization or a device-specific friction landscape.

**Application to single-bit erasure:** Specializing eq. (thermo-length-bound to the bistable potential of eq. (bistable, with the control parameter representing barrier height and tilt, gives a quantitative form of the leading correction. For an overdamped Langevin system with mobility  $\mu$  executing the canonical erasure protocol, the dissipated work obeys

$$\langle W \rangle - k_B T \ln 2 \geq \frac{C(\mathcal{P})}{\tau} \quad \text{as } \tau \rightarrow \infty, \quad (29)$$

where is a constant determined by the geometry of the protocol and is bounded below by for the geodesic protocol. [16,17] The scaling of eq. 1-over-tau is the classical signature of the leading finite-time correction, and was observed experimentally already in the first generation of Landauer verifications [29,33].

**Connection to overdamped optimal transport:** A distinct but related finite-time result arises for overdamped Langevin dynamics when the initial and final probability distributions are prescribed, and the optimization is performed over admissible transport protocols. Under these assumptions, Aurell et al. [15,16] showed that

The minimum dissipation for a transformation  $\rho_i \rightarrow \rho_f$  completed in time  $\tau$  can be written as

$$\langle W_{\text{diss}} \rangle_{\text{min}} = \frac{\mathcal{W}_2(\rho_i, \rho_f)^2}{\mu \tau}, \quad (30)$$

where is the Wasserstein distance between the initial and final probability distributions and is the mobility. The corresponding optimal transport is described by displacement interpolation in probability-distribution space [38].

This overdamped optimal-transport result should be

## Thermodynamic geometry: geodesic protocols minimize dissipation

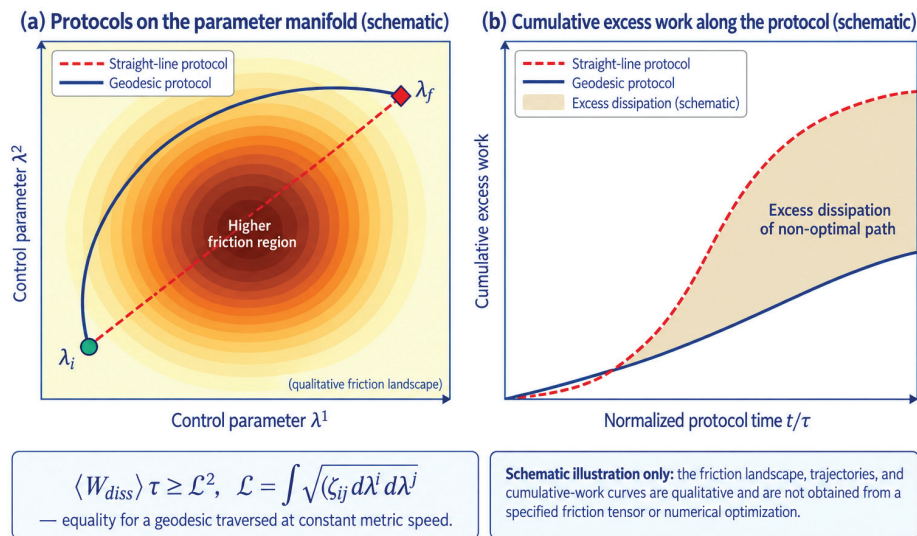


Figure 4: Thermodynamic geometry of finite-time control (schematic).

distinguished from the Sivak–Crooks thermodynamic-length construction. The latter is a slow-driving, near-equilibrium linear-response result formulated on the manifold of externally controlled parameters through the friction tensor, whereas the optimal-transport result is formulated for overdamped probability dynamics between prescribed endpoint distributions. The two geometric descriptions are related under appropriate dynamical assumptions, but they should not be identified without an explicit mapping between the Wasserstein metric and the control-space friction metric [39].

**Status of the geometric framework:** The thermodynamic-geometry approach has matured into a complete and predictively useful theory of leading-order finite-time dissipation [17,40,41,]. For protocols that do not stray far from quasi-equilibrium, the predictions of Eq.  $W_{diss}$ -integrated, eq:wasserstein-bound are in quantitative agreement with experiment, including the precision tests of Jun et al. [30] and the more recent quantum-dot realization of Scandi et al. [42], which directly verified the thermodynamic-length scaling in a mesoscopic erasure protocol.

### The tight bound in the highly irreversible regime

The thermodynamic-length result is a slow-driving expansion. A complementary bound, valid for Markov jump processes beyond the linear-response regime, was derived by Lee, Lee, Kwon, and Park [18]. For a discrete process, let

$$\frac{1}{2} \sum_n |p_n(\tau) - p_n(0)| \quad (31)$$

be the total-variation distance between the endpoint distributions, and let

$$A_{\text{tot}} \int_0^\tau \sum_{n \neq m} R_{nm}(t) p_m(t) \quad (32)$$

be the time-integrated dynamical activity. The ratio measures the average net distributional change per jump.

For perfect erasure of an initially unbiased two-state memory, the symmetric-relative-entropy result of Lee et al. gives

$$\frac{\langle Q_{\text{bath}} \rangle}{k_B T} \geq \ln 2 + \tanh^{-1}(v). \quad (33)$$

The factor is explicit in eq. (LLKP-bound; the source uses units with  $k_B = 1$ . The finite-time excess is therefore

$$\frac{Q_{\text{excess}}}{k_B T} \geq \tanh^{-1}(v). \quad (34)$$

For  $v \ll 1$ ,

$$\tanh^{-1}(v) = v + \mathcal{O}(v^3), \quad v = \frac{1}{2\tau \langle A \rangle_\tau} \quad \text{for perfect one-bit erasure,} \quad (35)$$

where  $\langle A \rangle_\tau = A_{\text{tot}} / \tau$ . In the opposite limit,

$$\tanh^{-1}(v) = -\frac{1}{2} \ln(1-v) + \frac{1}{2} \ln 2 + o(1), \quad (36)$$

so the correction diverges as the process approaches the speed-limit boundary.

**Theorem 3.1 (Tight finite-time erasure bound)** Under the Markov jump-process assumptions of [18], perfect erasure of an initially unbiased two-state memory obeys eq. LLKP-bound. The result is tight for the optimal dynamics identified in that work.

**Remark 3.2 (Restricted domain of application)** *thm: LLKP constrains a specified two-state erasure dynamics. Applying it to CMOS, a neural-network layer, or an LLM query requires a coarse-grained Markov model from which it can be calculated or measured. Clock period alone does not determine it  $V$ , and a numerical value should not be assigned without such a model.*

## Related developments

The geometric and tight-bound frameworks reviewed above are the two principal threads of recent work on finite-time Landauer bounds, but they do not exhaust the literature. We briefly note three further developments that will be relevant in subsequent sections.

**Quantum extensions:** Van Vu and Saito [22] extended the finite-time Landauer bound to the quantum regime by establishing a quantum classical speed limit; their result, which incorporates quantum coherence as an additional source of dissipation, will be developed in sec:quantum. Earlier work by Zhen, Egloff, Modi, and Dahlsten [43] established the universality of the  $1/\tau$  scaling across a wide class of quantum processes, providing the conceptual bridge between the classical and quantum geometric frameworks.

**Bit-erasure with finite error tolerance:** Practical computing operates not at vanishing erasure error but at a finite, controlled error rate  $\mathcal{E}$ . Ma, Chen, Sun, and Dong [44] derived the minimum energetic cost of bit erasure as a function of erasure error  $\mathcal{E}$ , finding that even small finite errors can substantially relax the cost of fast erasure. The trade-off between energy cost and erasure error—between dissipation and computational fidelity—is a structural feature of the hierarchy and will reappear prominently in our discussion of architectural cost (sec:architecture:neural).

**Generalized Landauer bounds:** Several authors have proposed generalizations of the Landauer bound that extend its scope to settings beyond simple erasure: absolutely irreversible dynamics [45], multi-partite systems with structured correlations [46], and finite-size baths with explicit quantum corrections [47]. These generalizations are not in tension with the framework of this section; they refine the bound in specific physical scenarios that depart from the canonical single-bit single-bath setting.

## Implications for computing hardware

Finite-time results provide an important warning against treating it as an engineering target that can be approached independently of speed and reliability. They do not, however, determine the finite-time cost of a modern logic gate from its clock rate alone. The relevant relaxation time, controlled state space, transition activity, error criterion, and thermodynamic boundary must all be specified.

For conventional digital hardware, the experimentally observed energy per operation is many orders of magnitude larger than  $k_B T \ln 2$ . In that regime, a finite-time erasure

correction of a few would be numerically hidden inside device and architectural costs. This observation should not be interpreted as a quantitative estimate of Level 1 for a GPU. It means only that finite-time bounds become operationally binding after higher-level implementation losses have been reduced sufficiently.

The appropriate research programme is therefore empirical and model-specific:

1. identify the information-bearing degrees of freedom of a device;

2. infer or measure a stochastic transition model;

estimate  $\ell$ ,  $A_{\text{tot}}$ , and the erasure error;

compare the measured heat with both eq:thermo-length-bound, eq:LLKP-bound; and

report which energy used by drivers, clocks, sensing, and control lies outside the modelled subsystem.

Without these steps, finite-time theory should be presented as a process-specific lower-bound framework rather than as a numerical contribution to the energy of an AI workload.

## Quantum and strong-coupling corrections

The classical finite-time framework reviewed in sec:finite-time rests on two structural assumptions that are increasingly difficult to maintain as one descends to the device scales relevant for modern computation: that the information-bearing degree of freedom can be treated classically, and that its coupling to the surrounding bath is weak. At the dimensions and temperatures characteristic of contemporary nanoscale devices—transistors of 23 gate length, single-electron transistors, superconducting qubits, nanomagnetic memories operating in the few- $k_B T$  regime—both assumptions become problematic. Quantum coherence enters non-trivially into the dynamics; system-bath coupling can be of the same order as internal energy scales; and the Markovian approximation, which underpins the classical bound, fails generically.

This section reviews the corrections to Landauer's bound that arise from relaxing these assumptions. We organize the discussion around three contemporary results that, taken together, define the current state of the art at this level of the hierarchy: the quantum finite-time bound of Van Vu and Saito [22], which identifies quantum coherence as an additional source of dissipation within the weak-coupling Lindblad framework; the geometric framework of Rolandi and Perarnau-Llobet [21] for finite-time erasure beyond weak coupling, where strong coupling and non-Markovianity are treated exactly for a fermionic mode; and the recent many-body verification of Aïme, Tajik et al. [35], which provides the first experimental access to Landauer's principle in a regime where the system is genuinely a quantum field. We close with a discussion of the implications for the energetic gap and a comparative positioning against the recent review of Chattopadhyay et al. [48] on the same topic.

## Quantum stochastic thermodynamics: framework and conventions

A complete development of quantum stochastic thermodynamics is beyond the scope of this review; the canonical references are the topical survey of Goold et al. [20], the introduction of Vinjanampathy and Anders [49], the systematic treatment in the volume edited by Binder et al. [50], and the recent review of irreversible entropy production by Landi and Paternostro [51]. We record here only the elements needed for our purposes, with explicit choices of notation and convention.

**Open system in weak coupling:** For a quantum system  $S$  with Hamiltonian  $H_S(t)$ , weakly coupled to a thermal bath  $B$  at inverse temperature  $\beta$ , the reduced density matrix  $\rho_S(t)$  evolves under a Lindblad master equation,

$$\frac{d\rho_S}{dt} = -\frac{i}{\hbar}[H_S(t), \rho_S] + \mathcal{D}_t[\rho_S], \quad (37)$$

where  $\mathcal{D}_t$  is the dissipator constructed from the bath correlation functions in the Born–Markov–secular approximation [52]. The instantaneous equilibrium state is the Gibbs state  $\rho_S^{\text{eq}}(t) = e^{-\beta H_S(t)} / Z(t)$ , which  $\mathcal{D}_t$  annihilates by construction.

**Work, heat, and entropy production:** For an open quantum system in the weak-coupling regime, the average work performed on the system between  $t=0$  and  $t=\tau$  and the average heat absorbed from the bath are defined by

$$\langle W \rangle = \int_0^\tau dt \text{Tr} \left[ \rho_S(t) \frac{dH_S(t)}{dt} \right], \quad (38)$$

$$\langle Q \rangle = \int_0^\tau dt \text{Tr} [\dot{\rho}_S(t) H_S(t)], \quad (39)$$

respectively, satisfying the first law  $\Delta \langle H_S \rangle = \langle W \rangle + \langle Q \rangle$ . The von Neumann entropy of the system is  $S_{\text{vN}}[\rho_S] = -\text{Tr}[\rho_S \ln \rho_S]$  (in nats), related to the thermodynamic entropy by the same factor as in the classical case.

The total entropy production along the protocol takes the Spohn–Lindblad form

$$\langle \Sigma \rangle = \Delta S_{\text{vN}}[\rho_S] - \beta \langle Q \rangle \geq 0, \quad (40)$$

which is the quantum analogue of the second law in stochastic thermodynamics [51,53]. Our sign convention (heat *absorbed* from the bath is positive) follows Ref. [51].

**Notation caveat:** We emphasize that the definitions (38)–(39) are valid only in the weak-coupling regime where the energy of the system is well-defined independently of the bath. In the strong-coupling regime to be considered in sec:quantum:strong-coupling, this decomposition becomes

ambiguous and a careful choice must be made, which we discuss explicitly there. Different choices yield different “works” and “heats,” and any quantitative comparison with the classical bounds of sec:finite-time requires that this choice be made consistently [54,55].

## Finite-time quantum Landauer bounds at weak coupling

For information encoded in an open quantum system governed by a Markovian Lindblad equation, Van Vu and Saito derived finite-time lower bounds that hold for arbitrary operation time and control protocol [56]. In the notation of this review their result can be represented as

$$\beta \langle Q_{\text{bath}} \rangle \geq \Delta S_{\text{erase}} + \mathcal{B}_\tau, \quad \mathcal{B}_\tau \geq 0, \quad (41)$$

where  $\Delta S_{\text{erase}}$  is the entropy reduction in nats and  $\mathcal{B}_\tau$  is the finite-time correction specified by the dynamical quantities in the original theorem. It scales inversely with the operation time in the relevant regime.

The same work derives a second lower bound that relates additional heat to coherence generated in the instantaneous energy eigenbasis. Schematically,

$$\beta \langle Q_{\text{bath}} \rangle \geq \Delta S_{\text{erase}} + \mathcal{B}_{\text{coh}}, \quad \mathcal{B}_{\text{coh}} \geq 0. \quad (42)$$

Equations (41) and (42) are lower bounds on the same heat. They should not be added unless a theorem establishes joint additivity for the chosen dynamics. Their common implication is that rapid erasure and coherence generation can each obstruct saturation of the quasistatic Landauer limit.

Neither  $\mathcal{B}_\tau$  nor  $\mathcal{B}_{\text{coh}}$  is a universal correction coefficient. Their numerical values depend on the system Hamiltonian, the Lindblad generator, bath temperature, the controlled protocol, operation time, and, for coherence-related contributions, the relevant instantaneous energy basis. These terms must therefore be regarded as model-dependent quantum bounds rather than universal additive corrections to Landauer’s principle.

**Remark 4.1 (Scope)** *These results apply to the specified open-quantum-system model. They hold beyond the slow-driving approximation, but their numerical evaluation requires the Lindblad generator and the controlled quantum trajectory. They do not yield a generic correction for a classical processor.*

## Strong coupling and non-Markovianity: the Rolandi–Pernau-Llobet framework

The Lindblad framework of the previous subsection is the workhorse of quantum thermodynamics, but it relies on the Born–Markov approximation that is generically invalid at strong system–bath coupling. For a fermionic mode whose tunnel coupling to a fermionic reservoir is of the same order as  $k_B T$ , neither the Markovian master equation nor the standard splitting into “system” and “bath” contributions remains

adequate. The benchmark result in this regime is the work of Rolandi and Perarnau-Llobet [57], who treat a single fermionic mode coupled to a continuum reservoir exactly, without Born–Markov assumptions.

**Setup:** Consider a single fermionic mode (the “bit”) of energy  $\varepsilon(t)$  coupled to a structureless fermionic reservoir at inverse temperature  $\beta$ , with tunnel-coupling rate  $\Gamma(t)$  that is itself a control parameter. The Hamiltonian decomposes as  $H = H_S(t) + H_B + H_{SB}(t)$ , where

$$H_S(t) = \varepsilon(t)c^\dagger c, \quad H_{SB}(t) = \sum_k [V_k(t)c^\dagger b_k + \text{h.c.}], \quad (43)$$

and the coupling spectral density encodes  $\Gamma(t)$ . Both  $\varepsilon(t)$  and  $\Gamma(t)$  are taken as control parameters [57].

**Choice of work definition:** At strong coupling, several inequivalent definitions of work and heat exist in the literature [54,55]. Rolandi and Perarnau-Llobet adopt the convention in which the work is associated with changes in the full Hamiltonian (system plus interaction),

$$\langle W \rangle = \int_0^\tau dt \langle \dot{H}(t) \rangle, \quad (44)$$

and the heat is the energy flow through the system–bath interface defined consistently with the strong-coupling free-energy decomposition of [58] (cf. Sec. III of Ref. [57]). This choice ensures that the second law holds in the form  $\langle W \rangle \geq \Delta \mathcal{F}$ , with  $\mathcal{F}$  the appropriately defined non-equilibrium free energy of the system.

**Slow-driving expansion:** In the slow-driving regime (small  $\dot{\varepsilon}, \dot{\Gamma}$ ), the dissipated work admits a quadratic expansion in protocol velocity,

$$\langle W_{\text{diss}} \rangle = \int_0^\tau dt \dot{\lambda}^\top g(\lambda(t)) \dot{\lambda} + \mathcal{O}(\dot{\lambda}^3), \quad (45)$$

where  $\lambda = (\varepsilon, \Gamma)^\top$  and  $g(\lambda)$  is the strong-coupling thermodynamic metric, computed analytically by Rolandi and Perarnau-Llobet using non-equilibrium Green’s function techniques [57]. The metric is positive semidefinite and reduces to the Sivak–Crooks friction tensor of eq:friction-tensor in the appropriate weak-coupling limit.

**The strong-coupling Landauer bound:** The minimum dissipated work over all protocols of duration  $\tau$  implementing a fixed initial-to-final transformation in  $(\varepsilon, \Gamma)$ -space is, by the same geometric argument as eq:thermo-length-bound, bounded below by

$$\langle W_{\text{diss}} \rangle \tau \geq \mathcal{L}_g^2, \quad \mathcal{L}_g = \int_{\lambda_i}^{\lambda_f} \sqrt{g_{ij} d\lambda^i d\lambda^j}, \quad (46)$$

where is the geodesic length in the strong-coupling metric. The optimal protocol is a geodesic with constant excess power.

**Key qualitative finding:** The quantitative analysis of [57] yields a result that is conceptually important for the hierarchy of this review: *strong system–bath coupling is required to saturate the finite-time bound*. Specifically, in the weak-coupling limit

$\Gamma \ll k_B T / \hbar$ , the relaxation time diverges and the geodesic length in Eq. (RP-bound grows correspondingly, increasing the minimum dissipation at fixed  $\tau$ . The optimal erasure protocol exploits a non-trivial trajectory in the plane, modulating the coupling strength to track the geodesic. This means that strong-coupling effects can in principle reduce dissipation, provided they are exploited via an optimal protocol.

**Remark 4.2 (Implication for nanoscale devices)** *The Rolandi–Perarnau-Llobet result inverts a common assumption in device design, where strong system–bath coupling is treated as parasitic. At fixed operation time, dissipation is minimized by tunable coupling that modulates between strong and weak as required by the optimal trajectory. This observation has direct implications for the design of next-generation memory devices operating in the few- $k_B T$  regime, where engineering of the coupling strength is increasingly feasible [59]. The connection between this microscopic control freedom and the architectural cost analysed in sec:architecture is, to our knowledge, not yet developed in the literature.*

**Quadratic convergence via interacting environments:** A complementary and more recent result, due to Lipka-Bartosik and Perarnau-Llobet [60], shows that engineered interactions within the bath can qualitatively change the convergence toward Landauer’s bound. Concretely, for erasure in contact with a finite reservoir of constituents, the residual entropy production of any protocol employing a non-interacting reservoir decays at most linearly  $n$ ; when the reservoir constituents interact and are prepared near a critical point, an explicit protocol achieves the quadratic scaling.

$$\Sigma_{\text{min}} \propto \frac{1}{n^2}, \quad (47)$$

which is, in fact, the best possible [60]. The qualitative significance is twofold: it demonstrates that the generic linear convergence to the Landauer bound is not universal but reflects properties of the bath, and it identifies engineered, interacting reservoirs as a genuine thermodynamic resource that relaxes the resource–dissipation trade-off of erasure.

## Many-body verification: the Aimet–Tajik experiment

The theoretical developments above raise a natural question: can Landauer’s principle be probed in regimes where the system itself cannot be reduced to a single mode, but is genuinely a quantum field? This question has now been answered experimentally by Aimet, Tajik et al. [35], in a result of methodological as well as physical significance.

**Experimental setup:** The experimental platform is a quantum field simulator based on ultracold atoms trapped in a quasi-one-dimensional geometry. Two tunnel-coupled condensates, prepared in a thermal state of the massive Klein–Gordon Hamiltonian, are subjected to a *global mass quench*: the tunnel-coupling barrier is rapidly ramped, taking the system from the massive to the massless Klein–Gordon model. The post-quench evolution is reconstructed by dynamical tomographic measurements of the field correlators [35].

**Generalized Landauer principle:** The theoretical framework underlying the experiment relates the thermodynamic and information-theoretic descriptions of a system–environment partition. Let denote the system and its environment, and let be a reference Gibbs state of the environment at inverse temperature  $\beta_E$ . Defining

$$\Delta S = S(\rho_S(t)) - S(\rho_S(0)), \quad (48)$$

together with

$$\Delta I = I(S : E)_t - I(S : E)_0, \quad \Delta D = D(\rho_E(t) \parallel \gamma_E^{\beta_E}) - D(\rho_E(0) \parallel \gamma_E^{\beta_E}), \quad (49)$$

the generalized entropy–production balance can be written as

$$\Delta \Sigma = \beta_E \Delta E_E + \Delta S = \Delta I + \Delta D, \quad (50)$$

where is the change in the environment energy,  $I(S : E)$  is the quantum mutual information between system and environment, and is the quantum relative entropy of the environment state with respect to the reference Gibbs state [35].

The commonly used Landauer form is recovered as a special case when the initial system and environment are uncorrelated, and the environment initially coincides with the reference thermal state. Then, and so that

$$\beta_E \Delta E_E = -\Delta S + I(S : E)_t + D(\rho_E(t) \parallel \gamma_E^{\beta_E}) \geq -\Delta S. \quad (51)$$

This special-case inequality should be distinguished from the more general balance in eq:generalized-landauer. In the experimentally considered partitions, initial system–environment correlations and deviations of the environment from the reference thermal state are retained explicitly through [35].

**Experimental result:** Aïmet, Tajik et al. [35] perform dynamical tomographic reconstruction for a family of system–environment partitions of the many-body quantum field. Measurements of field correlations are used to reconstruct the corresponding covariance matrices, from which the reduced states and the relevant information-theoretic quantities are determined. This procedure allows the experiment to evaluate the different terms entering the generalized entropy–production balance of Eq. generalized-landauer, including changes in system entropy, system–environment mutual information, and the relative entropy of the environment with respect to the chosen thermal reference state. Importantly, the analysis retains the contributions associated with initially correlated partitions and with an environment that need not initially coincide with the reference Gibbs state.

**Methodological significance:** The methodological significance of the experiment therefore lies in the combination of dynamical tomography and information-theoretic thermodynamic accounting, rather than in using dissipated

heat as a standalone estimator of entropy. The experimentally measured field correlations provide access to the covariance matrices of the many-body state, from which entropies, mutual information, and relative-entropy contributions can be reconstructed. Energy and information terms can then be compared within the generalized Landauer balance. Thus, Landauer’s principle serves as a framework for organizing the relation between energetic and informational changes, but the entropy change itself is not inferred from heat alone [35].

## Implications for computation

The quantum results reviewed in this section refine the conditions under which a process-specific Landauer bound can be approached. They do not supply a universal multiplicative correction for CMOS logic or neural inference.

Coherence, strong coupling, non-Markovianity, and many-body correlations alter the thermodynamic state variables and the admissible control protocols. Their importance depends on the physical substrate, bath spectrum, temperature, operation time, and the division between system and environment. A correction calculated for a fermionic mode or an ultracold Bose gas cannot be transferred numerically to a room-temperature GPU without a physical mapping.

Three conclusions are nevertheless robust.

1. Weak-coupling Markovian bounds should not be assumed to remain tight at nanoscale or in deliberately correlated substrates.
2. Strong coupling can be a control resource rather than only a source of loss; the direction and magnitude of the effect are protocol-dependent [57].
3. Many-body information, including correlations and mutual information, must be included when the information-bearing degrees of freedom are collective [35].

Accordingly, Level 2 of the hierarchy is represented in later sections as a family of substrate-specific lower bounds, not as a universal numerical coefficient assigned to an AI query.

## Positioning against existing reviews

Two recent reviews cover material that overlaps with this section, and we make our positioning explicit. Goold et al. [20], published in 2016, covers the foundational quantum-information thermodynamics framework, but predates the finite-time and strong-coupling developments reviewed here. The recent review of Chattopadhyay et al. [48], published in *Reports on Progress in Physics* in 2025, covers Landauer’s principle and the thermodynamics of computation with substantial overlap with our treatment of finite-time and quantum corrections in sec:finite-time, sec: quantum.

The present review is complementary in three respects. First, it extends the discussion to architectural, data-movement, and facility boundaries (sec: architecture,

sec: overhead). Second, it makes the accounting status of each quantity explicit: theorem, process-specific model, reference normalization, estimate, or measurement. Third, it connects the thermodynamic literature to current AI-serving measurements and to hardware–algorithm co-design. We do not claim a unique numerical decomposition of an AI workload; the contribution is a framework for preventing incompatible quantities from being combined.

## The cost of control: Protocol overhead

The bounds developed in sec:finite-time, sec: quantum treat the controlled system in isolation: an external protocol is prescribed, and the dissipation incurred by the system as it follows that protocol is computed. This idealization implicitly assigns zero cost to the implementation of the protocol itself—to the actuators, controllers, measurement devices, and clock distribution networks that realize it in any physical setting. For a quasi-static laboratory experiment, this idealization is harmless. For a computing device executing operations per second, it is not: a substantial fraction of the operational energy budget is consumed not by the information-bearing degrees of freedom themselves, but by the machinery that controls them.

This section reviews thermodynamic costs associated with implementing and operating a controlled computational process, focusing primarily on Levels 3 and 4 of the conceptual hierarchy in Fig. 1. The first is the cost of *shortcuts to thermodynamic computing*: the additional dissipation incurred when one drives a system rapidly along a prescribed trajectory in state space, beyond the natural relaxation dynamics [61]. The second is the cost of *measurement and feedback*: when a protocol depends on real-time information about the system, the thermodynamic accounting must include the cost of acquiring that information [62,63]. The third—and conceptually the most important for the remainder of this review—is the *mismatch cost* of Kolchinsky and Wolpert [23, 24], which quantifies the additional dissipation incurred when the actual distribution over inputs to a process differs from the distribution for which the process is energetically optimized. Mismatch cost is the structural mechanism by which the architectural and algorithmic choices of a computing device propagate, with full thermodynamic force, into its energy budget.

## Shortcuts to adiabaticity and the boundary of control

Rapid driving generally creates non-equilibrium populations and coherences. Counterdiabatic or transitionless driving adds control terms designed to follow a reference adiabatic trajectory at finite speed [64–66]. For a non-degenerate reference Hamiltonian  $H_0(t)$  with instantaneous eigenstates  $|n(t)\rangle$ , one standard form is

$$H_{CD}(t) = i\hbar \sum_n \left[ |\partial_t n(t)\rangle \langle n(t)| - \langle n(t)| \partial_t n(t)\rangle |n(t)\rangle \langle n(t)| \right]. \quad (52)$$

Classical analogues use auxiliary forces or potentials to steer a probability density along a desired path [67].

The phrase “cost of counterdiabatic driving” is not

represented by one universal scalar. Different studies quantify field amplitude, operator norm, work fluctuations, control power, or total energy used by the controller [68]. A control term can reduce dissipation inside the target subsystem while requiring additional resources outside it. The net result depends on the chosen cost functional and on whether the controller, power supply, and eventual reset are included in the system boundary.

Shortcuts to thermodynamic computing make the same accounting issue explicit for information processing [61]. An auxiliary protocol may make the information-bearing subsystem evolve more quickly or faithfully, but this does not establish a reduction in total energy until the implementation of the auxiliary control is included.

The defensible conclusion for the hierarchy is therefore a boundary rule: a finite-time bound on the controlled subsystem and an energetic model of the controller must be reported separately before they are combined. Control costs are not assumed to be positive additions of a universal size, nor can they be omitted when a subsystem-level advantage is claimed.

## Measurement, feedback, and controller bookkeeping

Feedback can reduce the work required of a controlled subsystem by using information acquired in a measurement. For a feedback protocol with outcome  $M$ , a standard generalized second-law form is

$$\langle W \rangle - \Delta F \geq -k_B T I_{\text{nat}}(X; M), \quad (53)$$

where  $I_{\text{nat}}$  is mutual information in nats [63,69]. If mutual information is reported in bits, the right-hand side contains an additional factor of  $\ln 2$ .

Equation (53) is a statement about the controlled subsystem. Complete thermodynamic bookkeeping must also include the physical memory and apparatus that perform the measurement, store the outcome, implement the feedback, and are eventually restored to a reusable state. There is no universal standalone law of the form.

$$W_{\text{measurement}} \geq k_B T I$$

that is independent of the memory Hamiltonian and the measurement protocol.

For the standard idealized case of a memory with a degenerate Hamiltonian, resetting a memory whose logical-state entropy is  $H_{\text{nat}}(M)$  obeys the familiar Landauer bound

$$W_{\text{reset}} \geq k_B T H_{\text{nat}}(M). \quad (54)$$

The degeneracy assumption is important because the logical states then carry no free-energy difference associated with their energies. For a non-degenerate or time-dependent memory Hamiltonian, the minimum work depends additionally on changes in the memory free energy and, when present, on correlations between the memory and the measured system.

A fully reusable measurement–feedback device must restore not only the logical memory distribution but also the memory Hamiltonian, the relevant control parameters, and any correlations required for the initial operating state. Only the complete thermodynamic cycle can therefore be compared consistently with the ordinary second law [62,63].

Experiments with single-electron and colloidal systems have confirmed the information-dependent work relations in controlled settings [70,71]. Their lesson for computation is methodological: subsystem savings must not be reported while controller, readout, and memory–reset costs are omitted.

Training algorithms can be viewed abstractly as feedback processes, but the mutual information in a gradient does not by itself determine the electrical energy of a training step. A thermodynamic application would require a physical model of measurement, parameter storage, update, and reset. Consequently, this section does not assign a numerical Level 3 contribution to neural-network training.

### Mismatch cost: Distribution dependence of a physical process

A physical implementation is generally optimized for a set of operating conditions rather than for every possible input distribution. Mismatch cost formalizes the additional entropy production that can arise when the actual input distribution differs from a distribution for which the process is thermodynamically optimal [23,72,73].

Let denote a stochastic map induced by a fixed physical process, let be the actual input distribution, and let be an optimizing input distribution for that process. Under the decomposition assumptions of Kolchinsky and Wolpert, the excess entropy production is

$$\Sigma_{\text{mis}}(\rho; \varphi, \Phi) = D_{\text{KL}}(\rho \| \varphi) - D_{\text{KL}}(\Phi(\rho) \| \Phi(\varphi)) \geq 0. \quad (55)$$

Non-negativity follows from the data-processing inequality. The expression is measured in nats; multiplying by converts it to thermodynamic entropy, and multiplying by gives an energy scale for an isothermal setting.

**Theorem 5.1 (Mismatch-cost contraction)** *Under the hypotheses of the mismatch-cost decomposition [23], the excess entropy production of a process operated on rather than an optimizing distribution is given by eq. mismatch-cost. It vanishes precisely when the channel preserves the relative entropy of the pair  $(\rho, \varphi)$ :*

$$D_{\text{KL}}(\rho \| \varphi) = D_{\text{KL}}(\Phi(\rho) \| \Phi(\varphi)). \quad (56)$$

Equality can occur for injective deterministic maps, and more generally whenever the channel is sufficient or recoverable for the relevant pair. Injectivity is therefore a sufficient, not a necessary, condition.

The result is distribution- and process-dependent. The prior cannot be inferred merely from the truth table of a Boolean gate or from the statement that a circuit is CMOS. It

must be derived from a specified physical implementation or estimated experimentally. In particular, a uniform prior should not be assumed without evidence.

**Circuits:** For circuit models in which gates are thermodynamically modular and each gate  $\mathcal{G}$  is assigned an input distribution  $\rho_{\mathcal{G}}$  and an optimizing distribution  $\varphi_{\mathcal{G}}$ , the circuit mismatch contribution can be written schematically as

$$\Sigma_{\text{mis}}^{\text{circuit}} = \sum_{\mathcal{G} \in \mathcal{G}} \left[ D_{\text{KL}}(\rho_{\mathcal{G}} \| \varphi_{\mathcal{G}}) - D_{\text{KL}}(\Phi_{\mathcal{G}}(\rho_{\mathcal{G}}) \| \Phi_{\mathcal{G}}(\varphi_{\mathcal{G}})) \right]. \quad (57)$$

This is an excess contribution, not the total entropy production of the circuit. The total also contains the baseline dissipation of each gate and any coupling, clocking, communication, or control costs outside the modular model [74,75].

Different circuits computing the same function can therefore have different mismatch costs because their intermediate distributions and topology differ. Recent work has begun to formulate this dependence for circuit families [75,76], but exact comparisons require the gate priors and workload distribution. Speed, gate count, and mismatch cost need not be ordered in the same way (Figure 5).

Ripple-carry and carry-lookahead adders compute the same arithmetic function but induce different intermediate distributions, fan-in patterns, and modular decompositions. These structural differences can produce different mismatch costs under a specified set of gate priors. The figure is qualitative: no universal numerical ratio between the two architectures follows from topology alone. A quantitative comparison requires an explicit workload distribution, physical gate model, and an optimizing prior for every gate [76,77].

### Wolpert's synthesis: stochastic thermodynamics as the organizing language for energy costs

The mismatch-cost framework, together with the broader programme of stochastic thermodynamics of computation [73,74,78,79,], has been argued by Wolpert [25] to constitute the appropriate theoretical framework for understanding the energy costs of computation more broadly. In a recent perspective in PNAS, Wolpert [25] identifies four classes of contributions to the operational cost of any computational system, each of which corresponds to a distinct level of the hierarchy proposed in this review:

- The *Landauer cost* of the computation itself, set by the Shannon-entropy decrease of the logical operation (thm:landauer).
- The *finite-time cost* arising from the speed of operation relative to the relaxation time of the implementation (sec:finite-time).
- The *mismatch cost* arising from the difference between the actual input distribution and the implicit optimization prior of the device (thm: mismatch).

### Mismatch cost depends on circuit topology, not just function

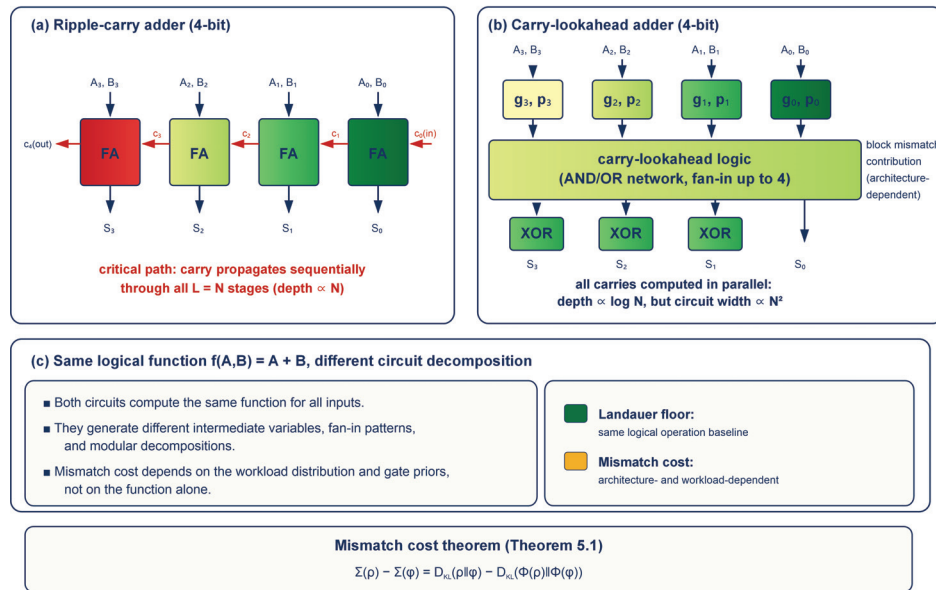


Figure 5: Circuit topology and mismatch accounting.

- The *residual entropy production* arising from microscopic implementation details not captured by any of the above (sec: overhead).

The framework of this review can be read, in part, as a quantitative extension and unification of Wolpert's classification: the hierarchy of Fig. 1 is structurally compatible with Wolpert's decomposition, but adds further granularity at the architectural and infrastructural levels and applies the resulting framework explicitly to the operational gap of contemporary AI hardware. Where Wolpert's exposition is programmatic and conceptual, we aim to render the classification quantitative and operationally testable.

### Implications for AI systems

Control theory and mismatch cost identify terms that are absent from a bare Landauer calculation, but neither can be assigned a universal numerical value for an accelerator.

Counterdiabatic control and feedback are relevant when a device actively uses such protocols. Their energetic benefit must be evaluated together with the controller, measurement apparatus, and reset of any memory used by the controller. Moving those components outside the system boundary can make a subsystem appear to beat a bound while merely relocating the cost.

Mismatch cost is potentially relevant to AI because layer-input distributions are structured and change with data, training, quantization, and model state. Nevertheless, three measurements are required before a thermodynamic mismatch energy can be reported:

- [leftmargin=2.2em,itemsep=0.25em] the stochastic map implemented by the physical device;

- its optimizing input distribution or family of distributions; and
- the actual input distribution at the same physical granularity.

Activation histograms alone do not determine a gate-level optimizing prior, and a KL divergence computed in software is not automatically a heat measurement. The most defensible present use of mismatch cost is therefore as a design principle and experimental hypothesis: hardware–algorithm co-design should reduce the divergence between workload statistics and the operating distribution for which the physical process is least dissipative.

This perspective yields a concrete research agenda. Profiling tools can estimate workload distributions; calibrated device models can infer candidate priors; and energy measurements can test whether reductions in the predicted contraction of relative entropy correlate with lower dissipation. Until such validation exists, Level 4 remains a distribution-dependent theoretical contribution rather than a closed term in an LLM energy budget.

### Architectural cost: From gates to neural workloads

The preceding sections concern lower bounds on specified physical processes. A processor, however, is not an idealized erasure device. It contains transistors, interconnects, clock distribution, memories, converters, control logic, and software-visible dataflow. Their combined energy is Level 5 of the hierarchy. This level is measured operationally and should not be inferred by multiplying  $k_B T \ln 2$  by an assumed number of “bit-equivalent operations.”

## Switching energy and reliability margins

For a capacitive node in CMOS, a standard first-order model for dynamic energy is

$$E_{\text{dyn}} \simeq a C_{\text{eff}} V_{\text{dd}}^2, \quad (58)$$

where  $a$  is an activity factor,  $C_{\text{eff}}$  is the effective switched capacitance, and  $V_{\text{dd}}$  is the supply voltage [80,81]. Static leakage, short-circuit current, clocking, and voltage conversion add further terms.

Thermal noise and target error probability constrain how closely signal energies may approach  $k_B T$ , but there is no universal decomposition of a modern processor into fixed factors such as a  $10^2$  “noise margin,” a  $10^3$  “voltage gap,” and a  $10^1$  “dataflow overhead.” Those ratios depend on technology node, circuit family, frequency, operating voltage, error tolerance, and the chosen logical and physical reference operation [82,83]. Such factors may be useful in a device-specific engineering audit, but they are not general thermodynamic constants.

The correct comparison is therefore two-tiered. A physical model provides a lower bound or switching-energy estimate for a defined device operation. Measurement then determines the realized energy of a circuit or workload. The ratio between them is meaningful only when the numerator and denominator refer to the same operation and system boundary.

## Arithmetic throughput: FLOPs are not erasures

Hardware data sheets report peak throughput under explicit conventions. For example, an H100-class accelerator may draw approximately 700 while advertising dense FP8 tensor throughput close to  $2 \times 10^{15}$  FLOP/s under favourable conditions [84]. Dividing power by peak throughput gives

$$e_{\text{FLOP}}^{\text{peak}} \simeq \frac{700}{2 \times 10^{15}} \approx 0.35. \quad (59)$$

If a fused multiply-add is counted as two FLOPs, the corresponding peak arithmetic energy is approximately 0.70 per MAC. These are peak ratios, not workload measurements. Lower utilization raises the effective energy per useful operation because device and host power do not scale linearly to zero.

Equation (59) can be normalized to show the separation between an operational arithmetic scale and the one-bit Landauer quantum. The result is only a reference ratio. A FLOP is not equivalent to the erasure of one bit, and a MAC cannot be assigned a Landauer minimum without specifying its input distribution, output retention, overwrite policy, and physical realization.

## Memory, communication, and conversion

Arithmetic is only one component of accelerator energy. Moving operands through the memory hierarchy,

communicating across chips, refreshing memories, distributing clocks, and converting between analogue and digital representations can equal or exceed arithmetic energy [39]. The balance changes with model architecture, batch size, sequence length, precision, sparsity, cache reuse, and parallelization strategy.

This observation has two implications for thermodynamic accounting.

First, a query-level estimate based only on FLOP count is incomplete. At minimum it

should state whether the estimate includes high-bandwidth memory, host processors, interconnect, idle capacity, and facility overhead.

Second, an apparent reduction in arithmetic energy can shift cost elsewhere. Analogue in-memory computing may reduce repeated weight movement while introducing ADC/DAC and calibration costs. Photonic multiplication may make the passive linear transformation inexpensive while leaving laser, modulator, detector, conversion, and memory energy outside the optical core. Neuromorphic hardware may exploit event sparsity but incur routing and state maintenance costs. System-level comparison must include these boundaries.

## Neural-network operations

**Matrix multiplication:** Dense matrix multiplication dominates many transformer workloads. Its operational energy is determined by the number and precision of arithmetic operations, data reuse, memory traffic, kernel efficiency, and hardware utilization. Parameter count alone is not a complete operation count: attention, mixture-of-experts routing, recurrent decoding, and context length change the amount of active computation per token.

**Activation functions:** A non-injective activation such as ReLU loses logical information at the abstract mathematical level. This fact does not imply a universal thermodynamic cost of one half bit per dimension. For continuous inputs, ReLU produces a mixed distribution with a point mass at zero and a continuous positive component; its entropy change depends on the quantization, measure, input distribution, and physical representation. A Landauer calculation is possible only after those choices are fixed.

**Training and backpropagation:** Backpropagation reuses intermediate information and contains algebraic steps that can be embedded in reversible computations, but a practical training system is not thermodynamically reversible. It overwrites buffers, updates optimizer state, communicates gradients, and repeatedly loads and stores parameters and activations. Checkpointing and recomputation change the time-memory-energy trade-off [85]. The relevant metric is therefore the measured energy of a training step under a stated implementation, not a claim that gradient computation is intrinsically close to Landauer’s bound.

## Alternative architectures

**Neuromorphic computing:** Neuromorphic systems combine event-driven operation, local state, and sparse communication. They have demonstrated large advantages on selected spatiotemporal and sparse workloads [86–88]. Those results should not be generalized to dense transformer training without workload-matched benchmarks. The hierarchy predicts that gains arise primarily at Level 5 through sparsity, locality, and reduced data movement; a Level 4 interpretation additionally requires a measured device prior.

**Analogue and in-memory computing:** Crossbar and compute-in-memory systems reduce the separation between storage and arithmetic [89–92]. Their total energy must include programming, peripheral circuits, ADC/DAC conversion, error correction, calibration, and the precision required by the application.

**Photonic computing:** Optical systems can perform linear transformations with high parallelism and low passive propagation loss [27,93,94]. End-to-end advantage depends on sources, modulators, detectors, electronic memory, nonlinearities, and conversion between domains. The absence of Boolean gates in a passive optical core does not make mismatch or control costs universally zero.

**Reversible and adiabatic computing:** Logically reversible circuits can avoid mandatory erasure within an ideal computation [8,95,96]. Practical implementations trade energy against speed, error, retained garbage information, clock generation, and eventual reset. Their relevance to AI depends on whether reversibility can be maintained through memory, communication, and input–output boundaries.

**Thermodynamic computing:** Thermodynamic computers use stochastic physical relaxation as a computational resource. Recent work has demonstrated a small-scale stochastic processing unit for sampling and linear-algebra primitives [73]. This is an important proof of principle, not yet evidence of a production-scale energy advantage. Programming, equilibration, readout, connectivity, accuracy, and digital interface costs must be included in comparisons.

## Implications for the hierarchy

Level 5 is where current AI energy can be measured most directly and where the largest near-term engineering gains are likely to occur. This conclusion does not require claiming that architecture contributes a fixed number of orders of magnitude. It follows from the fact that contemporary devices operate far above process-specific thermodynamic bounds and that measured energy changes substantially with precision, sparsity, utilization, memory placement, and dataflow.

For AI researchers, the practical output of this section is a reporting discipline:

[leftmargin=2.2em,itemsep=0..25em]

1. report the arithmetic convention, precision, and active operation count;

2. report context and output lengths, batch size, and utilization;
3. separate accelerator, memory, host, network, idle, and facility energy where possible;
4. compare alternatives on the same model, task, accuracy, and accounting boundary; and
5. use Landauer normalization only as a labelled reference scale unless logical entropy reduction has been quantified.

## Thermal and infrastructural overhead

Level 6 changes the accounting boundary from the computing equipment to the facility that delivers power and rejects heat. It is conceptually different from the lower levels: it does not determine the logical cost of computation, but multiplies or adds to the energy already consumed by the IT system.

## Full-stack accounting

For a query served over a time interval  $\mathcal{T}$ , a direct measurement can be written as

$$E_{IT}^{\text{query}} = \int_{\mathcal{T}} \left[ P_{\text{accelerator}}(t) + P_{\text{host}}(t) + P_{\text{memory}}(t) + P_{\text{network}}(t) + P_{\text{idle allocation}}(t) \right] dt. \quad (60)$$

The facility energy is then

$$E_{\text{facility}}^{\text{query}} = m_{\text{fac}} E_{IT}^{\text{query}}, \quad m_{\text{fac}} \geq 1, \quad (61)$$

when a multiplicative allocation is appropriate. Power–usage effectiveness is a common annual or facility-level proxy,

$$\text{PUE} = \frac{E_{\text{facility}}}{E_{IT}}. \quad (62)$$

PUE is not a property of the model, and it does not include losses already counted inside the IT boundary. Utilization and idle–capacity allocation must therefore not be hidden inside PUE if they are already represented in eq:query-it-energy.

Published PUE values vary by facility and accounting practice. The industry average remains substantially above the best hyperscale facilities [81][]. Query-level work should report the actual multiplier used, its time period, and whether it is measured or assumed.

## Thermodynamics of heat rejection

Let be the device junction temperature,  $T_C$  the coolant or cold-side temperature, and the ambient or heat-sink temperature. In an ordinary cooling path, the ordering is

$$T_J > T_C > T_A \quad (63)$$

when heat ultimately flows to a cooler ambient without refrigeration. Pumps, fans, and finite thermal resistances still consume work, but there is no Carnot requirement to pump heat from cold to hot.

Active refrigeration is required when heat must be rejected to a hotter sink, for example when  $T_C < T_A$ . For an ideal refrigerator operating between cold temperature  $T_C$  and hot temperature  $T_A$ , the Carnot coefficient of performance is

$$\text{COP}_{\text{Carnot}} = \frac{T_C}{T_A - T_C}, \quad (64)$$

and the minimum work per unit heat removed from the cold side is

$$\frac{W_{\min}}{Q_C} = \frac{T_A - T_C}{T_C}. \quad (65)$$

Real systems attain a fraction of the Carnot COP, so their work is larger than eq:cooling-carnot. This relation constrains a specified refrigeration cycle; it does not imply a universal irreducible PUE of 1.15–1.20. PUE also includes power delivery, lighting, fans, pumps, and other facility loads.

Warmer liquid cooling can reduce refrigeration requirements and improve heat-reuse opportunities, but the benefit depends on climate, coolant temperature, heat-exchanger design, and the temperature limits of the hardware. Facility optimization is therefore significant, yet it cannot remove energy already dissipated by arithmetic and data movement.

## Operational inference energy

There are two defensible routes to query energy.

**Measurement:** The preferred route integrates the full serving stack as in eq:query-it-energy. Elsworth et al. report that a median Gemini Apps text prompt used approximately 0.24 in their measured production environment, including active accelerators, host systems, idle capacity, and data-centre overhead [4]. The result is specific to the measured workload distribution and infrastructure.

**Model-based estimation:** When direct measurement is unavailable, the energy may be estimated from throughput and serving assumptions:

$$E_{\text{query}} = N_{\text{out}} \left[ E_{\text{decode/token}} + \frac{E_{\text{prefill}}}{N_{\text{out}}} \right], \quad (66)$$

where each term should include arithmetic, memory, communication, host, idle-capacity, and facility boundaries consistently. Oviedo et al. provide a model-based full-stack methodology and report a broad distribution of sub-watt-hour energies for large-model inference [6]. An informal estimate of approximately 0.3 per typical ChatGPT query [2][ ] is broadly consistent with that scale but should not be described as an audited measurement.

A simplified accelerator-only arithmetic estimate

$$E_{\text{arith}} = N_{\text{FLOP}} e_{\text{FLOP}}, \quad (67)$$

is useful for sensitivity analysis but is not a reconstruction of the full query unless all omitted terms are shown to be negligible. The number of active FLOPs must be derived from a specified model, context length, attention implementation, number of active experts, precision, and output length. Parameter count alone is insufficient.

## Where uncertainty enters

The principal uncertainties in public estimates are:

[leftmargin=2.2em,itemsep=0.25em]

- model identity and active parameter count;
- prefill and decoding sequence lengths;
- batch size, speculative decoding, and cache reuse;
- arithmetic convention and numerical precision;
- accelerator and host utilization;
- memory and interconnect traffic;
- allocation of idle capacity;
- facility multiplier; and
- whether energy is a mean, median, tail value, or scenario.

These variables can easily move the result by factors of several. Consequently, agreement within a factor of three may establish only an order-of-magnitude scale. It does not demonstrate that a proposed thermodynamic decomposition has accounted for the measured energy.

## Implications for efficiency programmes

Facility improvements can deliver immediate economic and environmental benefits, but their maximum percentage effect is limited by the fraction of energy already outside the IT load. A reduction of PUE from 1.5 to 1.1 is important; it cannot compensate for an inefficient model, low accelerator utilization, or excessive memory movement.

For AI research, the hierarchy therefore favours a full-stack order of operations: reduce unnecessary computation and data movement, improve utilization and workload-hardware matching, then minimize facility overhead. This is an accounting conclusion, not a claim that cooling is thermodynamically negligible.

## Synthesis: A unified hierarchical framework

The preceding sections establish a hierarchy of modelling and measurement questions rather than an additive partition of computational energy. At each level, one must identify the physical process under consideration, the quantity being

bounded or measured, the relevant input distribution and operating conditions, and the system boundary within which the energy is evaluated.

The purpose of the framework is therefore not to assign every observed joule to a unique thermodynamic term. Its purpose is to distinguish process-specific lower bounds from realized implementation costs and facility-level accounting quantities, while clarifying when different results may legitimately be compared.

### Aggregation of compatible lower bounds

A mathematically valid aggregation statement is possible only when all bounds refer to the same physical process, energy quantity, and system boundary.

#### Theorem 8.1 (Aggregation of co-normalized lower bounds)

Let  $\mathcal{P}$  be a specified physical implementation of a computational process, and let  $E_{\text{chip}}(\mathcal{P})$  denote its average chip-level energy under a specified input distribution and set of operating conditions.

Suppose that

$$L_0(\mathcal{P}), \dots, L_m(\mathcal{P})$$

are valid lower bounds on that same quantity, expressed using the same unit, physical boundary, input distribution, logical coarse-graining, temperature, and operating conditions. Then

$$E_{\text{chip}}(\mathcal{P}) \geq \max_{0 \leq i \leq m} L_i(\mathcal{P}). \quad (68)$$

If the allocated facility-level energy is related to the chip-level energy by

$$E_{\text{facility}}(\mathcal{P}) = m_{\text{fac}} E_{\text{chip}}(\mathcal{P}), \quad m_{\text{fac}} \geq 1, \quad (69)$$

then

$$E_{\text{facility}}(\mathcal{P}) \geq m_{\text{fac}} \max_{0 \leq i \leq m} L_i(\mathcal{P}). \quad (70)$$

*Proof.* By hypothesis, each  $L_i$  is a lower bound on the same quantity  $E_{\text{chip}}(\mathcal{P})$ . Therefore,

$$E_{\text{chip}}(\mathcal{P}) \geq L_i(\mathcal{P})$$

for every  $i \in \{0, \dots, m\}$ , which immediately implies eq:compatible-bound-maximum. Multiplication by the non-negative factor yields eq:facility-compatible-bound.

The significance of thm: hierarchy lies primarily in its hypotheses. Two quantities cannot be combined merely because both are reported in joules. A one-bit erasure bound, a strong-coupling bound for a fermionic mode, a measured accelerator energy per operation, and a facility-level energy per prompt do not automatically constrain the same physical quantity.

Similarly, lower bounds are not generally additive. Two inequalities may constrain the same realized dissipation from different physical or informational descriptions. Unless an independent derivation establishes additivity, the strongest compatible lower bound is given by their maximum rather than their sum.

The theorem should therefore be understood as an aggregation rule for co-normalized bounds. The substantive contribution of the framework is to identify the physical conditions under which such co-normalization is possible.

### The seven levels of the framework

The hierarchy classifies contributions according to both their physical origin and their epistemic status.

[leftmargin=2.8em,style=nextline,itemsep=0.5em]

This level contains the process-specific Landauer term

$$E_L = k_B T \ln 2 \Delta H_{\text{erase}}. \quad (71)$$

It constitutes a lower bound only after the information-bearing degrees of freedom, the logical coarse-graining, and the entropy reduction have been defined. It is not a universal energy assigned to every instruction, FLOP, MAC, gate evaluation, or transistor transition.

This level contains additional process-specific bounds associated with finite protocol duration, thermodynamic length, dynamical activity, error tolerance, and the stochastic dynamics of the implementation. These results apply only when the state space, transition dynamics, and protocol are explicitly specified (sec:finite-time).

This level includes modifications associated with coherence, quantum correlations, many-body degrees of freedom, finite-size reservoirs, non-Markovianity, and non-negligible system-bath interactions (sec: quantum). Such corrections are process-dependent and cannot be assigned numerically to a conventional processor without an appropriate microscopic model.

This level accounts for the energy required to implement a control protocol, prepare and drive the physical substrate, measure its state, process feedback information, and reset controller memories (sec: control). These costs may lie outside the subsystem used to derive a lower bound and must therefore be included explicitly when the system boundary is expanded.

This level contains excess entropy production caused by operating a fixed physical process under an input distribution different from the distribution for which that process is thermodynamically optimized. Its evaluation requires a physical channel, an optimizing prior, and an actual operating distribution defined at the same granularity.

This level contains the realized energy associated with switching, arithmetic, memory access, interconnects, communication, numerical conversion, clocking,

synchronization, and software-visible dataflow (sec: architecture). It is normally determined by measurement or device-level modelling rather than by multiplying an operation count by the Landauer quantum.

This level expands the accounting boundary from the chip or IT equipment to the facility. It includes power conversion and distribution, cooling, pumping, fans, heat rejection, idle allocation, and other infrastructure energy (sec: overhead). It is commonly represented through a facility multiplier, but the value and meaning of that multiplier depend on the adopted allocation method.

The hierarchy does not imply the additive identity

$$E_{\text{op}} = \sum_{\ell=0}^6 E_{\ell}. \quad (72)$$

Some levels contain alternative lower bounds on the same physical quantity. Others contain excess costs defined relative to a reference process, measured implementation energies, or accounting multipliers. Consequently, the appropriate mathematical relationship must be determined separately for each physical implementation and system boundary.

### Geometric and information-theoretic structures

Several results associated with the lower levels of the framework share a geometric formulation. In the slow-driving regime, excess work can be expressed as a quadratic form in protocol velocity, with a friction tensor defining a thermodynamic metric and thermodynamic length [17,40,41]. Optimal-transport formulations introduce a geometry on the space of probability distributions [67]. Related geometric structures also arise in quantum-state manifolds and quantum-control problems.

Mismatch cost is usually formulated instead through the contraction of relative entropy under a physical channel. This gives it an explicitly information-theoretic structure. Information geometry may provide connections between these descriptions, but a general theorem unifying finite-time thermodynamic length, quantum-control geometry, and circuit-level mismatch cost has not yet been established.

The distinction should therefore not be interpreted as proving that higher levels lack geometric structure. Rather, it identifies a gap between currently available theories. A unified treatment may require a combination of Riemannian geometry, optimal transport, graph structure, information geometry, and stochastic control.

Architecture- and facility-level quantities introduce further difficulties. They involve discrete circuit topologies, heterogeneous components, software schedules, shared resources, workload-dependent utilization, and accounting allocations. Any future geometric description of these levels must preserve the physical meaning and system boundary of each metric rather than combining formally similar quantities without a common operational interpretation.

### Open problems and research programme

The hierarchy suggests seven research problems that would make the connection between thermodynamic bounds and AI systems more quantitative and falsifiable.

#### Problem 1: logical-entropy audits of AI primitives.

For a quantized arithmetic primitive, activation function, memory update, communication step, or inference subroutine, define the information-bearing state, input distribution, retained outputs, discarded variables, overwritten memory, and total logical entropy reduction. Determine under what conditions a workload-level bound can be calculated or bounded without double-counting intermediate states.

Such an audit is necessary before a process-specific Landauer bound can be assigned to a complete neural computation. Operation counts alone are insufficient.

#### Problem 2: finite-time models of computing devices.

Construct calibrated stochastic models for representative memory, logic, and analogue-computing devices. Measure or infer the quantities entering thermodynamic-length and activity-based bounds, including transition rates, endpoint distributions, dynamical activity, relaxation times, protocol duration, and error probability.

These experiments should report both the energy of the modelled subsystem and the energy of the controllers, drivers, sensors, and readout electronics. This would establish whether Level 1 corrections remain experimentally observable after the full control boundary is included.

#### Problem 3: realistic quantum and strong-coupling substrates.

Determine whether coherence, correlations, many-body effects, or tunable system-bath coupling can reduce total device energy under realistic temperature, speed, connectivity, and error-tolerance requirements. Comparisons must include state preparation, control, stabilization, readout, reset, and any refrigeration needed by the substrate.

A reduction in the energy of an isolated elementary transformation does not by itself imply a reduction in system-level energy.

#### Problem 4: empirical measurement of mismatch cost.

Infer or experimentally determine optimizing distributions for physical logic gates, analogue arrays, stochastic processors, and neuromorphic devices. Compare those distributions with measured workload distributions and calculate the corresponding relative-entropy contraction.

The central empirical question is whether the predicted mismatch term correlates quantitatively with measured excess heat or electrical energy under controlled changes of the input distribution. Such an experiment would provide a direct test of the relevance of Level 4 to computing hardware.

### Problem 5: architecture-matched AI benchmarks.

Compare digital, neuromorphic, analogue in-memory, photonic, reversible, and thermodynamic systems using identical tasks, accuracy targets, numerical precision, latency constraints, and system boundaries. The benchmark should separately report:

[leftmargin=2.2em,itemsep=0.2em]

1. arithmetic or core-computation energy;
2. memory and data-movement energy;
3. conversion and peripheral-circuit energy;
4. host and control energy;
5. idle-capacity allocation; and
6. facility-level overhead.

Without these controls, comparisons between an efficient core operation and a complete conventional system are not physically meaningful.

### Problem 6: reproducible full-stack inference accounting.

Develop open measurement protocols that report distributions, rather than only a single average, of joules per token and watt-hours per prompt. At minimum, the protocol should disclose context length, output length, batching, numerical precision, active parameters, accelerator utilization, memory and communication traffic, host-system power, idle allocation, and facility multiplier [2,4,6].

Where proprietary constraints prevent disclosure of model or infrastructure details, uncertainty intervals and explicit accounting assumptions should be reported.

### Problem 7: falsifiable predictions for emerging paradigms.

For each proposed computing paradigm, identify:

[leftmargin=2.2em,itemsep=0.2em]

1. the level or component expected to decrease;
2. the physical mechanism responsible for the decrease;
3. the components expected to remain unchanged or increase;
4. the adopted system boundary;
5. the reference conventional system; and
6. a numerical system-level prediction stated before measurement.

A paradigm is not validated by demonstrating a low-energy core operation when conversion, programming, control, readout, memory, or data movement dominates the complete workload.

### Implications for emerging computing paradigms

The hierarchy provides a common language for assessing alternative computing technologies. Different paradigms target different parts of the energy stack and should therefore not be evaluated using a single device-level metric.

**Neuromorphic systems:** The primary hypothesis is a reduction of Level 5 energy through event-driven operation, local memory, sparse communication, and co-location of state and computation. These advantages are expected to be strongest for sparse, event-based, and spatiotemporal workloads. They may be smaller for dense matrix multiplication unless the model, data representation, and training procedure are redesigned for the hardware.

Claims of neuromorphic efficiency should include the energy of encoding, routing, learning, host communication, and any conversion between conventional and event-based representations.

**Analogue in-memory systems:** Analogue in-memory computing primarily targets Level 5 by reducing the movement of weights and intermediate activations between memory and arithmetic units. Its end-to-end advantage depends on programming energy, peripheral circuits, ADC and DAC conversion, calibration, device variability, precision, endurance, refresh, and error correction.

Claims involving Level 4 additionally require an experimentally grounded physical prior and a measured workload distribution. Analogue operation alone does not establish a reduced mismatch cost.

**Photonic systems:** Passive or weakly driven optical transformations can reduce the energy of selected linear operations and communication tasks. However, end-to-end efficiency depends on light sources, modulators, detectors, nonlinear elements, optical memory, electronic control, and repeated conversion between optical and electronic domains.

The appropriate benchmark must therefore include the complete electro-optical system rather than the energy of an isolated optical matrix multiplication.

**Reversible systems:** Logical reversibility can reduce or eliminate Level 0 entropy reduction within an idealized internal computation. A practical benefit, however, depends on protocol speed, error control, garbage retention and uncomputation, clocking, signal restoration, input preparation, and output reset.

The relevant comparison is therefore the complete algorithmic cycle, including initialization and reuse of memory, rather than an isolated reversible gate.

**Thermodynamic systems:** Thermodynamic and stochastic computing systems exploit physical relaxation, fluctuations, or equilibrium sampling as computational resources. They may be naturally suited to sampling, probabilistic inference, and optimization tasks.

Existing experiments demonstrate small-scale feasibility [73], but scaling, programmability, connectivity, equilibration time, control, readout, accuracy, and task-level benchmarking remain open. Their evaluation should distinguish the energy of passive relaxation from the energy required to establish, maintain, and interrogate the physical energy landscape.

## The framework as a technology-assessment tool

An energy-efficiency claim should be accompanied by four explicit statements:

[leftmargin=2.2em,itemsep=0.3em]

1. the level, component, or physical process whose energy is expected to decrease;
2. the system boundary, workload, operating conditions, and comparison baseline;
3. the epistemic status of the reported quantity—a rigorous lower bound, analytical model, numerical estimate, laboratory measurement, or full-stack operational measurement; and
4. the costs that are shifted to other levels, components, or accounting boundaries.

This reporting discThe framework prevents three recurrent errors: treating as a universal cost per operation, summing lower bounds that constrain different physical quantities, and comparing an efficient core with a full-stack conventional system.

It also makes the hierarchy directly useful to AI researchers. Rather than functioning as a purported decomposition of every joule consumed by an AI system, the framework becomes a checklist for experimental design, reporting, benchmarking, and hardware–algorithm co-design.

The central conclusion of the synthesis is therefore methodological: fundamental bounds, physical implementations, computer architectures, and facilities must be connected through explicit system boundaries and measurable quantities. Only then can thermodynamic reasoning provide a reliable basis for assessing the energetic limits and opportunities of neural computation.

## Conclusions and outlook

Landauer's principle remains the correct starting point for the thermodynamics of logically irreversible information processing. Its use in AI energy analysis, however, requires a distinction that is often lost: the cost associated with erasing one unbiased bit under specified assumptions is not a universal price attached to every FLOP, MAC, instruction, activation, or transistor transition.

This review developed a hierarchy that preserves that distinction. Levels 0–4 describe process-specific thermodynamic terms: logical entropy reduction, finite-time constraints, quantum and strong-coupling structure, control

costs, and mismatch. Level 5 describes realized architecture, including arithmetic, memory, communication, conversion, and dataflow. Level 6 extends the boundary to facility power and heat rejection. The levels do not form a universal additive decomposition. Compatible lower bounds combine through a maximum, while empirical overheads require direct measurement or an explicit accounting model.

The principal practical conclusion is that present AI systems are governed primarily by architectural and systems variables. Precision, active parameters, sequence length, batching, utilization, memory traffic, network topology, and idle-capacity allocation strongly affect joules per token and watt-hours per prompt. Published full-stack studies place representative text-prompt energy in the sub-watt-hour range, but also demonstrate that no single query-energy number is universal [4,6].

Fundamental thermodynamic results remain strategically important because they identify which reductions are possible in principle and which trade energy against time, error, control, or information retention. Their near-term value is greatest when coupled to experiments on actual devices, not when assigned speculative numerical coefficients in a workload budget.

For AI researchers, the hierarchy provides a concrete reporting standard: define the physical and computational boundary, distinguish measurement from estimation, state the arithmetic convention, expose uncertainty, and label Landauer normalization as a reference ratio unless logical entropy reduction has been audited. For hardware research, it provides a common language for comparing neuromorphic, in-memory, photonic, reversible, and thermodynamic approaches without omitting conversion or control costs.

The central open challenge is to connect these levels quantitatively without conflating them. Achieving that goal will require calibrated stochastic device models, workload-specific entropy audits, empirical measurement of physical optimization priors, and reproducible full-stack energy benchmarks. Such a programme would turn the thermodynamics of AI from an order-of-magnitude analogy into a falsifiable science of computational energy.

## Appendix

### Operational accounting and Landauer reference scales

This appendix presents calculations that can be reproduced from stated quantities. Its purpose is to keep three conceptually distinct objects separate:

[leftmargin=2.2em,itemsep=0.25em]

1. a process-specific thermodynamic lower bound;
2. a Landauer-normalized reference ratio; and
3. an operational energy estimate or measurement.

## The one-bit reference scale

At temperature  $T$ ,

$$E_{\text{bit,L}}(T) = k_B T \ln 2. \quad (73)$$

Using  $k_B = 1.380649e-23$  and  $T = 300$  gives

$$E_{\text{bit,L}}(300) \approx 2.87e-21. \quad (74)$$

For a process that erases bits, the lower bound is  $E_{\text{bit,L}} \Delta H_{\text{erase}}$ . No operation-count conversion is made unless has been established.

## Peak accelerator arithmetic

For an accelerator with power  $P_{\text{acc}}$  and stated peak throughput  $\Theta_{\text{FLOP}}$ ,

$$e_{\text{FLOP}}^{\text{peak}} = \frac{P_{\text{acc}}}{\Theta_{\text{FLOP}}}. \quad (75)$$

Using an illustrative H100 specification of  $P_{\text{acc}} = 700$  and  $\Theta_{\text{FLOP}} = 1.98 \times 10^{15}$  dense FP8 FLOP/s gives

$$e_{\text{FLOP}}^{\text{peak}} \approx 3.54e-13 = 0.354. \quad (76)$$

With two FLOPs per fused multiply-add, the corresponding peak MAC ratio is approximately.

$$e_{\text{MAC}}^{\text{peak}} \approx 0.708. \quad (77)$$

These values assume peak throughput. An effective workload value should be computed from measured energy and useful operations:

$$e_{\text{FLOP}}^{\text{eff}} = \frac{\int T P_{\text{acc}}(t) dt}{N_{\text{FLOP}}^{\text{useful}}}. \quad (78)$$

For reference only,

$$R_{\text{L,FLOP}}^{\text{peak}} = \frac{e_{\text{FLOP}}^{\text{peak}}}{k_B T \ln 2} \approx 1.2 \times 10^8 \quad (T = 300). \quad (79)$$

This number is not a claim that one FLOP erases bits or that its realizable minimum is  $k_B T \ln 2$ .

## Query-level measurement

A measured query energy should be reported with its workload distribution. For the median production prompt reported by Elsworth et al.,

$$E_{\text{query}}^{\text{median}} \approx 0.24 = 864. \quad (80)$$

Its Landauer-normalized reference ratio at 300 is

$$R_{\text{L,query}} = \frac{864}{2.87e-21} \approx 3.0 \times 10^{23}. \quad (81)$$

Again, this is only a reference ratio. It is not an estimate of the number of logical bit erasures in the prompt.

Oviedo et al. report a distribution of model-based full-stack estimates rather than one universal value [6]. When comparing with such work, the manuscript should quote the median and uncertainty interval and should not describe the estimate as a direct measurement.

## Parameterized model-based estimate

For a specified serving configuration, a reproducible estimate can be organized as

$$E_{\text{accelerator}} = N_{\text{FLOP}} e_{\text{FLOP}}^{\text{eff}}, \quad (82)$$

$$E_{\text{IT}} = E_{\text{accelerator}} + E_{\text{host}} + E_{\text{memory}} + E_{\text{network}} + E_{\text{idle}}, \quad (83)$$

$$E_{\text{facility}} = m_{\text{fac}} E_{\text{IT}}. \quad (84)$$

The following inputs are mandatory: Table 1.

A value should be derived from this configuration. Parameter count alone does not establish the number of MACs performed per token; the operation count must be derived from the specified architecture, active parameters, sequence lengths, and serving configuration.

## Uncertainty and comparison

Let denote the vector of assumptions in tab:app-required-inputs. A model-based result should be reported as

$$E_{\text{query}} = f(\theta), \quad \theta \in \Theta, \quad (85)$$

with a sensitivity analysis over a plausible set  $\Theta$ . Agreement within a factor of several can support an order-of-magnitude comparison, but it does not validate a unique thermodynamic decomposition.

The numerical values associated with Levels 1–4 are

**Table 1:** Minimum information for a model-based inference-energy estimate.

| Quantity   | Required specification                                  |
|------------|---------------------------------------------------------|
| Model      | architecture, total and active parameters, precision    |
| Sequence   | input length, output length, attention implementation   |
| Serving    | batch size, decoding strategy, cache reuse, utilization |
| Operations | FLOP/MAC convention and derivation of active operations |
| Hardware   | accelerator type, count, power or measured energy       |
| Full stack | host, memory, network, idle-capacity allocation         |
| Facility   | measured or assumed multiplier/PUE and time period      |
| Statistic  | mean, median, percentile, or scenario                   |

intentionally omitted from this appendix. They require device-specific stochastic models, control boundaries, or physical priors that are not available for the LLM serving scenario. Assigning order-one coefficients would create precision without identifiability.

## Acknowledgments

The author thanks colleagues at the Graviton Scientific Society and the Federal University of Piauí for valuable discussions during the preparation of this review.

## Declarations

**Conflict of interest:** The author declares no competing financial or non-financial interests.

**Data availability:** No new datasets were generated or analyzed in this study. The numerical examples were derived from publicly available sources cited in the article.

**Declaration of generative AI and AI-assisted technologies:** Generative AI tools were used exclusively to assist with the translation of the manuscript into English and its formatting in LaTeX. They were not used to generate scientific content, data, results, analyses, interpretations, or conclusions. The author reviewed the entire manuscript and assumes full responsibility for its accuracy and integrity.

## References

- International Energy Agency. Energy and AI. Paris: International Energy Agency; 2025.
- Henderson P, Hu J, Romoff J, Brunskill E, Jurafsky D, Pineau J. Towards the systematic reporting of the energy and carbon footprints of machine learning. *J Mach Learn Res*. 2020;21(248):1-43. Available from: <https://jmlr.org/papers/v21/20-312.html>
- Schwartz R, Dodge J, Smith NA, Etzioni O. Green AI. *Commun ACM*. 2020;63(12):54-63. Available from: <https://dl.acm.org/doi/10.1145/3381831>
- Elsworth C, Huang K, Patterson D, Schneider I, Sedivy R, Goodman S, et al. Measuring the environmental impact of delivering AI at Google scale. *arXiv [Preprint]*. 2025;2508.15734. Available from: <https://arxiv.org/abs/2508.15734>
- Luccioni AS, Jernite Y, Strubell E. Power hungry processing: watts driving the cost of AI deployment? In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. New York: ACM; 2024. p. 85-99. Available from: <https://dl.acm.org/doi/10.1145/3630106.3658542>
- Oviedo F, Kazhamiaka F, Choukse E, Kim A, Luers A, Nakagawa M, et al. Energy use of AI inference, efficiency pathways, and test-time scaling. *Joule*. 2026;102430. Available from: <https://www.sciencedirect.com/science/article/pii/S2542435126001145>
- Wilhelm P, Wittkopp T, Kao O. Beyond test-time compute strategies: advocating energy-per-token in LLM inference. In: *Proceedings of the 5th Workshop on Machine Learning and Systems (EuroMLSys '25)*. New York: ACM; 2025. p. 208-215. Available from: <https://dl.acm.org/doi/abs/10.1145/3721146.3721953>
- Bennett CH. Logical reversibility of computation. *IBM J Res Dev*. 1973;17(6):525-532. Available from: <https://dl.acm.org/doi/10.1147/rd.176.0525>
- Bennett CH. The thermodynamics of computation—a review. *Int J Theor Phys*. 1982;21:905-940. Available from: <https://link.springer.com/article/10.1007/BF02084158>
- Landauer R. Irreversibility and heat generation in the computing process. *IBM J Res Dev*. 1961;5(3):183-191. Available from: <https://ieeexplore.ieee.org/document/5392446>
- Parrondo JMR, Horowitz JM, Sagawa T. Thermodynamics of information. *Nat Phys*. 2015;11:131-139. Available from: <https://www.nature.com/articles/nphys3230>
- Jarzynski C. Equalities and inequalities: irreversibility and the second law of thermodynamics at the nanoscale. *Annu Rev Condens Matter Phys*. 2011;2:329-351. Available from: <https://www.annualreviews.org/content/journals/10.1146/annurev-conmatphys-062910-140506>
- Seifert U. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Rep Prog Phys*. 2012;75(12):126001. Available from: <https://iopscience.iop.org/article/10.1088/0034-4885/75/12/126001>
- Van den Broeck C, Esposito M. Ensemble and trajectory thermodynamics: a brief introduction. *Physica A*. 2015;418:6-16. Available from: <https://www.sciencedirect.com/science/article/abs/pii/S037843711400346X>
- Aurell E, Gawłdzki K, Mejía-Monasterio C, Mohayaei R, Muratore-Ginanneschi P. Refined second law of thermodynamics for fast random processes. *J Stat Phys*. 2012;147:487-505. Available from: <https://arxiv.org/abs/1201.3207>
- Aurell E, Mejía-Monasterio C, Muratore-Ginanneschi P. Optimal protocols and optimal transport in stochastic thermodynamics. *Phys Rev Lett*. 2011;106:250601. Available from: <https://arxiv.org/abs/1012.2037>
- Sivak DA, Crooks GE. Thermodynamic metrics and optimal paths. *Phys Rev Lett*. 2012;108:190602. Available from: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.108.190602>
- Lee JS, Lee S, Kwon H, Park H. Speed limit for a highly irreversible process and tight finite-time Landauer's bound. *Phys Rev Lett*. 2022;129:120603. Available from: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.129.120603>
- Proesmans K, Ehrich J, Bechhoefer J. Finite-time Landauer principle. *Phys Rev Lett*. 2020;125:100602. Available from: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.125.100602>
- Goold J, Huber M, Riera A, del Rio L, Skrzypczyk P. The role of quantum information in thermodynamics—a topical review. *J Phys A Math Theor*. 2016;49(14):143001. Available from: <https://iopscience.iop.org/article/10.1088/1751-8113/49/14/143001>
- Rolandi A, Perarnau-Llobet M. Finite-time Landauer principle beyond weak coupling. *Quantum*. 2023;7:1161. Available from: <https://quantum-journal.org/papers/q-2023-11-03-1161/>
- Van Vu T, Saito K. Finite-time quantum Landauer principle and quantum coherence. *Phys Rev Lett*. 2022;128:010602. Available from: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.128.010602>
- Kolchinsky A, Wolpert DH. Dependence of integrated, instantaneous, and fluctuating entropy production on the initial state in quantum and classical processes. *Phys Rev E*. 2021;104:054107. Available from: <https://journals.aps.org/pre/abstract/10.1103/PhysRevE.104.054107>
- Wolpert DH. The stochastic thermodynamics of computation. *J Phys A Math Theor*. 2019;52(19):193001. Available from: <https://iopscience.iop.org/article/10.1088/1751-8121/ab0850>
- Wolpert DH. Is stochastic thermodynamics the key to understanding the energy costs of computation? *Proc Natl Acad Sci U S A*. 2024;121. Available from: <https://www.pnas.org/doi/10.1073/pnas.2321112121>

26. Marković D, Mizrahi A, Querlioz D, Grollier J. Physics for neuromorphic computing. *Nat Rev Phys*. 2020;2:499-510. Available from: <https://www.nature.com/articles/s42254-020-0208-2>
27. Shastri BJ, Tait AN, Ferreira de Lima T, Pernice WHP, Bhaskaran H, Wright CD, et al. Photonics for artificial intelligence and neuromorphic computing. *Nat Photonics*. 2021;15:102-114. Available from: <https://www.nature.com/articles/s41566-020-00754-y>
28. Esposito M, Van den Broeck C. Second law and Landauer principle far from equilibrium. *Europhys Lett*. 2011;95(4):40004. Available from: <https://iopscience.iop.org/article/10.1209/0295-5075/95/40004>
29. Bérut A, Arakelyan A, Petrosyan A, Ciliberto S, Dillenschneider R, Lutz E. Experimental verification of Landauer's principle linking information and thermodynamics. *Nature*. 2012;483:187-189. Available from: <https://www.nature.com/articles/nature10872>
30. Jun Y, Gavrilov M, Bechhoefer J. High-precision test of Landauer's principle in a feedback trap. *Phys Rev Lett*. 2014;113:190601. Available from: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.113.190601>
31. Hong J, Lambson B, Dhuey S, Bokor J. Experimental test of Landauer's principle in single-bit operations on nanomagnetic memory bits. *Sci Adv*. 2016;2. Available from: <https://www.science.org/doi/10.1126/sciadv.1501492>
32. Koski JVR, Maisi VF, Sagawa T, Pekola JP. Experimental observation of the role of mutual information in the nonequilibrium dynamics of a Maxwell demon. *Phys Rev Lett*. 2014;113:030601. Available from: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.113.030601>
33. Yan LL, Xiong TP, Rehan K, Zhou F, Liang DF, Chen L, et al. Single-atom demonstration of the quantum Landauer principle. *Phys Rev Lett*. 2018;120:210601. Available from: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.120.210601>
34. Landauer R. Information is physical. *Phys Today*. 1991;44(5):23-29. Available from: <https://www.scirp.org/reference/referencespapers?referenceid=4314028>
35. Aïmet S, Tajik M, Tournaire G, Schüttelkopf P, Sabino J, Sotiriadis S, et al. Experimentally probing Landauer's principle in the quantum many-body regime. *Nat Phys*. 2025;21:1326-1331. Available from: <https://www.nature.com/articles/s41567-025-02930-9>
36. Schmiedl T, Seifert U. Optimal finite-time processes in stochastic thermodynamics. *Phys Rev Lett*. 2007;98:108301. Available from: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.98.108301>
37. Benamou JD, Brenier Y. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numer Math*. 2000;84:375-393. Available from: <https://link.springer.com/article/10.1007/s002110050002>
38. Dechant A, Sasa SI. Geometric decomposition of entropy production in out-of-equilibrium systems. *Phys Rev Res*. 2021;3.
39. Frim AG, DeWeese MR. Geometric bound on the efficiency of irreversible thermodynamic cycles. *Phys Rev Lett*. 2022;128:230601. Available from: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.128.230601>
40. Scandi M, Perarnau-Llobet M. Thermodynamic length in open quantum systems. *Quantum*. 2019;3:197. Available from: <https://quantum-journal.org/papers/q-2019-10-24-197/>
41. Scandi M, Barker D, Lehmann S, Dick KA, Maisi VF, Perarnau-Llobet M. Minimally dissipative information erasure in a quantum dot via thermodynamic length. *Phys Rev Lett*. 2022;129:270601. Available from: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.129.270601>
42. Epoch AI. How much energy does ChatGPT use? *Epoch AI Gradient Updates*. 2025.
43. Zhen YZ, Egloff D, Modi K, Dahlsten O. Universal bound on energy cost of bit reset in finite time. *Phys Rev Lett*. 2021;127:190602. Available from: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.127.190602>
44. Ma YH, Chen JF, Sun CP, Dong H. Minimal energy cost to initialize a bit with tolerable error. *Phys Rev E*. 2022;106:034112. Available from: <https://journals.aps.org/pre/abstract/10.1103/PhysRevE.106.034112>
45. Buffoni L, Campisi M. Generalized Landauer bound from absolute irreversibility. *arXiv [Preprint]*. 2024;2310.05449. Available from: <https://arxiv.org/abs/2310.05449>
46. Lorenzo S, McCloskey R, Ciccarello F, Paternostro M, Palma GM. Landauer's principle in multipartite open quantum system dynamics. *Phys Rev Lett*. 2015;115:120403. Available from: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.115.120403>
47. Reeb D, Wolf MM. An improved Landauer principle with finite-size corrections. *New J Phys*. 2014;16:103011. Available from: <https://iopscience.iop.org/article/10.1088/1367-2630/16/10/103011>
48. Chattopadhyay P, Misra A, Pandit T, Paul G. Landauer principle and thermodynamics of computation. *Rep Prog Phys*. 2025;88:086001. Available from: <https://arxiv.org/abs/2506.10876>
49. Vinjanampathy S, Anders J. Quantum thermodynamics. *Contemp Phys*. 2016;57(4):545-579. Available from: <https://www.tandfonline.com/doi/full/10.1080/00107514.2016.1201896>
50. Binder F, Correa LA, Gogolin C, Anders J, Adesso G, editors. *Thermodynamics in the quantum regime*. Vol. 195. Cham: Springer; 2018.
51. Landi GT, Paternostro M. Irreversible entropy production: from classical to quantum. *Rev Mod Phys*. 2021;93:035008. Available from: <https://journals.aps.org/rmp/abstract/10.1103/RevModPhys.93.035008>
52. Breuer HP, Petruccione F. *The theory of open quantum systems*. Oxford: Oxford University Press; 2002.
53. Spohn H. Entropy production for quantum dynamical semigroups. *J Math Phys*. 1978;19:1227-1230. Available from: <https://pubs.aip.org/aip/jmp/article-abstract/19/5/1227/460049/Entropy-production-for-quantum-dynamical?redirectedFrom=fulltext>
54. Rivas A. Strong coupling thermodynamics of open quantum systems. *Phys Rev Lett*. 2020;124:160601. Available from: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.124.160601>
55. Strasberg P, Schaller G, Schmidt TL, Esposito M. Fermionic reaction coordinates and their application to an autonomous Maxwell demon in the strong-coupling regime. *Phys Rev B*. 2018;97:205405. Available from: <https://journals.aps.org/prb/abstract/10.1103/PhysRevB.97.205405>
56. Bormashenko E. Landauer bound in the context of minimal physical principles: meaning, experimental verification, controversies and perspectives. *Entropy*. 2024;26:423. Available from: <https://www.mdpi.com/1099-4300/26/5/423>
57. Bormashenko E. Landauer's principle: past, present and future. *Entropy*. 2025;27:437. Available from: <https://www.mdpi.com/1099-4300/27/4/437>
58. Seifert U. First and second law of thermodynamics at strong coupling. *Phys Rev Lett*. 2016;116:020601. Available from: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.116.020601>
59. Maroney OJE. Generalizing Landauer's principle. *Phys Rev E*. 2009;79:031105. Available from: <https://journals.aps.org/pre/abstract/10.1103/PhysRevE.79.031105>
60. Lipka-Bartosik P, Perarnau-Llobet M. Minimizing dissipation via interacting environments: quadratic convergence to Landauer bound. *Phys Rev Lett*. 2025;135:170404. Available from: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.135.170404>

61. Boyd AB, Patra A, Aifer B, Crutchfield JP. Shortcuts to thermodynamic computing: the cost of fast and faithful information processing. *J Stat Phys*. 2022;187:17. Available from: <https://link.springer.com/article/10.1007/s10955-022-02871-0>
62. Sagawa T, Ueda M. Minimal energy cost for thermodynamic information processing: measurement and information erasure. *Phys Rev Lett*. 2009;102:250602. Available from: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.102.250602>
63. Sagawa T, Ueda M. Generalized Jarzynski equality under nonequilibrium feedback control. *Phys Rev Lett*. 2010;104:090602. Available from: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.104.090602>
64. Berry MV. Transitionless quantum driving. *J Phys A Math Theor*. 2009;42:365303. Available from: <https://iopscience.iop.org/article/10.1088/1751-8113/42/36/365303>
65. Demirlak M, Rice SA. Adiabatic population transfer with control fields. *J Phys Chem A*. 2003;107:9937-9945. Available from: <https://pubs.acs.org/jpcfh/article-abstract/107/46/9937/3779261/Adiabatic-Population-Transfer-with-Control-Fields?redirectedFrom=fulltext>
66. Guéry-Odelin D, Ruschhaupt A, Kiely A, Torriontgui E, Martínez-Garaot S, Muga JG. Shortcuts to adiabaticity: concepts, methods, and applications. *Rev Mod Phys*. 2019;91:045001. Available from: <https://journals.aps.org/rmp/abstract/10.1103/RevModPhys.91.045001>
67. Li G, Quan HT, Tu ZC. Shortcuts to isothermality and nonequilibrium work relations. *Phys Rev E*. 2017;96:012144. Available from: <https://journals.aps.org/pre/abstract/10.1103/PhysRevE.96.012144>
68. Funo K, Zhang JN, Chatou C, Kim K, Ueda M, del Campo A. Universal work fluctuations during shortcuts to adiabaticity by counterdiabatic driving. *Phys Rev Lett*. 2017;118:100602. Available from: <https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.118.100602>
69. Uptime Institute. Global data center survey results 2024. Uptime Institute Annual Survey. 2024.
70. Koski JVR, Maisi VF, Pekola JP, Averin DV. Experimental realization of a Szilard engine with a single electron. *Proc Natl Acad Sci U S A*. 2014;111:13786-13789. Available from: <https://www.pnas.org/doi/10.1073/pnas.1406966111>
71. Toyabe S, Sagawa T, Ueda M, Muneyuki E, Sano M. Experimental demonstration of information-to-energy conversion and validation of the generalized Jarzynski equality. *Nat Phys*. 2010;6:988-992. Available from: <https://www.nature.com/articles/nphys1821>
72. Kolchinsky A, Wolpert DH. Dependence of dissipation on the initial distribution over states. *J Stat Mech Theory Exp*. 2017;2017:083202. Available from: <https://iopscience.iop.org/article/10.1088/1742-5468/aa7ee1/pdf>
73. Melanson D, Abu Khater M, Aifer M, Donatella K, Gordon MH, Ahle T, et al. Thermodynamic computing system for AI applications. *Nat Commun*. 2025;16:3757. Available from: <https://www.nature.com/articles/s41467-025-59011-x>
74. Wolpert DH, Kolchinsky A. Thermodynamics of computing with circuits. *New J Phys*. 2020;22:063047. Available from: <https://iopscience.iop.org/article/10.1088/1367-2630/ab82b8>
75. Yadav A, Yousef M, Wolpert DH. Minimal thermodynamic cost of computing with circuits. *arXiv [Preprint]*. 2025;2504.04031. Available from: <https://arxiv.org/html/2504.04031v1>
76. Yadav A, Caravelli F, Wolpert DH. Mismatch cost of computing: from circuits to algorithms. *arXiv [Preprint]*. 2024;2411.16088. Available from: <https://arxiv.org/html/2411.16088v1>
77. Manzano G, Fagas G, Esposito M. Quantum-classical correspondence principle for work distributions in a stochastic Jarzynski equality. *Phys Rev E*. 2018;98:052130.
78. Riechers PM, Gu M. Initial-state dependence of thermodynamic dissipation for any quantum process. *Phys Rev E*. 2021;103:042145. Available from: <https://journals.aps.org/pre/abstract/10.1103/PhysRevE.103.042145>
79. Horowitz M. Computing's energy problem (and what we can do about it). In: 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC). 2014. p. 10-14. Available from: [https://www.researchgate.net/publication/271463146\\_11\\_Computing's\\_energy\\_problem\\_and\\_what\\_we\\_can\\_do\\_about\\_it](https://www.researchgate.net/publication/271463146_11_Computing's_energy_problem_and_what_we_can_do_about_it)
80. Rabaey JM, Chandrakasan A, Nikolic B. Digital integrated circuits: a design perspective. 2nd ed. Pearson; 2003.
81. Theis TN, Wong HSP. The end of Moore's law: a new beginning for information technology. *Comput Sci Eng*. 2017;19(2):41-50. Available from: <https://ieeexplore.ieee.org/document/7878935>
82. Zhirnov VV, Cavin RK, Hutchby GI, Bourianoff GI. Limits to binary logic switch scaling—a gedanken model. *Proc IEEE*. 2003;91(11):1934-1939. Available from: <https://ieeexplore.ieee.org/document/1240081>
83. NVIDIA Corporation. NVIDIA H100 Tensor Core GPU architecture whitepaper. NVIDIA Technical Whitepaper. 2022.
84. Korthikanti VA, Casper J, Lym S, McAfee L, Andersch M, Shoeybi M, et al. Reducing activation recomputation in large transformer models. *arXiv [Preprint]*. 2022;2205.05198. Available from: <https://arxiv.org/abs/2205.05198>
85. Davies M, Srinivasa N, Lin TH, China G, Cao Y, Choday SH, et al. Loihi: a neuromorphic manycore processor with on-chip learning. *IEEE Micro*. 2018;38(1):82-99.
86. Davies M, Wild A, Orchard G, Sandamirskaya Y, Guerra GAF, Joshi P, et al. Advancing neuromorphic computing with Loihi: a survey of results and outlook. *Proc IEEE*. 2021;109(5):911-934. Available from: <https://ieeexplore.ieee.org/document/9395703>
87. Intel Corporation. Intel builds world's largest neuromorphic system to enable more sustainable AI. Intel Newsroom press release. 2024.
88. Nowotny T, et al. Eventprop training for efficient neuromorphic applications. *arXiv [Preprint]*. 2025;2503.04341. Available from: <https://arxiv.org/abs/2503.04341>
89. Ielmini D, Wong HSP. In-memory computing with resistive switching devices. *Nat Electron*. 2018;1:333-343. Available from: <https://www.nature.com/articles/s41928-018-0092-2>
90. Sebastian A, Le Gallo M, Khaddam-Aljameh R, Eleftheriou E. Memory devices and applications for in-memory computing. *Nat Nanotechnol*. 2020;15:529-544. Available from: <https://www.nature.com/articles/s41565-020-0655-z>
91. Wan W, Kubendran R, Schaefer C, Eryilmaz SB, Zhang W, Wu D, et al. A compute-in-memory chip based on resistive random-access memory. *Nature*. 2022;608:504-512. Available from: <https://www.nature.com/articles/s41586-022-04992-8>
92. Xue CX, et al. Embedded 1Mb ReRAM-based computing-in-memory macro with multibit input and weight for CNN-based AI edge processors. *IEEE J Solid State Circuits*. 2020;55:203-215. Available from: <https://ieeexplore.ieee.org/document/8910375>
93. Shen Y, Harris NC, Skirlo S, Prabhu M, Baehr-Jones T, Hochberg M, et al. Deep learning with coherent nanophotonic circuits. *Nat Photonics*. 2017;11:441-446. Available from: <https://www.nature.com/articles/nphoton.2017.93>
94. Wetzstein G, Ozcan A, Gigan S, Fan S, Englund M, Soljačić M, et al. Inference in artificial intelligence with deep optics and photonics. *Nature*. 2020;588:39-47. Available from: <https://pubmed.ncbi.nlm.nih.gov/33268862/>

95. Frank MP. The physical limits of computing. Comput Sci Eng. 2002;4(3):16-26. Available from: <https://ieeexplore.ieee.org/abstract/document/998637>

96. Manipatruni S, Nikonov DE, Lin CC, Gosavi TA, Liu H, Prasad B, et al. Scalable energy-efficient magnetoelectric spin-orbit logic. Nature. 2019;565:35-42. Available from: <https://www.nature.com/articles/s41586-018-0770-2>

### Discover a bigger Impact and Visibility of your article publication with Peertechz Publications

#### Highlights

- ❖ Signatory publisher of ORCID
- ❖ Signatory Publisher of DORA (San Francisco Declaration on Research Assessment)
- ❖ Articles archived in worlds' renowned service providers such as Portico, CNKI, AGRIS, TDNet, Base (Bielefeld University Library), CrossRef, Scilit, J-Gate etc.
- ❖ Journals indexed in ICMJE, SHERPA/ROMEO, Google Scholar etc.
- ❖ OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting)
- ❖ Dedicated Editorial Board for every journal
- ❖ Accurate and rapid peer-review process
- ❖ Increased citations of published articles through promotions
- ❖ Reduced timeline for article publication

**Submit your articles and experience a new surge in publication services**

<https://www.peertechzpublications.org/submission>

*Peertechz journals wishes everlasting success in your every endeavours.*