IPI Letters

ISSN 2976 - 730X
IPI Letters 2026, Vol 4 (2):13-20
https://doi.org/10.59973/ipil.335

Received: 2026-02-12
Accepted: 2026-03-16
Published: 2026-03-20

Article

Local Entropy Inversion in Large-Scale AI Systems: Landauer Bounds on Algorithmic Compression

Boris Kriger[1,2,*]

[1] Information Physics Institute, Gosport, Hampshire, United Kingdom, www.informationphysicsinstitute.org
[2] Institute of Integrative and Interdisciplinary Research Toronto, Ontario, Canada, https://interdisciplinary-research.institute/

*Corresponding author: boris.kriger@informationphysicsinstitute.net

Abstract - We apply Landauer's principle to the training of large language models (LLMs), framing the process as a physically irreversible compression of high-entropy data distributions into low-entropy structured representations stored in model weights. This yields a lower bound on the minimum energy required for AI training, expressed in terms of the information-theoretic compression achieved. Empirical analysis of contemporary AI systems - GPT-3, PaLM, and LLaMA-2 - reveals that current implementations operate approximately 10^{21} times above this Landauer limit. We introduce a *demon efficiency* metric to quantify this gap and examine how it varies across systems and baseline assumptions. We discuss an instructive analogy between LLM training and Maxwell's demon that provides physical intuition for the entropy-reducing character of the training process. We present a sensitivity analysis showing that while the absolute value of the efficiency metric depends on the choice of entropy baseline, the order-of-magnitude gap to the Landauer limit is robust across reasonable choices. These results provide a physical perspective on the energy requirements of artificial intelligence, though we emphasize that the Landauer bound is a direct consequence of well-established thermodynamic principles rather than a new theoretical result.

Keywords - Information thermodynamics; Landauer's principle; Large language models; Algorithmic compression; Minimum description length; Thermodynamic efficiency.

1 Introduction

The training of large language models (LLMs) involves a process that admits a natural thermodynamic interpretation: the conversion of high-entropy data distributions into low-entropy structured representations stored in model weights. When a system such as GPT-3 or PaLM processes hundreds of billions of tokens and encodes the resulting statistical regularities in a set of neural network parameters, it performs a physically irreversible operation whose energy cost is bounded from below by fundamental thermodynamic principles.

The connection between information processing and thermodynamics was established through the resolution of Maxwell's thought experiment [1] by Szilard [2], Landauer [3], and Bennett [4]. Landauer's principle states that erasing one bit of information in a system at temperature T requires a minimum heat dissipation of

$$Q_{\min} = k_B T \ln 2 \approx 2.87 \times 10^{-21} \text{ J at } T = 300 \text{ K} \tag{1}$$

where k_B is Boltzmann's constant. This result, experimentally confirmed by Bérut et al. [5], connects the logical irreversibility of computation to physical entropy production. Vopson [6] proposed a speculative mass-energy-information equivalence principle suggesting that stored information may possess an effective mass $m_{\text{bit}} = k_B T \ln 2/c^2$. While this proposal remains under discussion, it illustrates the broader perspective that information may have physical manifestations beyond the established Landauer energy cost.

A typical large language model training run involves processing 10^{12}-10^{13} tokens of training data, performing 10^{23}-10^{25} floating-point operations, and consuming 10^{13}-10^{14} joules of electrical energy, corresponding to estimated carbon emissions on the order of 10^5-10^6 kg CO_2 equivalent [7]. The resulting model stores learned regularities in roughly 10^{11}-10^{12} bits of model weights. It is important to note that this is not compression in the archival sense-the training data cannot be reconstructed from the model weights. Rather, the model learns a conditional probability distribution that captures statistical structure in the data, achieving lower cross-entropy than a native baseline.

This paper applies Landauer's principle to this process to obtain a lower bound on the minimum energy required for AI training (Section 2), validates this bound against publicly available data from contemporary AI systems (Section 3), develops an instructive analogy between LLM training and Maxwell's demon (Section 4), and discusses implications and limitations (Section 5).

2 Theoretical Framework

2.1 Information-Theoretic Preliminaries

Let $\mathcal{X}$ denote a finite alphabet and $\mathcal{X}^*$ the set of all finite strings over $\mathcal{X}$. For a probability distribution P over $\mathcal{X}^*$, the Shannon entropy is

$$H(P) = -\sum_{x \in \mathcal{X}^*} P(x) \log_2 P(x) \tag{2}$$

and the Kullback-Leibler divergence between distributions P and Q is

$$D_{\text{KL}}(P\|Q) = \sum_x P(x) \log_2 \frac{P(x)}{Q(x)} \geq 0 \tag{3}$$

with equality if and only if $P = Q$ almost everywhere. The minimum description length (MDL) principle [8,9] selects the model M that minimizes the total code length $L(M) + L(D|M)$, where $L(M)$ is the description length of the model and $L(D|M)$ is the description length of data D given model M.

2.2 Thermodynamics of Computation

Following Landauer [3] and Bennett [4], we state the connection between logical and physical irreversibility.

Theorem 1 (Landauer's Principle - Extended Form)**.** *Any logically irreversible computational operation that transforms a system from a state of entropy S_1 to a state of entropy $S_2 < S_1$ must dissipate heat to the environment of at least*

$$Q_{\text{min}} = T(S_1 - S_2) = k_B T \ln 2 \cdot \Delta I \tag{4}$$

where $\Delta I = (S_1 - S_2)/(k_B \ln 2)$ is the information erased in bits.

Proof. Consider a system coupled to a thermal reservoir at temperature T. The second law of thermodynamics requires $\Delta S_{\text{total}} = \Delta S_{\text{system}} + \Delta S_{\text{reservoir}} \geq 0$. For the reservoir, $\Delta S_{\text{reservoir}} = Q/T$ where Q is the heat absorbed by the reservoir. If the system's entropy decreases by $\Delta S_{\text{system}} = S_2 - S_1 < 0$, then $Q \geq T(S_1 - S_2) = k_B T \ln 2 \cdot \Delta I$. This minimum is achieved only for quasi-static, reversible processes. □

2.3 AI Training as Information Compression

We now formalize the AI training process in thermodynamic terms. Let $\mathcal{D}_{\text{train}} = \{x_1, x_2, \ldots, x_N\}$ be the training dataset consisting of N tokens from vocabulary V with $|V| = \mathcal{V}$.

Definition 2 (Model Cross-Entropy)**.** For a language model M with distribution P_M, the cross-entropy on the training data is

$$H_{\text{cross}}(P_{\text{data}}, P_M) = -\frac{1}{N}\sum_{i=1}^{N} \log_2 P_M(x_i|x_{<i}) \tag{5}$$

The cross-entropy relates to perplexity as $\text{PPL} = 2^{H_{\text{cross}}}$. State-of-the-art LLMs achieve $H_{\text{cross}} \approx 3.5$-$4.5$ bits per token on the WikiText-103 benchmark [10], corresponding to perplexity values of approximately 11-23.

To apply Landauer's principle, we require a baseline entropy H_{prior} representing the system's initial state before training. The choice of this baseline is not unique and constitutes a modeling assumption that affects the absolute value of the resulting efficiency metric. We consider three natural choices and present results for each:

(a) *Uniform baseline*: $H_{\text{uniform}} = \log_2 \mathcal{V} \approx 15$-$17$ bits/token for typical vocabulary sizes $\mathcal{V} \approx 32{,}000$–$128{,}000$. This represents maximum ignorance but is not physically realized by any practical model, since softmax normalization and architectural biases ensure that even random initializations produce non-uniform output distributions.

(b) *Untrained model baseline*: $H_{\text{random}} \approx 10$-$12$ bits/token, measured empirically as the cross-entropy of a randomly initialized transformer on the training data. This is the most physically meaningful baseline, as it represents the actual starting state of the training process.

(c) *Unigram baseline*: $H_{\text{unigram}} \approx 10$-$12$ bits/token, representing the cross-entropy achievable by a simple frequency-based model with no contextual information.

The near-coincidence of baselines (b) and (c) reflects the fact that randomly initialized transformers, before training, perform comparably to simple frequency-based models. Throughout this paper, we adopt $H_{\text{prior}} \approx 10$ bits/token (the untrained model baseline) as our primary reference, while reporting the sensitivity of our results to this choice. We emphasise that the effective compression $\Delta H = H_{\text{prior}} - H_{\text{cross}}$ measures the improvement in predictive cross-entropy achieved by training, not a reduction in physical entropy of the data itself.

2.4 Application of the Landauer Bound to AI Training

Applying Theorem 1 to the training process yields a lower bound on the required energy.

Proposition 3 (Landauer Bound Applied to AI Training)**.** *For a model trained on N tokens that achieves cross-entropy reduction $\Delta H = H_{\text{prior}} - H_{\text{cross}}$ bits per token, the Landauer limit on the minimum energy dissipated is*

$$E_{\min} = N \cdot k_B T \ln 2 \cdot \Delta H \tag{6}$$

This follows directly from Theorem 1 by identifying $\Delta I = N \cdot \Delta H$ as the total number of bits of predictive uncertainty eliminated during training. We note that this is a straightforward application of Landauer's principle rather than a new theoretical result; its value lies in making explicit the thermodynamic lower bound and comparing it quantitatively with actual energy consumption.

Corollary 4 (Fundamental Efficiency Limit)**.** *The thermodynamic efficiency of AI training, defined as the ratio of the Landauer minimum to actual energy consumed, is*

$$\eta = \frac{E_{\min}}{E_{\text{actual}}} = \frac{N \cdot k_B T \ln 2 \cdot \Delta H}{E_{\text{actual}}} \tag{7}$$

For contemporary AI systems operating at $T \approx 350 \pm 30$ K (the sustained junction temperature of GPU dies under training workloads; using the conventional $T = 300$ K would reduce $E_{\min}$ by approximately 15%, making the gap to actual consumption marginally larger) processing $N \approx 10^{12}$ tokens with $\Delta H \approx 6 \pm 1$ bits per token under the untrained-model baseline, the minimum energy evaluates to

$$E_{\min} \approx 10^{12} \times 1.38 \times 10^{-23} \times 350 \times 0.693 \times 6 \approx 2.0 \times 10^{-8} \text{ J} = 20 \text{ nJ} \tag{8}$$

with uncertainty range $E_{\min} \in [1.4, 2.8] \times 10^{-8}$ J due to variations in T and ΔH. Under the uniform baseline ($\Delta H \approx 12$ bits/token), this doubles to approximately 40 nJ; under a bigram baseline ($\Delta H \approx 3$ bits/token, as shown in Table 2), it halves to approximately 10 nJ. In all cases, the Landauer minimum remains on the order of 10^{-8} J.

Actual energy consumption for training such models is $E_{\text{actual}} \approx 10^{13}$-$10^{14}$ J [7]. Taking $E_{\text{actual}} \approx 10^{13}$ J as a representative mid-range estimate yields

$$\eta_{\text{current}} \approx \frac{2 \times 10^{-8}}{10^{13}} \approx 2 \times 10^{-21} \tag{9}$$

with plausible range $\eta_{\text{current}} \in [5 \times 10^{-22}, 5 \times 10^{-21}]$ across different systems and baseline choices.

The ~ 21 order-of-magnitude gap between theoretical minimum and actual energy consumption is unsurprising given the known inefficiencies of current hardware and algorithms. Following the analysis of Frank [11], we can decompose this gap approximately as: transistor switching losses contribute a factor of $\sim 10^6$ above Landauer per logic operation; memory access and data movement add $\sim 10^3$-10^4 overhead; the redundant computation inherent in gradient descent (multiple passes over the data, trial-and-error optimization) contributes $\sim 10^3$-10^6; and cooling infrastructure adds a factor of $\sim$ 1.1-1.5 (PUE). Multiplied together, these factors account for the observed gap. This decomposition, while approximate, suggests that the largest gains would come from more energy-efficient switching (e.g., reversible or adiabatic logic) and more data-efficient training algorithms.

We note an important subtlety: the quantity $\Delta I = N \cdot \Delta H$ represents the *net* information stored in the trained model, but the *total* information erased during the training process is vastly larger. Each gradient descent step involves logically irreversible operations - intermediate activations are computed and discarded, gradients are accumulated and consumed, and optimizer states are repeatedly overwritten. Over the full course of training (typically 10^5-10^6 gradient steps, each involving the full forward and backward pass over a mini-batch), the cumulative number of bits erased may exceed ΔI by many orders of magnitude. The true Landauer lower bound, accounting for all irreversible operations performed during training rather than only the net result, would therefore be substantially higher than the value given by Proposition 3. Our bound should thus be understood as a *lower* bound on the lower bound - the absolute thermodynamic minimum corresponding to an idealized, perfectly efficient process that achieves the same net compression in a single reversible step.

3 Empirical Analysis

3.1 Data Collection and Methodology

We analyze publicly available data on training costs and performance for three AI systems with well-documented training details: GPT-3 [12], LLaMA-2 [13], and PaLM [14]. Training energy is estimated from

$$E_{\text{train}} = \text{FLOPs} \times \frac{\text{J}}{\text{FLOP}} \times \text{PUE} \tag{10}$$

where typical values are 10–50 pJ/FLOP for the relevant GPU generations (V100, A100) and PUE of 1.1-1.5 for efficient data centers. FLOPs counts are taken from the original publications where available.

We define the *demon efficiency* η_D as the ratio of the Landauer minimum energy to the actual energy consumed:

$$\eta_D = \frac{E_{\min}}{E_{\text{actual}}} = \frac{N \cdot k_B T \ln 2 \cdot \Delta H}{E_{\text{actual}}} \tag{11}$$

A perfect thermodynamic process achieves $\eta_D = 1$. The thermodynamic cost per bit of cross-entropy reduction is $C_{\text{AI}} = E_{\text{actual}}/(N \cdot \Delta H)$.

3.2 Results

Table 1 summarizes the thermodynamic characteristics of the analyzed systems. Perplexity values are reported on the WikiText-103 benchmark. Energy estimates are derived from FLOPs counts and GPU specifications reported in the original publications [12,13,14], with PUE = 1.2 applied uniformly. All efficiency values use the untrained-model baseline ($H_{\text{prior}} \approx 10$ bits/token).

Table 1: Thermodynamic characteristics of selected AI systems. All values use the untrained-model baseline ($H_{\text{prior}} \approx 10$ bits/token). Uncertainties reflect estimated ±30% on energy and ±0.3 bits/token on H_{cross}.

Model	Params (10^9)	Tokens (10^{12})	Energy (MWh)	PPL (WT-103)	C_{AI} (J/bit)	η_D (10^{-21})
GPT-3	175	0.3	1,287 ± 400	20.5	2.8 ± 0.9	1.1 ± 0.4
PaLM	540	0.78	3,400 ± 1,000	15.3	2.1 ± 0.7	1.4 ± 0.5
LLaMA-2 70B	70	2.0	685 ± 200	16.9	0.22 ± 0.07	14 ± 4

Table 2 shows the sensitivity of the demon efficiency to baseline choice for GPT-3. While the absolute value of η_D varies by a factor of $\sim$ 2-3× across baselines, the order of magnitude ($\sim 10^{-21}$) is stable, confirming that the gap to the Landauer limit is robust.

Table 2: Sensitivity of GPT-3 demon efficiency to baseline entropy choice.

Baseline	H_{prior} (bits/token)	ΔH (bits/token)	η_D (10^{-21})
Uniform ($\log_2 \mathcal{V}$)	~ 15	~ 10.6	~ 2.3
Untrained model	~ 10	~ 5.6	~ 1.1
Unigram	~ 10	~ 5.6	~ 1.1
Bigram	~ 7.5	~ 3.1	~ 0.6

Several patterns emerge. First, all systems operate at demon efficiencies of order 10^{-21}, confirming the vast gap between current practice and the Landauer limit regardless of baseline choice. Second, LLaMA-2 70B achieves roughly an order of magnitude higher efficiency than GPT-3 and PaLM, consistent with the compute-optimal scaling analysis of Hoffmann et al. [15], which showed that training smaller models on more data is more efficient than scaling model size alone. Third, thermodynamic costs per bit range from approximately 0.22 to 2.8 J/bit - all roughly 10^{21} times the Landauer minimum of $\sim 3 \times 10^{-21}$ J/bit.

For comparison, biological neural computation achieves approximately 10^{-12}-10^{-10} J per synaptic operation [16], and estimates of the energy cost of learning in the human brain are on the order of 10^{-9}–10^{-8} J/bit [17]. Current AI systems are therefore 10^9-10^{10} times less efficient than biological systems per bit of cross-entropy reduction, though they achieve higher throughput.

3.3 Connection to Scaling Laws

Kaplan et al. [18] and Hoffmann et al. [15] established empirical scaling laws relating model cross-entropy to the number of parameters, training tokens, and compute budget. These scaling laws can be reinterpreted thermodynamically: since $\Delta H = H_{\text{prior}} - H_{\text{cross}}$, the well-documented power-law relationship between compute and cross-entropy implies a corresponding power-law relationship between compute and the effective compression achieved. Specifically, the diminishing returns in cross-entropy improvement with increasing compute imply that the marginal thermodynamic cost per additional bit of compression increases as models approach the irreducible entropy of natural language (estimated at $\sim$ 1-2 bits/token based on human prediction experiments [19]). A detailed quantitative fit to these scaling laws in thermodynamic terms would require substantially more data points than the three systems analyzed here and is left for future work.

3.4 Efficiency Trends

The demon efficiency of AI systems has improved over the period 2020-2023, with LLaMA-2 (2023) achieving roughly 17× higher efficiency than GPT-3 (2020). While the data are too sparse to establish a precise trend with confidence, this improvement is broadly consistent with a doubling time of approximately 2-3 years, reflecting combined advances in hardware efficiency and training algorithms. Even with continued exponential improvement at this rate, current efficiencies would remain far below the Landauer limit for decades. Practical limits on transistor energy efficiency suggest that conventional silicon technology can at best reach $\eta_D \sim 10^{-12}$–10^{-10} [11], leaving a gap of 10–12 orders of magnitude to be closed through qualitatively different approaches such as reversible or neuromorphic computing.

4 The Maxwell's Demon Analogy

The thermodynamic analysis above admits an instructive analogy to Maxwell's demon. We present this as a conceptual framework rather than a formal equivalence, as the mapping between the two domains is metaphorical at several points.

Maxwell's original thought experiment [1] proposed an intelligent being capable of sorting fast and slow gas molecules, apparently decreasing entropy without work. The resolution, completed by Bennett [4], showed that the demon must acquire, store, and eventually erase information about each molecule, with the erasure step dissipating at least $k_B T \ln 2$ of heat per bit by Landauer's principle [3], exactly compensating the entropy decrease of the gas.

LLM training admits the following analogical mapping: the high-entropy data distribution plays the role of the unsorted gas; the training process (gradient descent) plays the role of the demon's sorting; the model weights play the role of the demon's memory in which information about data regularities is stored; and the training energy dissipated as heat plays the role of the thermodynamic cost that the second law demands.

This analogy is useful because it provides physical intuition for a key feature of the training process: the local reduction of informational entropy (improved predictions) is necessarily accompanied by a global increase in thermodynamic entropy (heat dissipation), as the second law requires. However, we note important disanalogies. The demon performs explicit measurements on individual molecules, whereas gradient computation is a deterministic mathematical operation over batches of data; the "information extracted per gradient step" is a statistical aggregate, not a discrete measurement. The demon's memory is eventually erased to complete the thermodynamic cycle, whereas trained model weights persist and are used during inference without erasure. Despite these limitations, the analogy provides a useful conceptual bridge between information-theoretic and thermodynamic descriptions of the training process.

5 Discussion

The analysis presented here yields several observations about the physics of AI training, together with important caveats.

The Landauer bound applied to AI training (Proposition 3) establishes that the energy cost of training has a fundamental physical floor determined by the amount of cross-entropy reduction achieved. This bound is not new -it follows directly from Landauer's principle [3] - but applying it explicitly to AI systems makes the scale of current thermodynamic inefficiency concrete: approximately 10^{21} times above the fundamental limit. As discussed in Section 2.4, this gap can be approximately decomposed into contributions from transistor physics, memory hierarchy, algorithmic redundancy, and infrastructure overhead.

A key limitation of our analysis is the dependence on the choice of baseline entropy H_{prior}. As Table 2 demonstrates, different baselines yield demon efficiency values differing by factors of 2-3×. More fundamentally, the entire framework hinges on the assumption that the reduction in predictive cross-entropy (ΔH) can be identified with a reduction in physical entropy in the Landauer sense. This identification is not exact: ΔH measures the improvement in the model's predictive distribution, not a directly measurable change in the physical entropy of the hardware. The connection to Landauer's principle is therefore indirect - we treat the bits of predictive uncertainty eliminated during training as a proxy for the bits of information that must be physically stored in the low-entropy configuration of model weights. While this proxy is reasonable (the model weights do physically encode learned information, and writing them into an ordered configuration is an irreversible thermodynamic process), it should not be confused with a rigorous measurement of physical entropy change in the GPU substrate. Recent work on non-equilibrium bounds for learning in physical systems [22] has explored related questions from a more rigorous statistical-mechanical perspective; extending such approaches to the scale of LLM training remains an open challenge.

Our results connect to recent developments in information physics. Vopson and Lepadatu [20] introduced the second law of information dynamics, showing that information entropy in physical systems tends to minimize over time - a behavior observed across domains from genetic information to digital data storage. The systematic reduction of informational entropy during AI training, as quantified in this paper, is consistent with this principle and extends its applicability to artificial learning systems. The thermodynamic costs we calculate represent the physical price of such entropy reduction.

The comparison with biological systems is instructive, though approximate. The estimates of 10^{-9}–10^{-8} J/bit for brain learning [17] suggest that biological evolution has found information-processing architectures that operate 10^9-10^{10} times more efficiently than current silicon-based AI. This gap is expected, given that biological neural computation has been optimized by natural selection over billions of years, whereas digital computing optimizes for speed and flexibility at the expense of energy efficiency. Neuromorphic computing architectures [21] represent one approach to narrowing this gap.

Finally, we note that the MDL principle [8,9] provides a connection between compression and generalization: a model that achieves low cross-entropy on held-out data has, in an information-theoretic sense, captured genuine regularities in the data distribution rather than merely memorizing the training set. This connects the thermodynamic analysis to questions about what LLMs have "learned" though a detailed exploration of this connection is beyond the scope of the present paper.

6 Conclusion

We have applied Landauer's principle to the training of large language models, obtaining a lower bound on the minimum energy required as a function of the cross-entropy reduction achieved. Empirical analysis of GPT-3, PaLM, and LLaMA-2 reveals that current systems operate approximately 10^{21} times above this fundamental limit (Proposition 3), with the gap approximately attributable to transistor physics, memory hierarchy overhead, algorithmic redundancy, and infrastructure costs. The Maxwell's demon analogy (Section 4) provides useful physical intuition for the entropy-reducing character of AI training, though we have noted its limitations as a formal equivalence.

The main value of this analysis is in making explicit the physical constraints on AI training and providing a benchmark framework - the demon efficiency η_D -against which future improvements can be measured. The key open questions include: how much of the 10^{21} gap is reducible through improved algorithms versus requiring fundamentally different hardware; whether neuromorphic or reversible computing architectures can approach biological efficiency ($\sim 10^{-10}$ J/bit); and how the thermodynamic costs of training compare to those of inference at scale, particularly as deployment grows. Of the various paths to closing the remaining 10-12 orders of magnitude above the practical silicon limit, reversible and adiabatic computing [11] appears to be the only approach capable of reducing the dominant transistor-switching contribution by multiple orders of magnitude.

Acknowledgments

The author thanks the anonymous reviewer for detailed and constructive feedback that substantially improved the rigor and presentation of this work.

References

[1] James Clerk Maxwell, Theory of Heat, Longmans, Green, and Co., London (1871)

[2] Leo Szilard, Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen, Zeitschrift für Physik, Vol. 53, No. 11–12, pp. 840–856 (1929)

[3] Rolf Landauer, Irreversibility and heat generation in the computing process, IBM Journal of Research and Development, Vol. 5, No. 3, pp. 183–191 (1961)

[4] Charles H. Bennett, The thermodynamics of computation—a review, International Journal of Theoretical Physics, Vol. 21, No. 12, pp. 905–940 (1982)

[5] Antoine Bérut, Artak Arakelyan, Artyom Petrosyan, Sergio Ciliberto, Raoul Dillenschneider, and Eric Lutz, Experimental verification of Landauer's principle linking information and thermodynamics, Nature, Vol. 483, pp. 187–189 (2012)

[6] Melvin M. Vopson, The mass-energy-information equivalence principle, AIP Advances, Vol. 9, No. 9, 095206 (2019)

[7] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean, Carbon emissions and large neural network training, arXiv preprint arXiv:2104.10350 (2021)

[8] Jorma Rissanen, Modeling by shortest data description, Automatica, Vol. 14, No. 5, pp. 465–471 (1978)

[9] Peter D. Grünwald, The Minimum Description Length Principle, MIT Press (2007)

[10] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher, Pointer sentinel mixture models, arXiv preprint arXiv:1609.07843 (2016)

[11] Michael P. Frank, The physical limits of computing, Computing in Science and Engineering, Vol. 4, No. 3, pp. 16–26 (2002)

[12] Tom Brown, Benjamin Mann, Nick Ryder, et al., Language models are few-shot learners, Advances in Neural Information Processing Systems, Vol. 33, pp. 1877–1901 (2020)

[13] Hugo Touvron, Louis Martin, Kevin Stone, et al., LLaMA 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023)

[14] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, et al., PaLM: Scaling language modeling with pathways, arXiv preprint arXiv:2204.02311 (2022)

[15] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, et al., Training compute-optimal large language models, arXiv preprint arXiv:2203.15556 (2022)

[16] Peter Lennie, The cost of cortical computation, Current Biology, Vol. 13, No. 6, pp. 493–497 (2003)

[17] Simon Laughlin, Rob de Ruyter van Steveninck, and John Anderson, The metabolic cost of neural information, Nature Neuroscience, Vol. 1, No. 1, pp. 36–41 (1998)

[18] Jared Kaplan, Sam McCandlish, Tom Henighan, et al., Scaling laws for neural language models, arXiv preprint arXiv:2001.08361 (2020)

[19] Claude E. Shannon, Prediction and entropy of printed English, Bell System Technical Journal, Vol. 30, No. 1, pp. 50–64 (1951)

[20] Melvin M. Vopson and Serban Lepadatu, Second law of information dynamics, AIP Advances, Vol. 12, No. 7, 075310 (2022)

[21] Steve B. Furber, Francesco Galluppi, Steve Temple, and Luis A. Plana, The SpiNNaker project, Proceedings of the IEEE, Vol. 102, No. 5, pp. 652–665 (2014)

[22] Jeremy A. Owen, Artemy Kolchinsky, and David H. Wolpert, The fundamental thermodynamic costs of communication, arXiv preprint arXiv:2302.04320 (2023)