Physics in Next-token Prediction

Hongjun An [* 1] Yiliang Song [* 1] Xuelong Li [1]

Abstract

We discovered the underlying physics in Next-token Prediction (NTP). We identified the law of information conservation within NTP and proposed the **First Law of Information Capacity (IC-1)**, demonstrating that the essence of intelligence emergence in auto-regressive models is fundamentally a process of information transfer. We also introduced Landauer's Principle into NTP, formulating the **Second Law of Information Capacity (IC-2)**, which establishes the relationship between auto-regressive model training and energy consumption. Additionally, we presented several corollaries, which hold practical significance for production practices. Finally, we validated the compatibility and complementarity of our findings with existing theories.

1. Introduction

Currently, the state-of-the-art (SOTA) artificial intelligence models predominantly employ an auto-regressive architecture (OpenAI, 2024). Utilizing the Next-token Prediction (NTP) approach, these models seamlessly integrate various modalities, including text, images, audio, and video. This remarkable versatility and intelligence are reshaping human production and lifestyle in profound ways.

However, beneath this promising landscape looms a substantial "*scientific cloud.*" Guided by the Scaling Law (Kaplan et al., 2020), researchers are relentlessly constructing ever-larger training datasets and expending increasing amounts of energy to train larger auto-regressive models in pursuit of more "*intelligent*" systems. Why do big data and immense computational power lead to the emergence of "*intelligent*"? What is the essence behind this phenomenon? Where does the Scaling Law ultimately lead? What awaits us on the horizon?

In this paper, we delve deeply into these questions and un-

---
[*]Equal contribution [1]Institute of Aritificial Intelligence (TeleAI), China Telecom, P. R. China. Correspondence to: Xuelong Li <xuelong_li@ieee.org>.

The 1st Law of Information Capacity

$$\eta = \frac{D\left(H - L\right)}{N}$$

The 2nd Law of Information Capacity

$$\eta = \frac{E_0}{k_B N T \ln 2}$$

η: information capacity
N: parameter size (in bits)
L: average cross-entropy training loss
D: number of trained tokens
H: entropy of the entire dataset
E_0: minimum energy required to train
k_B: the Boltzmann constant
T: the Kelvin temperature of radiator

Figure 1. We propose the **First Law of Information Capacity (IC-1)** and the **Second Law of Information Capacity (IC-2)**, reveals the principle of information conservation and the energy relationship within NTP.

cover the underlying physics in NTP. We identify the law of information conservation (Baocheng et al., 2013) within NTP and derive the **First Law of Information Capacity (IC-1)**, demonstrating that the essence of intelligence emergence in auto-regressive models is fundamentally a process of information transfer. Additionally, we introduce Landauer's Principle (Landauer, 1961) into NTP, formulating the **Second Law of Information Capacity (IC-2)**, which reveals the relationship between the training process of auto-regressive models and the principles of energy in the real world. Based on our findings, we derive several corollaries that can effectively guide our daily practices and production activities. Finally, we validate the compatibility and com-

plementarity of our insights with existing theories, attesting to the rationality of our theoretical framework.

The contribution of this paper can be summarized as follows:

- We identified the conservation law of information in NTP and proposed the IC-1.

- We introduced the Landauer's Principle into NTP and proposed the IC-2.

- Based on IC-1 and IC-2, we derived several corollaries that can guide practical applications in human production.

- We validated the compatibility and complementarity of our perspectives with existing theories.

2. The First Law of Information Capability: the Conservation of Information in NTP

2.1. Preliminaries

In 2023, Rae conceptualized the process of NTP as a mechanism for compressing information of dataset (Rae, 2023), elucidating the relationship between the compression process and the emergence of intelligence: given a token vocabulary $\mathcal{V}$ and a dataset $\mathcal{D} = \{x_1, x_2, ..., x_{|\mathcal{D}|}\}(x_i \in \mathcal{V})$, our goal is to transmit $\mathcal{D}$ from $\mathcal{A}$ to $\mathcal{B}$ token-by-token as accurately as possible. During transmission, both parties can share an encoding and decoding function f. In this section, we will demonstrate that the intelligence level of function f is related to its ability to compress $\mathcal{D}$.

2.1.1. BASELINE: NON-INTELLIGENT TRANSMISSION

Assuming we have transmitted $\mathcal{D}_t = \{x_{1:t}\}(t < |\mathcal{D}|)$ and are about to transmit x_{t+1}, for a non-intelligent f_0, according to information theory (Shannon, 1948), the length of the code $z_{t+1}^{f_0} = f_0(x_{t+1}|x_{1:t})$ is at least $-\log P(x_{t+1}|x_{1:t})$ (Eq. 1), which is its self-information.

$$|z_{t+1}^{f_0}| = I(x_{t+1}|x_{1:t}) = -\log P(x_{t+1}|x_{1:t}), \quad (1)$$

where the initial condition is $P(x_1|x_0) = P(x_1)$.

Thus, the total cost required by this method is as indicated by $I(\mathcal{D})$ (Eq. 2).

$$I(\mathcal{D}) = \sum_{t=1}^{|\mathcal{D}|} |z_t^{f_0}| = \sum_{t=1}^{|\mathcal{D}|} -\log P(x_{t+1}|x_{1:t}) \\ = |\mathcal{D}|H(\mathcal{D}). \quad (2)$$

2.1.2. INTELLIGENT TRANSMISSION BASED ON COMPRESSION

When auto-regressive models, such as large language models (LLMs), are applied to predict the next token, at each iteration, they input $x_{1:t}$ and predict $P(x_{t+1}|x_{1:t}, f_a)$. So the total cost required by this method to transmit $\mathcal{D}$ is at least $I(\mathcal{D}|f_a)$ (Eq. 3).

$$I(\mathcal{D}|f_a) = \sum_{t=0}^{|\mathcal{D}|-1} -\log P(x_{t+1}|x_{1:t}, f_a). \quad (3)$$

It is noteworthy that the Eq. 3 is in exact correspondence with the objective loss function, cross-entropy loss (Bishop, 1995), which is optimised during the training phase (Eq. 4).

$$\ell(f_a) = \frac{1}{|\mathcal{D}|} \sum_{t=0}^{|\mathcal{D}|-1} -\log P(x_{t+1}|x_{1:t}, f_a) \\ = \frac{1}{|\mathcal{D}|} I(\mathcal{D}|f_a). \quad (4)$$

Thus, the model training process, characterized by the reduction of the loss function, can be regarded as a compression process of the dataset $\mathcal{D}$. The higher the compression, the more intelligent of the model becomes.

2.2. A Derivation of the First Law of Information Capability

If we conduct a meticulous examination of Eq. 2 and Eq. 4, we may uncover an even more intriguing insight. That is, when f_a is sufficiently powerful, $I(\mathcal{D}|f_a)$ is absolutely likely to be smaller than $I(\mathcal{D})$. Does this imply that some information disappears into thin air (Eq. 5)?

$$I(\mathcal{D}) - I(\mathcal{D}|f_a) = ? \quad (5)$$

In fact, from the perspective of physics, **information is conserved** (Baocheng et al., 2013), these pieces of information do not vanish; rather, they are transferred into the model f_a (Eq. 6).

$$I(f_a^+) = I(\mathcal{D}) - I(\mathcal{D}|f_a), \quad (6)$$

where $I(f_a^+)$ represents the effective information stored in f_a pertaining to task $\mathcal{D}$. We introduce the concept of **Information Capacity** (Li & Zhao, 2021), denoted by η, defined as the ratio of effective information $I(f_a^+)$ to its data size N (approximately equal to the number of parameter size, measured in bits). Thus, we obtain Eq. 7

$$\eta = \frac{I(f_a^+)}{N}. \quad (7)$$

After performing equivalent transformations to the Eq. 7, we arrive at the **First Law of Information Capacity (IC-1)**, as shown in Eq. 8.

$$\eta = \frac{D(H - L)}{N}. \tag{8}$$

where $L = \ell(f_a)$ denotes the average cross-entropy loss, $D = |\mathcal{D}|$ denotes the number of tokens in dataset $\mathcal{D}$, and $H = H(\mathcal{D})$ denotes the entropy of the entire dataset.

Therefore, the process of model training is essentially a process of compressing dataset $\mathcal{D}$, leading to a reduction in the loss function, the transfer of information to model f_a, and an increase in its information capacity η. This transfer is driven by the model training process, particularly the back propagation algorithm. (Rumelhart et al., 1986). The energy for this transfer comes from electrical resources in the real world (Sec. 3).

2.3. A Dynamic Perspective to the Process of Intelligence Emergence

Currently, Eq. 8 remains **static**. In particular, within our derivation, D represents the total number of tokens in the dataset, H denotes the overall entropy of the dataset, and L is the overall average loss after training has concluded. From the perspective of the law of conservation of information, it is imperative that conservation be observed not only at the terminal state, but throughout the entirety of the **dynamic** training process. Therefore, we will restate the meanings of the variables in Eq. 8:

- H: the overall entropy of the dataset, is a constant.

- N: parameter size of the model, measured in bits. Once the model architecture is established, it will become a fixed constant.

- D: the token numbers that **has been trained**. This is a **variable** that monotonically increases as training progresses.

- η, L: η is the information capacity of the model, and L is the dynamic average cross-entropy loss. Both change dynamically as D increases.

It is therefore possible to describe the entirety of the model training process from a dynamic perspective based on ITE:

Initial State. Training has not yet begun, with no information transfer, thus $\eta = 0$.

Training State. As training progresses, D gradually increases, leading to a decrease in L. To satisfy the equation, η must inevitably increase. During this dynamic process, the information gradually transfers to the f_a, prompting the model to *learn*.

Terminal State. When $\eta = \eta_{\max}$ (determined by the model architecture), the information that the model parameters can store reaches saturation. At this point, continuing the training process does not enable the model to learn more tokens; L converges, and training concludes.

3. The Second Law of Information Capacity: the Energy Relationship in NTP

In Sec. 2.2, the IC-1 indicates that the learning process in auto-regressive models is fundamentally an information transfer process. The driving force behind this transfer comes from the back-propagation (Rumelhart et al., 1986) algorithm, while the energy is sourced from the power of the physical world. One might wonder, what is the minimum amount of energy required to complete this information transfer process?

In 1961, Landauer proposed that the energy required to erase a single bit is at least $k_B T \ln 2$, known as the **Landauer's Principle** (Landauer, 1961). Therefore, according to Eq. 8, when we transfer information $I(f_a^+)$, at least energy $E_0 = I(f_a^+)k_B T \ln 2$ must be consumed. Thus, the training process of auto-regressive models establishes an energy relationship with the physical world. We can derive the **Second Law of Information Capacity (IC-2)** (Eq. 9).

$$\eta = \frac{E_0}{k_B N T \ln 2}, \tag{9}$$

where E_0 is the minimum energy required to complete the information transfer, $k_B \approx 1.38 \times 10^{-23}$J/K is the Boltzmann constant, and T is the temperature of the heat reservoir in Kelvin.

4. Corollaries to Guide Practice

4.1. Estimating the Entropy of the Dataset

From Sec. 2.3, it can be inferred that when the model training process is in its **initial state**, $\eta \approx 0$. From this, we can derive the following corollary: the Data Entropy Estimation Theorem (Corollary 4.1).

Corollary 4.1. *The entropy of the dataset is approximately equal to the initial loss of the model training.*

Proof.

$$\eta N = D(H - L), D > 0, \eta \approx 0, \\ \Rightarrow H \approx L. \tag{10}$$

It is crucial to acknowledge that this corollary is exclusively applicable for estimation purposes. This is due to the fact that the model itself incorporates initial information following its initialisation, which ultimately leads to the emergence of $\eta > 0$.

4.2. Estimating the Quality of the Dataset

When the number of tokens in the dataset is fixed, a higher entropy of the dataset indicates that it can provide more information for the model to learn. Therefore, we can consider entropy H as a metric for evaluating dataset quality, with the assessment method outlined in Corollary 4.1.

4.3. Matching Suitable Model Size with Dataset Size

In the practice of large model production, determining the appropriate model size, the amount of training data required, and the duration of training is a critical issue. In previous work, the Knowledge Capacity Scaling Laws indicated that current large language models (LLMs) generally store only 2 bits of information per parameter (Allen-Zhu & Li, 2024). For models using the FP16/BF16 or INT8 formats to store parameters, the information capacity η is approximately $0.125 \sim 0.25$. Additionally, according to the technical reports of well-known large models (Kaplan et al., 2020), it has been inferred that the entropy of datasets typically ranges from 10 to 15 approximately. Thus, if N and L are given, the required dataset size D can be estimated; if D and L are given, the required model size can be estimated.

Sec. 5.1 indicates that the IC-1 is compatible with the Scaling Law proposed by the OpenAI team in 2020 (Kaplan et al., 2020). Their advantage lies in providing power-law relationships between L and N, L and D, as well as L and C (computation cost) respectively through extensive experiments. However, these power-law relationships are validated only for large models based on the Transformer architecture (Vaswani et al., 2017). From the IC-1, we cannot derive the relationships between L and N or D respectively; we can only calculate their corresponding values from a macro perspective. Nevertheless, the IC-1 is applicable to all auto-regressive models, regardless of architecture, because its underlying principle is based on the physical principle of information conservation. Therefore, we are compatible and complementary with the OpenAI's Scaling Law.

4.4. Identify the Energy Limits Required for Training Auto-regressive Models

As chip technology advances and autore-gressive model learning algorithms are optimized, the energy required to train models achieving similar intelligence levels is expected to decline. Additionally, future developments may lead to a shift from GPUs to quantum computers, which could further reduce energy consumption. Nevertheless, regardless of technological advancement, there is an inherent limitation to the amount of energy that can be consumed for the transmission of information, known as Landauer's Limit, which is represented by IC-2 (Eq. 9).

5. Consistency with Existing Theories

5.1. Consistency with the Scaling Law of Neural Language Models

In this section, we will substitute the experimental data from (Kaplan et al., 2020) into the IC-1 to verify the consistency between the IC-1 and the Scaling Law of Neural Language Models. According to Sec. 2.3, we can know that the H is a constant.

When L is fixed, it indicates that model training has stopped. At this point, for a given set of N and D—that is, for a specific model trained on a defined number of tokens—η becomes a constant. Consequently, $(H - L)/\eta$ is also a constant, resulting in N being proportional to D.

$$N = kD, k = \frac{H - L}{\eta};$$
$$\Rightarrow N \propto D. \tag{11}$$

When D is fixed, it implies that the same number of tokens is being trained. If the model structure is altered to increase N, the change in the utilization rate η cannot be directly determined. Consequently, the relationship between L and N cannot be directly established (Eq. 12).

When N is fixed, it implies that the model is deterministic. If D is increased, meaning that the model is trained on a larger number of tokens, η will also increase correspondingly. However, the relationship between the change rate of the two cannot be directly determined, and thus the relationship between L and D also cannot be established directly (Eq. 12).

$$L = H - \eta \frac{N}{D}. \tag{12}$$

In summary, to verify whether the IC-1 aligns with the empirical formula measured by the OpenAI team in 2020, we can approach the analysis from the perspective of fixed L. Here is the empirical formula (Eq. 13 and Eq. 14):

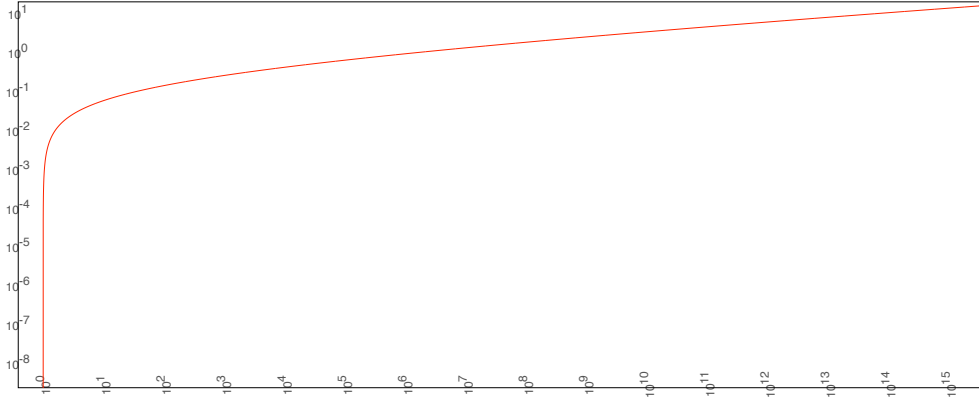$$L = \left(\frac{D}{5.4 \times 10^{13}}\right)^{-0.095}, \tag{13}$$

Figure 2. The plotted curve of $f(x) = x^{0.095} - x^{0.076}$. When $0 < x \ll 10^{15}$, $0 < f(x) \ll 10$.

$$L = \left(\frac{\frac{1}{16} N}{8.8 \times 10^{13}} \right)^{-0.076}. \tag{14}$$

The factor of $\frac{1}{16}$ in the Eq. 14 arises because the parameter count in the original paper refers to the quantity of FP/BF16 numbers, whereas N in this paper is expressed in bits. By setting (13)=(14), we obtain:

$$\left(\frac{5.4 \times 10^{13}}{D} \right)^{0.095} = \left(\frac{8.8 \times 10^{13}}{\frac{1}{16} N} \right)^{0.076}.$$

By constructing the function $f(x) = x^{0.095} - x^{0.076}$ and plotting its curve (Fig. 2), it becomes evident that when $0 < x \ll 10^{15}$, $0 < f(x) \ll 10$. Therefore, it can be inferred that the bases on both sides of the equation are nearly equal, with the difference in exponents attributed to observational errors.

$$\frac{5.4 \times 10^{13}}{D} \approx \frac{8.8 \times 10^{13}}{\frac{1}{16} N}.$$

Thus, we can deduce that:

$$N \approx 26.08 D.$$

This is consistent with the IC-1. Furthermore, when $\eta = 0$, meaning the model is untrained, we have $L_0 = H \approx 10$ (as visually inferred from Figure 2 in the paper (Kaplan et al., 2020)); when the model converges, $L \in [2, 7]$. Thus, we obtain:

$$\eta = \frac{H - L}{k} \in \left[\frac{10 - 7}{26.08}, \frac{10 - 3}{26.08} \right] = [0.115, 0.268].$$

The information capacity η falls within a normal range (< 1), thus indicating that the IC-1 is consistent with OpenAI's empirical formula.

5.2. Consistency with the Knowledge Capacity Scaling Laws

As noted in Sec. 4.3, the information capacity of LLMs in the Knowledge Capacity Scaling Laws should generally be less than $0.125 \sim 0.25$, which is consistent with the calculated value of $[0.115, 0.268]$ in Sec. 5.1.

6. Conclusion

This paper has revealed the fundamental physical principles that underpin NTP by establishing the IC-1 and the IC-2. These laws elucidate not only the conservation of information and the energy requirements in auto-regressive model training, but also offer practical corollaries that can guide the development and training of intelligent systems. By aligning our findings with existing theories, we have demonstrated the broad applicability and significance of our theoretical framework, thereby paving the way for more efficient and sustainable advancements in artificial intelligence.

References

Allen-Zhu, Z. and Li, Y. Physics of Language Models: Part 3.3, Knowledge Capacity Scaling Laws, 2024.

Baocheng, Z., Qingyu, C., Mingsheng, Z., and You, L. Information conservation is fundamental: recovering the lost information in Hawking radiation. *International Journal of Modern Physics D*, 22(12):1341014, 2013.

Bishop, C. M. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

Kaplan, J., McCandlish, S., Henighan, T., et al. Scaling Laws for Neural Language Models, 2020.

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

Li, X. and Zhao, B. Video distillation. *SCIENTIA SINICA Informationis*, 51(5):695–734, 2021.

OpenAI. Introducing OpenAI o1, 2024. URL `https://openai.com/o1`. Accessed: 2024-10-25.

Rae, J. Compression for AGI, 2023. URL `https://www.youtube.com/watch?v=dO4TPJkeaaU`. Accessed: 2024-10-23.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

Shannon, C. E. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

Vaswani, A., Shazeer, N., Parmar, N., et al. Attention is All you Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30. Curran Associates, Inc., 2017.