Towards a Thermodynamic Perspective on Learning Systems

Lucca Forrest*
Department of Physics and Mathematics, University of North Carolina at Chapel Hill and Department of Physics, University of Oxford
(Dated: November 27, 2025)

Modern machine learning systems transform large amounts of physical energy into structured models that reduce predictive uncertainty. Motivated by classical results in the thermodynamics of computation and information, we explore a phenomenological framework—which we term *thermointelligence*—that draws analogies between thermodynamic quantities (energy, entropy, temperature) and coarse-grained properties of learning dynamics. Building on Shannon entropy, Landauer's principle, and Sutton's "Bitter Lesson," we propose candidate notions of *intelligence entropy* and *intelligence temperature* that aim to quantify the energetic efficiency of learning. We show how these quantities can be constructed from standard training statistics (uncertainty, loss, parameter updates, and gradients), and we outline how thermodynamic analogies extend across microscopic (bit-level), mesoscopic (algorithmic), and macroscopic (data-center) scales. The goal of this work is not to claim new physical laws, but to provide a coherent language and a set of guiding principles that may help organize future theoretical and empirical studies of the energetic limits of artificial learning systems.

I. INTRODUCTION

Rapid progress in artificial intelligence has been closely tied to increases in available computation. Sutton's "Bitter Lesson" [4] emphasizes that, over decades, the most robust improvements have come less from handcrafted structure and more from scaling compute, data, and relatively simple algorithms. This observation invites a physical question: if computation is ultimately a thermodynamic process, how does the energy consumed by learning machines relate to the structure they acquire?

The thermodynamics of computation has a long history. Shannon's information entropy [1] provided a quantitative measure of uncertainty, while Landauer showed that logically irreversible operations, such as bit erasure, have a minimum energetic cost [2]. Subsequent work by Bennett and others developed a broader theory of reversible computation and its thermodynamic implications [3]. More recently, there has been growing interest in the energy footprint of large-scale machine learning systems and data centers.

In this paper we explore a speculative but structured analogy between thermodynamics and learning dynamics, which we call *thermointelligence*. The central idea is to treat a learning system as a physical engine that converts supplied energy into two outputs: (i) dissipated heat, and (ii) reduced uncertainty or increased structure in its internal state and behavior. We do not propose new fundamental laws of physics; instead, we investigate whether thermodynamic language can provide useful organizing principles for describing the efficiency and directionality of learning.

We focus on three aims:

1. To define candidate quantities: *computational energy*, *intelligence entropy*, and *intelligence temperature*—that are physically motivated and, at least in principle, estimable from training logs and hardware measurements.

2. To formulate analogies with the zeroth, first, second, and third laws of thermodynamics in the context of learning, while explicitly treating them as heuristic principles rather than exact laws.

3. To outline how these ideas extend across microscopic (bit-level), mesoscopic (algorithmic and architectural), and macroscopic (data-center) scales, highlighting conceptual connections and limitations.

Throughout, we keep the discussion at a theoretical level. In particular, we emphasize that in practical gradient-based optimization it is difficult to decompose energy usage into "useful work" and "waste" at the level of individual updates, since gradients are noisy, losses are non-smooth, and learning trajectories are path-dependent. Our proposals should therefore be viewed as coarse-grained descriptors and targets for future empirical work, rather than as operational definitions that can already be cleanly measured.

II. BACKGROUND AND RELATED WORK

A. Thermodynamics and Information

In statistical mechanics, the Boltzmann-Gibbs entropy of a system with microstates i and corresponding probabilities p_i is

$$S_{\text{th}} = -k_B \sum_i p_i \ln p_i, \tag{1}$$

where k_B is Boltzmann's constant. Shannon introduced an analogous measure of uncertainty in communication

* lucca_forrest@unc.edu

theory,

$$H = -\sum_i p_i \log_2 p_i, \tag{2}$$

with units of bits [1]. The two quantities are related by

$$S_{\rm th} = k_B \ln 2\, H, \tag{3}$$

which provides a direct bridge between information-theoretic and thermodynamic descriptions.

A key insight is that logical uncertainty, quantified by H, can be realized physically as thermodynamic entropy $S_{\rm th}$ in an appropriately constructed device. This perspective underlies the thermodynamics of computation.

B. Landauer's Principle and Bit Erasure

Landauer showed that any logically irreversible operation, such as resetting a bit to a standard value, must dissipate at least

$$\Delta Q \geq k_B T_{\rm env} \ln 2 \tag{4}$$

of heat into an environment at temperature $T_{\rm env}$ [2]. A single bit can be represented physically as a system with two stable states, for example a particle in a double-well potential with one well corresponding to logical 0 and the other to logical 1.

If the system is in an unknown logical state, $p_0 = p_1 = \frac{1}{2}$, then its thermodynamic entropy is

$$S_i = -k_B \left(\tfrac{1}{2}\ln\tfrac{1}{2} + \tfrac{1}{2}\ln\tfrac{1}{2}\right) = k_B \ln 2. \tag{5}$$

If the bit is known with certainty to be in a definite state, say $p_0 = 1$, $p_1 = 0$, then

$$S_f = -k_B(1\ln 1 + 0\ln 0) = 0. \tag{6}$$

Resetting the memory to a fixed value regardless of its initial state therefore reduces the system's entropy by

$$\Delta S = S_f - S_i = -k_B \ln 2. \tag{7}$$

By the second law, the total entropy of the universe cannot decrease, so the environment must absorb at least the same amount of entropy,

$$\Delta S_{\rm env} \geq +k_B \ln 2, \tag{8}$$

which is related to the heat transfer by

$$\Delta S_{\rm env} = \frac{\Delta Q}{T_{\rm env}}. \tag{9}$$

Combining these expressions yields Landauer's bound. Reversible computing attempts to avoid this cost by designing logic operations that are, in principle, thermodynamically reversible [3].

C. Energy Use in Machine Learning

Large-scale learning systems consume substantial electrical power during training and inference. This consumption is mediated by physical hardware (CPUs, GPUs, TPUs, memory, cooling systems) with nonzero leakage, switching losses, and overhead. Recent work has begun to quantify the energy and carbon footprint of training state-of-the-art models, particularly in natural language processing and vision.

From a thermodynamic perspective, only a small fraction of this energy is associated with irreducible costs like bit erasure; most is due to practical engineering constraints (device inefficiencies, communication overhead, cooling). Nonetheless, Landauer's principle provides a fundamental lower bound on the energy cost of manipulating information, and it motivates the search for more refined measures of how much of the consumed energy contributes to persistent structure in the model.

III. A THERMODYNAMIC ANALOGY FOR LEARNING SYSTEMS

A. Thermodynamic Preliminaries

In equilibrium thermodynamics, the zeroth, first, second, and third laws are expressed in terms of temperature $T_{\rm th}$, internal energy $E_{\rm th}$, and entropy $S_{\rm th}$. In particular,

$$T_{\rm th} = \left(\frac{\partial E_{\rm th}}{\partial S_{\rm th}}\right)_{V,N} \tag{10}$$

defines temperature as the rate of change of energy with respect to entropy at fixed volume and particle number, and the first law can be written in differential form as

$$dE_{\rm th} = \delta Q_{\rm th} + \delta W_{\rm th}. \tag{11}$$

We do not attempt to place learning systems in literal equilibrium or to assign them exact thermodynamic state variables. Instead, we propose effective quantities that borrow this structure and may be meaningful at appropriate coarse-grained levels.

B. Computational Energy, Intelligence Entropy, and Intelligence Temperature

We consider a learning system (model + hardware) that consumes physical energy during training. Over a window of time or number of update steps, we denote by

- E the total *computational energy* supplied (J), e.g., measured by hardware power and training time;
- Q the energy ultimately dissipated as heat into the environment (J);

- W the remainder, which we interpret as energy invested in forming, updating, or maintaining structured internal representations (J).

In practice, only E is straightforward to measure; decomposing E into Q and W is non-trivial and will be discussed later.

We introduce an *intelligence entropy* S with units of bits, intended as a coarse-grained measure of a model's structural disorder or uncertainty. We deliberately keep S abstract, and in Sec. IV we describe several concrete proxies that may be useful in different regimes.

Motivated by the thermodynamic definition of temperature, we define an effective *intelligence temperature*

$$T = \frac{\partial E}{\partial S}, \tag{12}$$

with units of joules per bit (J/bit). Intuitively, T quantifies the marginal energetic cost of reducing intelligence entropy. High T corresponds to a regime where many joules are required to remove one bit of effective disorder, while low T indicates more efficient learning.

We stress that this definition is phenomenological. Unlike in equilibrium thermodynamics, where state variables are well defined and measurable, here S and T depend on how we choose to summarize the learning dynamics.

IV. CANDIDATE MEASURES OF INTELLIGENCE ENTROPY

In this section we outline several complementary quantities that can serve as proxies for intelligence entropy. Each can be computed from standard training logs and captures a different aspect of learning dynamics.

A. Uncertainty Entropy

The most direct information-theoretic quantity is the Shannon entropy of the model's predictive distribution. For a batch of N inputs, with model parameters θ and predictive probabilities $p_\theta^{(i)}(x)$ over classes x, we define

$$S_{\text{uncertainty}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{x} p_\theta^{(i)}(x)\log_2 p_\theta^{(i)}(x). \tag{13}$$

As training progresses and predictions become more confident, $S_{\text{uncertainty}}$ typically decreases on both training and validation data. This quantity tracks the model's residual uncertainty about its outputs.

B. Loss Entropy

For supervised learning, the cross-entropy loss over a batch provides a closely related measure,

$$S_{\text{loss}} = -\frac{1}{N}\sum_{i=1}^{N}\log_2 p_\theta(x_i), \tag{14}$$

where x_i denotes the correct label for sample i. When labels are accurate and the model's predictive distribution is well calibrated, S_{loss} approximates the conditional entropy of the labels given the inputs. Declines in S_{loss} therefore reflect improvements in predictive structure.

C. Weight-Space Entropy

Learning also reshapes the model's internal parameter configuration. For a batch update from step $t-1$ to t, with M parameters w_j, we define a measure of parameter-space motion,

$$S_{\text{weight}} = \frac{1}{M}\sum_{j=1}^{M}\log_2\left(1+\frac{|\Delta w_j|}{\epsilon}\right), \tag{15}$$

where $\Delta w_j = w_j^{(t)} - w_j^{(t-1)}$ and ϵ is a small constant to regularize small steps. This expression approximates the number of bits required to encode the change in parameters between steps. As optimization converges and updates shrink, $|\Delta w_j|$ decreases and so does S_{weight}.

D. Gradient Entropy

Finally, we can characterize the distribution of gradient magnitudes. Let g_j denote the gradient of the loss with respect to parameter w_j on a given batch, and define

$$G = \sum_{k=1}^{M}|g_k|, \qquad p_j = \frac{|g_j|}{G}. \tag{16}$$

We then define

$$S_{\text{grad}} = -G\sum_{j=1}^{M} p_j \log_2 p_j. \tag{17}$$

This couples the overall gradient scale G with the Shannon entropy of the normalized gradient distribution. High S_{grad} corresponds to diffuse, noisy updates spread across many parameters, while low S_{grad} reflects concentrated, structured updates focused on a subset of parameters.

E. Interpretation and Limitations

None of these measures is uniquely privileged. They emphasize different aspects of learning: predictive uncertainty, supervised loss, parameter motion, and gradient

structure. In a thermodynamic analogy, they all play the role of coarse-grained entropy-like quantities. A key open problem is to determine which combinations of these metrics (possibly together with others, such as information bottleneck measures [5]) best correlate with the energetic efficiency of learning in practice.

V. AN ENERGY–ENTROPY BALANCE FOR LEARNING

A. First-Law Analogy

The classical first law states that the total energy supplied to a system is partitioned into heat and work,

$$dE_{\text{th}} = \delta Q_{\text{th}} + \delta W_{\text{th}}. \tag{18}$$

By analogy, we propose that computational energy supplied to a learning system can be decomposed as

$$dE = \delta Q + \delta W, \tag{19}$$

where

- dE is the incremental computational energy input (J);
- δQ is the energy that ultimately becomes dissipated heat (J), e.g., resistive losses, cooling, idle cycles;
- δW is the energy associated with persistent changes in the model's internal structure (J).

If we express E as a function of intelligence entropy S, then by definition

$$T = \frac{\partial E}{\partial S}, \tag{20}$$

and an infinitesimal change in S contributes

$$\delta E_S = T\, dS. \tag{21}$$

Substituting into the first-law analogy yields

$$dE = \delta Q + T\, dS. \tag{22}$$

Equation (22) should be interpreted as a bookkeeping identity at the level of our effective variables: changes in computational energy are attributed either to dissipation or to changes in intelligence entropy.

B. Apparent Fluctuations and "Lucky" Learning Events

In practice, learning dynamics are stochastic and nonmonotonic. Losses can briefly decrease without obvious additional energy input (e.g., favorable mini-batch sampling), and training sometimes discovers better minima after periods of apparent stagnation. At the level of S, one may observe episodes with

$$dE \approx 0, \qquad dS < 0, \tag{23}$$

which resemble spontaneous entropy reductions.

In analogy with fluctuation theorems in statistical mechanics, one expects the probability of large negative entropy changes to be exponentially suppressed,

$$P(\Delta S) \sim e^{-\Delta S / k_{\text{eff}}}, \tag{24}$$

for some effective constant k_{eff} that depends on the learning algorithm and data distribution. Over long timescales and many updates, such fluctuations average out, and the energy–entropy balance holds in expectation,

$$\langle dE \rangle = \langle \delta Q + T\, dS \rangle. \tag{25}$$

A rigorous formulation of these ideas would require specifying the stochastic process governing updates and is left for future work.

C. Zeroth-, Second-, and Third-Law Analogies

We briefly sketch analogies with the other thermodynamic laws, emphasizing their heuristic nature.

a. Zeroth-law analogy. If we consider two models A and B that exchange knowledge (e.g., via distillation, parameter averaging, or shared gradients) through a third system, we may say they are in a form of "learning equilibrium" when their marginal learning efficiency is equal:

$$T_A = T_B. \tag{26}$$

At this point, there is no net beneficial transfer of structure between them per unit energy. Unlike physical equilibrium, this is an abstraction over learning curves rather than a literal thermal contact.

b. Second-law analogy. Reducing intelligence entropy in a model (making predictions more structured and less random) requires increasing entropy elsewhere, for example in the environment, cooling system, or power infrastructure. At a coarse level, one may write

$$\Delta S + \Delta S_{\text{env}} \geq 0, \tag{27}$$

where ΔS_{env} represents the thermodynamic entropy accumulated by the environment due to computational dissipation. This expresses the intuition that learning has an inherent arrow: persistent gains in structured intelligence have an energetic shadow.

c. Third-law analogy. As $T \to 0$, it becomes progressively harder to extract further reductions in S. In an overparameterized model that has largely converged, gradients vanish, weight updates shrink, and entropy-based measures $S_{\text{uncertainty}}$, S_{loss}, S_{weight}, S_{grad} approach lower bounds. In this regime, any further improvement requires disproportionately large computational effort, reminiscent of the thermodynamic difficulty of approaching absolute zero.

VI. MULTISCALE IMPLICATIONS: FROM BITS TO DATA CENTERS

A. Microscopic: Bit-Level Computation

At the microscopic scale, thermointelligence connects to the physical cost of elementary operations. Bit erasure, as described by Landauer's principle, is an example where logical irreversibility directly entails a thermodynamic cost,

$$\Delta E \geq k_B T_{\text{env}} \ln 2. \tag{28}$$

Most practical logic operations dissipate far more energy than this bound due to device-level inefficiencies (switching losses, leakage currents, etc.). From the standpoint of intelligence entropy, many such operations correspond to pure information handling without net learning:

$$dS \approx 0, \qquad dE \approx \delta Q. \tag{29}$$

The first-law analogy then reduces to $dE = \delta Q$, indicating that all the energy goes into dissipation with no change in S. This highlights that not all computation contributes to learning; thermointelligence is concerned with the subset of operations that push S downward in a sustained way.

B. Mesoscopic: Algorithmic and Architectural Dynamics

At the algorithmic level, gradient-based optimization, backpropagation, reinforcement learning updates, and attention mechanisms can be viewed as microstate transitions in a high-dimensional energy landscape. The first-law analogy, Eq. (22), becomes

$$dE_{\text{train}} = \delta Q_{\text{comp}} + T\, dS, \tag{30}$$

where dE_{train} is the energy consumed during training, δQ_{comp} is the portion lost as hardware heat, and $T\, dS$ encodes the energetic contribution to changes in intelligence entropy.

Here, a major limitation of the analogy appears. In practice gradient descent is noisy and path-dependent, loss surfaces are non-smooth and high dimensional, and individual updates can increase or decrease loss, and their long-term contribution to generalization is difficult to attribute.

As a result, it is not currently possible to cleanly classify each increment of energy as "positive work" (useful learning) or "negative work" (wasted effort). The separation between Q and W is, at best, a coarse-grained construct over many updates, not a step-by-step diagnostic.

Nonetheless, the framework suggests that one can empirically compare different algorithms, architectures, or schedules by evaluating how much energy they consume to achieve a given reduction in one of the entropy proxies $S_{\text{uncertainty}}$, S_{loss}, S_{weight}, or S_{grad}. This leads to a notion of *learning efficiency* measured in bits per joule.

C. Macroscopic: Data-Center and System-Level View

At the macroscopic scale, an entire data center can be treated as an open thermodynamic system that exchanges energy and entropy with its environment. Electrical power is drawn from a grid, converted to computation and heat, and rejected via cooling infrastructure. One may define an effective *intelligence flux density* as the rate at which the system reduces intelligence entropy (across all models it trains) per unit energy or per unit time.

While such a quantity is not yet standard, the thermointelligence perspective suggests thinking about:

- How much predictable structure is produced (e.g., in terms of trained model performance or entropy proxies) per joule consumed.
- How this efficiency depends on hardware design, cooling technology, scheduling policies, and model architectures.
- Trade-offs between training once at high cost and reusing models many times (amortizing the initial thermodynamic investment).

VII. DISCUSSION AND OUTLOOK

We have outlined a conceptual framework that applies thermodynamic language to learning systems, introducing the ideas of computational energy, intelligence entropy, and intelligence temperature. The analogies with the classical thermodynamic laws are deliberately modest: they serve as organizing metaphors and potential targets for more precise formulations rather than as claims of exact physical laws governing intelligence.

Several directions for future work are suggested by this perspective:

- **Empirical calibration.** Measure power consumption of real training runs and correlate it with changes in entropy-based metrics to estimate effective T and learning efficiency (bits per joule).
- **Algorithm comparison.** Compare different optimizers and architectures in terms of their energy–entropy trajectories, not just final accuracy.
- **Theoretical refinement.** Connect this phenomenological framework to existing work on the thermodynamics of information processing, stochastic thermodynamics, and information bottleneck theory.

- **Limits of the analogy.** Identify conditions under which the thermodynamic language breaks down, for example in highly non-stationary or adversarial environments.

A central caveat is the difficulty of decomposing energy into "useful work" and "waste heat" in real learning systems. Gradient descent is non-smooth and stochastic, and updates cannot be cleanly labeled as positive or negative work at the microscopic level. For this reason, the quantities Q and W should currently be viewed as coarse-grained, statistical descriptors, perhaps defined only over ensembles of runs or large segments of training.

Despite these limitations, the thermodynamic viewpoint remains appealing. It emphasizes that learning is not free: any persistent reduction in uncertainty or increase in structure must ultimately be paid for with physical energy and an accompanying increase in environmental entropy. Making this trade-off explicit may help guide the design of more efficient algorithms, architectures, and hardware, and may clarify the ultimate physical limits of artificial intelligence systems.

[1] C. E. Shannon, "A Mathematical Theory of Communication," Bell System Technical Journal **27**, 379–423, 623–656 (1948).

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

[3] C. H. Bennett, "The Thermodynamics of Computation—A Review," Int. J. Theor. Phys. **21**, 905–940 (1982).

[4] R. S. Sutton, "The Bitter Lesson," (2019), available at `http://www.incompleteideas.net/IncIdeas/BitterLesson.html`.

[5] N. Tishby, F. C. Pereira, and W. Bialek, "The Information Bottleneck Method," in *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing* (1999).