

L²M: Mutual Information Scaling Law for Long-Context Language Modeling

Zhuo Chen^{1,2,3} Oriol Mayné i Comas^{1,3,4} Zhuotao Jin^{1,3,5} Di Luo^{1,2,3,5,6} Marin Soljačić^{1,2,3}

Abstract

We rigorously establish a *bipartite* mutual information scaling law in natural language that governs long-range dependencies. This scaling law, which we show is distinct from and scales independently of the conventional two-point mutual information, is the key to understanding *long-context language modeling*. Using this scaling law, we formulate the **Long-context Language Modeling (L²M)** condition, which relates a model’s capacity for effective long context length modeling to the scaling of its latent state size for storing past information. Our results are validated through experiments on both transformers and state space models. This work establishes a theoretical foundation that guides the development of large language models toward longer context lengths.

1. Introduction

Large language models (LLMs) have revolutionized natural language processing, achieving remarkable capabilities across a wide range of tasks (Brown et al., 2020; Chowdhery et al., 2022; Touvron et al., 2023a;b). Recent advances, particularly with models like GPT-4 (OpenAI et al., 2024), Claude-3.5, LLaMA-3 (Grattafiori et al., 2024), and DeepSeek-V3 (DeepSeek-AI et al., 2024), have demonstrated breakthroughs in tasks from code generation to mathematical problem solving, from text summarization to creative writing (Stiennon et al., 2020; Yuan et al., 2022; Gou et al., 2023; Wang et al., 2023). These models have become increasingly powerful and versatile, pushing the boundaries of what’s possible in natural language processing and marking significant steps toward artificial general intelligence (Bubeck et al., 2023; Ge et al., 2023; Kosinski, 2023).

¹{chenzhuo, omayne, jinzhta, diluo, soljacic}@mit.edu ²NSF AI Institute for Artificial Intelligence and Fundamental Interactions ³Massachusetts Institute of Technology ⁴Polytechnic University of Catalonia ⁵Harvard University ⁶University of California, Los Angeles.

Correspondence to: Zhuo Chen <chenzhuo@mit.edu>.

Preprint. Copyright 2025 by the author(s).

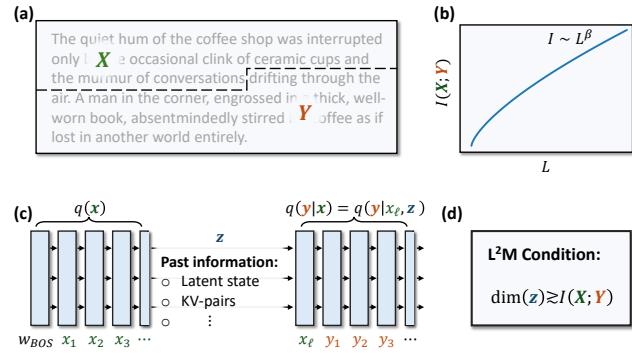


Figure 1. Illustration of the central idea of our paper. (a) A random block of text of length L is divided into two parts \mathbf{X} and \mathbf{Y} . (b) The (bipartite) mutual information between the random variables \mathbf{X} and \mathbf{Y} has a power-law scaling with respect to the length. (c) Autoregressive neural networks parameterize conditional distributions $q(\mathbf{y}|\mathbf{x}) = (\mathbf{y}|\mathbf{x}_\ell, \mathbf{z})$ via a hidden state \mathbf{z} that caches past information. (d) We formulate the Long-context Language Modeling (L²M) condition, which states that a model’s state size for storing past information must grow at least as the power-law scaling of the bipartite mutual information for effective long context length modeling.

In pushing these advances further, the ability to handle long contexts has become increasingly crucial. This ability is the key to document-level understanding, multi-turn dialogue, and complex reasoning. Models like GPT-o1 and DeepSeek-R1 often generate extensive chains of thought, sometimes spanning thousands of tokens to solve complex problems (Wei et al., 2022; Wang et al., 2024). However, processing long contexts remains a significant challenge. While transformer architectures have achieved remarkable success and recent advances like DeepSeek have dramatically improved the efficiency (DeepSeek-AI et al., 2024), they still suffer from the intrinsic computational cost quadratic in sequence length, resulting in challenges in long sequence length generation.

Although various architectures have been proposed to address the quadratic scaling (Katharopoulos et al., 2020; Gu et al., 2022a; Zhu et al., 2021; Beltagy et al., 2020; Ding et al., 2023; Gu & Dao, 2024; Dao & Gu, 2024; Peng et al., 2023; Sun et al., 2024), a fundamental gap persists in our theoretical understanding of how their effectiveness for cap-

turing long-range dependencies scales. In spite of existing efforts to characterize these dependencies through various statistical measures (Ebeling & Pöschel, 1994a; Debowski, 2011; Bentz et al., 2017; Futrell et al., 2019) theories guiding practical efficient architecture design have not yet been formulated.

In this work, we address the challenges of understanding long-context language modeling through the following contributions (Fig. 1).

1. We establish a bipartite mutual information scaling law, based on the relaxed Hilberg conjecture (Hilberg, 1990; Łukasz Debowski, 2015), which governs the true long-range dependencies of natural language and is the key to understanding long-context language modeling, fundamentally distinct from conventional two-point mutual information scaling.
2. We validate this new scaling law across various natural language datasets using state-of-the-art LLMs, including both LLaMA (Grattafiori et al., 2024) and DeepSeek (DeepSeek-AI et al., 2024) models, revealing strongly consistent power-law growth behavior.
3. We formulate the **Long-context Language Modeling (L²M)** condition and prove that a model’s state size for storing past information must scale faster than the bipartite mutual information scaling for effective long context length modeling, which is further empirically verified with transformer and state space models trained on varying sequence lengths.

Our theoretical foundation for understanding LLMs’ capability for long sequence length modeling provides crucial insights and quantitative analysis for improving LLM architectures, enabling more effective long-context applications, and guiding the development of more capable and scalable AI systems.

2. Related Works

2.1. Mutual Information Estimation and Application in Machine Learning

Mutual information estimation and optimization have been extensively studied in machine learning, with approaches including variational bounds (Poole et al., 2019), neural estimators (Belghazi et al., 2021), nearest-neighbor methods (Kraskov et al., 2004; Gao et al., 2015), and various upper bounds (Cheng et al., 2020). It has wide applications in feature selection (Brown et al., 2012), representation learning (Tschanneen et al., 2020), disentanglement (Chen et al., 2016), and generative modeling (Hjelm et al., 2019).

2.2. Statistical Properties of Natural Language

Natural language exhibits characteristic statistical scaling behaviors across different levels of analysis. Zipf’s law (Gelbukh & Sidorov, 2001) describes how word frequencies decay with their rank following a power-law distribution. Heaps’ law (Gelbukh & Sidorov, 2001) characterizes vocabulary growth, showing that the number of unique words scales sublinearly with text length. Hilberg’s conjecture and its relaxed version posit specific scaling laws for entropy and bipartite mutual information in natural language, respectively (Hilberg, 1990).

2.3. Neural Scaling Laws

Power-law relationships between model performance, architecture, and computational requirements have been established (Kaplan et al., 2020; Biderman et al., 2023), though theoretical understanding remains limited (Bahri et al., 2024; Bordelon et al., 2024). These empirical observations have guided the development of larger models, complementing our theoretical framework for context length scaling.

2.4. Architectures for Efficient Long-Context Modeling

Various approaches have been proposed for processing long sequences. Architectural innovations targeting the quadratic complexity include sparse attention (Child et al., 2019; Ding et al., 2023; Beltagy et al., 2020; Zaheer et al., 2020), recurrent mechanisms (Dai et al., 2019; Rae et al., 2019; Sukhbaatar et al., 2019), and alternative formulations (Katharopoulos et al., 2020; He et al., 2024; Gu et al., 2022a; Gu & Dao, 2024; Dao & Gu, 2024; Peng et al., 2023; Sun et al., 2024; Gu et al., 2022b; Beck et al., 2024; De et al., 2024). Memory-efficient implementations like Flash Attention (Dao et al., 2022; Dao, 2023; Shah et al., 2024), Lightning Attention (Qin et al., 2024), and Paged Attention (Kwon et al., 2023) have improved computational efficiency while maintaining the underlying complexity scaling.

2.5. Long-Form Reasoning and Context Utilization

Chain-of-thought prompting (Wei et al., 2022) and scratchpad methods (Nye et al., 2021) demonstrate the importance of extended context for complex reasoning tasks, emphasizing the need for effective long-range dependency modeling.

2.6. Information Theory and Physics-Inspired Approaches

Recent work has demonstrated how information-theoretic principles and physics-inspired approaches can guide machine learning (Tishby & Zaslavsky, 2015; Goldfeld & Polyanskiy, 2020; Chen & Luo, 2024), leading to novel architectures (Luo & Clark, 2019; Luo et al., 2023; Wang et al., 2021; Chen et al., 2022; Lami et al., 2022; Wu et al.,

2023; Chen et al., 2023; Dugan et al., 2024), training methods (Stokes et al., 2020; Chen et al., 2024b;a), and broad applications (Carleo et al., 2019; Lee et al., 2020; Luo et al., 2022; Moro et al., 2024; Chen et al., 2024b; Choi et al., 2024). Our work continues this direction by establishing concrete connections between information-theoretic principles and practical model architectures.

3. Preliminaries

3.1. Mutual Information

Mutual information $I(X; Y)$ quantifies the statistical dependence between random variables X and Y , defined as:

$$I(X; Y) = D_{KL}(p_{XY} \| p_X \otimes p_Y), \quad (1)$$

where $D_{KL}(\cdot \| \cdot)$ denotes the Kullback–Leibler (KL) divergence and p_{XY} is the joint distribution of X and Y . For discrete random variables, mutual information can be equivalently expressed as:

$$\begin{aligned} I(X; Y) &= H(X) + H(Y) - H(XY) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X), \end{aligned} \quad (2)$$

where $H(\cdot)$ denotes the (Shannon) entropy and $H(\cdot | \cdot)$ denotes the conditional entropy. For continuous distributions, these relations hold when entropy is replaced by differential entropy $h(\cdot)$. While differential entropy lacks some properties of Shannon entropy, the mutual information remains well-defined and maintains its key properties. This definition naturally extends to collections of random variables, which we denote as $I(X_{1:m}; Y_{1:n})$, where the notation $X_{i:j}$ represents the sequence (X_i, \dots, X_j) . For notational convenience, we use boldface notation $\mathbf{X} := X_{i:j}$ when the indices are clear from context. Similarly, we often drop the index of a single variable $X := X_i$ when appropriate.

3.2. Autoregressive Neural Networks

Modern large language models (LLMs) predominantly employ autoregressive neural architectures. An autoregressive neural network models a sequence of conditional probability distributions over tokens $\{q(w_i | w_{1:i-1}, w_{\text{BOS}})\}_{i=1}^L$, where w_{BOS} denotes the beginning-of-sentence token¹. Throughout this paper, we use q to denote model-generated probability distributions and p to denote the true underlying distributions; upper case letters denote random variables and lower case letters denote specific values or realizations of random variables. These conditional distributions jointly model the

¹Some early LLMs, such as the original GPT, did not include this token.

probability for a sequence of tokens given a prefix:

$$q(w_{\ell:L} | w_{1:\ell-1}, w_{\text{BOS}}) = \prod_{i=\ell}^L q(w_i | w_{1:i-1}, w_{\text{BOS}}). \quad (3)$$

When $\ell = 1$, this reduces to the unconditional generation distribution $q(w_{1:L} | w_{\text{BOS}})$. During inference, tokens are sampled sequentially from these conditional distributions to generate text or respond to prompts. Our theoretical analysis is independent of the specific sampling procedure employed.

4. Mutual Information Scaling Laws

In this section, we establish and empirically verify two distinct types of scaling laws for mutual information in natural language: *bipartite* mutual information scaling between adjacent text segments [Fig. 2 (a)] and *two-point* mutual information scaling between pairs of tokens [Fig. 2 (d)]. We then demonstrate that *bipartite* mutual information provides crucial insights into long-range dependencies in natural language in Section 5, and connect an LLM’s long context length modeling capacity to its ability to capture correct bipartite mutual information scaling, leading to the L²M condition.

4.1. Bipartite Mutual Information and Relaxed Hilberg Conjecture

While classical scaling laws in natural language (e.g., Zipf’s law and Heaps’ law) focus on token-level statistics, understanding language modeling requires analyzing dependencies between text segments. The *bipartite* mutual information between adjacent text blocks emerges as a particularly illuminating measure.

Definition 4.1 (*Bipartite* Mutual Information [Fig. 2(a)]). For a consecutive sequence of tokens (random variables) $W_{1:L}$ of length L , consider a bipartition of the tokens: $X_{1:\ell} := W_{1:\ell}$ and $Y_{1:L-\ell} := W_{\ell+1:L}$. The bipartite mutual information is the mutual information between the two parts $I_{\ell:L}^{\text{BP}} := I(X_{1:\ell}; Y_{1:L-\ell})$

When $L = 2\ell$, the bipartite mutual information in natural language is conjectured to follow a power-law growth (Hilberg, 1990; Łukasz Debowski, 2015),

Conjecture 4.2 (Relaxed Hilberg Conjecture²). *For equal lengths³ of \mathbf{X} and \mathbf{Y} , the bipartite mutual information scales as $I_{L/2;L}^{\text{BP}} \sim L^\beta$ for some $\beta \in [0, 1]$.*

We note this power-law growth is sometimes also referred

²The original conjecture does not concern tokens specifically, but the scaling should be similar when measured on symbols, tokens or words.

³Having equal lengths may not be necessary, but it presents in the original statement of the conjecture.

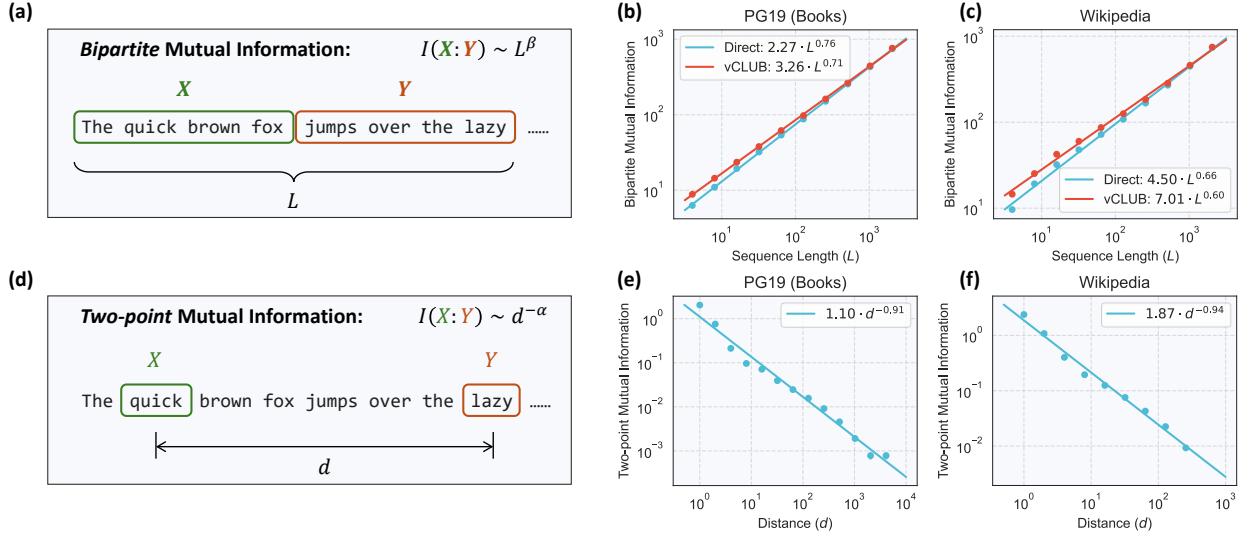


Figure 2. Illustration and estimations of bipartite and two-point mutual information in natural language. (a) The *bipartite* mutual information measures statistical dependence between two adjacent segments within a text block of length L . (b, c) Estimations using LLaMA 3.1 405B model (Meta) on PG19 dataset (Rae et al., 2020) of pre-1919 books and WIKIPEDIA (Foundation). Both direct estimation and vCLUB approximation (Cheng et al., 2020) show consistent results across datasets. See Appx. A.IV for results using DeepSeek and other LLaMA models. (d) The *two-point* mutual information measures statistical dependence between tokens separated by distance d . (e, f) Two-point mutual information estimations on PG19 dataset and WIKIPEDIA.

to as the sub-volume law. We will use the two terms interchangeably depending on context.

While empirical verification of this power-law growth is intuitively appealing, rigorous validation has proven challenging due to difficulties in estimating high-dimensional entropy and mutual information from samples. Early work by Hilberg (Hilberg, 1990) considered a stricter version of the conjecture (also see Appx. A.I) and found $\beta \approx 0.5$, with limited data and only for sequences up to 100 characters. Later studies using universal compression codes estimated $\beta \approx 0.95$ for millions of tokens (Łukasz Debowski, 2015), though these results may be sensitive to compression algorithm choice and likely overestimate the exponent (see Appx. A.II). Recent work using kNN and MINE estimators found a related scaling $I_{\ell;L}^{\text{BP}} \sim \ell^{0.82}$ for $\ell/L \leq 0.2$ and $L \leq 200$ (Lu et al., 2024), but suffered from dimensional bias issues⁴.

Despite these estimation challenges, all studies support the existence of power-law growth. In Sec. 4.3, we verify this conjecture using LLM approximations to the underlying distribution of natural language and demonstrate clear power-law scaling.

⁴The dimensional bias is severe enough that their measured $I_{50;100}^{\text{BP}}$ exceeds $I_{50;200}^{\text{BP}}$, violating basic mutual information inequalities.

4.2. Two-point Mutual Information

Before presenting our empirical results on bipartite mutual information scaling, we first address two-point mutual information, which was previously thought to indicate long-range language dependencies. In Sec. 4.4, we show why two-point mutual information is inadequate for capturing true multi-token dependencies, while bipartite mutual information provides a more complete understanding.

Definition 4.3 (*Two-point* Mutual Information [Fig. 2(d)]). The two-point mutual information measures the mutual information between two tokens (random variables) X and Y separated by a distance d : $I_d^{\text{TP}} = I(X;Y)$.

We note that the two-point mutual information has been observed to follow a power-law decay $I_d^{\text{TP}} \sim d^{-\alpha}$ (Ebeling & Pöschel, 1994b; Ebeling & Neiman, 1995; MONTEMURRO & PURY, 2002; Shen, 2019). Based on this scaling, it is often argued that natural language exhibits similar structures and long-range dependencies to critical physical systems (Lin & Tegmark, 2017), whose two-point correlation function follows similar decay (Stanley, 1995). We stress that this argument is misleading. The two-point mutual information does not capture the multivariate long-range dependencies in natural language. While the two-point mutual information for both natural language and critical physical systems follows a power-law decay, the bipartite mutual information scalings are drastically different—one-dimensional critical physical system only exhibits logarithmically growing bipartite mutual information (Alcaraz &

Rajabpour, 2013; Dai et al., 2020), much slower than the sub-volume law growth expected in natural language. We explain this in more detail in Sec. 4.4.

4.3. Empirical Verification of Mutual Information Scaling Laws

4.3.1. BIPARTITE MUTUAL INFORMATION

Measuring bipartite mutual information presents significant challenges without access to the underlying probability distribution p . Recent advances in LLMs provide a solution by offering high-quality approximations q to these distributions. As autoregressive neural networks, LLMs enable efficient computation of conditional probabilities (Sec 3.2) and their associated cross entropies (negative log likelihoods):

$$H(p_{\mathbf{Y}|\mathbf{X}}, q_{\mathbf{Y}|\mathbf{X}}) := -\mathbb{E}_{p_{\mathbf{X}\mathbf{Y}}} \log q(\mathbf{Y}|\mathbf{X}), \quad (4)$$

where the expectation is taken over samples from the true underlying distribution $p_{\mathbf{X}\mathbf{Y}}$. The cross entropy provides an upper bound to the true entropy:

$$H(p_{\mathbf{Y}|\mathbf{X}}, q_{\mathbf{Y}|\mathbf{X}}) = D_{KL}(p_{\mathbf{Y}|\mathbf{X}} || q_{\mathbf{Y}|\mathbf{X}}) + H^p(\mathbf{Y}|\mathbf{X}), \quad (5)$$

where the conditional cross entropy and KL-divergence always implicitly average over $p_{\mathbf{X}}$, and H^p (or H^q) denotes the entropy computed with respect to distribution p (or q).

Using these properties, we can construct a direct estimator for bipartite mutual information:

$$\begin{aligned} I_{\ell;L}^{\text{BP,direct}} &= \mathbb{E}_{p_{\mathbf{X}\mathbf{Y}}} [\log q(\mathbf{Y}|\mathbf{X}) - \log q(\mathbf{Y})] \\ &= H(p_{\mathbf{Y}}, q_{\mathbf{Y}}) - H(p_{\mathbf{Y}|\mathbf{X}}, q_{\mathbf{Y}|\mathbf{X}}) \\ &= I^p(\mathbf{X}; \mathbf{Y}) + \varepsilon(p, q), \end{aligned} \quad (6)$$

where $I^p(\mathbf{X}; \mathbf{Y})$ denotes mutual information with respect to p and $\varepsilon(p, q) = D_{KL}(p_{\mathbf{Y}} || q_{\mathbf{Y}}) - D_{KL}(p_{\mathbf{Y}|\mathbf{X}} || q_{\mathbf{Y}|\mathbf{X}})$. While this estimator no longer provides a bound, it preserves the key property that $\varepsilon(p, q) \rightarrow 0$ as $q \rightarrow p$.

We note this estimation method faces a specific challenge with modern LLMs: they model $q(w_i | w_{1:i-1}, w_{BOS})$ rather than $q(w_i | w_{1:i-1})$, where w_{BOS} denotes the BOS token. When sampling from the dataset, we can ensure \mathbf{X} starts at sentence beginnings, making $q(\mathbf{Y}|\mathbf{X}, w_{BOS}) \equiv q(\mathbf{Y}|\mathbf{X})$. However, \mathbf{Y} may start mid-sentence, creating a mismatch where $q(\mathbf{Y}) \neq q(\mathbf{Y}|w_{BOS})$. This introduces errors in estimating $H(p_{\mathbf{Y}}, q_{\mathbf{Y}})$. We address this using n -gram corrections for the first two tokens, from where most of biases come (see Appx. A.III).

To circumvent issues with estimating $q(\mathbf{Y})$, we also employ the vCLUB estimator (Cheng et al., 2020):

$$I_{\ell;L}^{\text{BP,vCLUB}} = \mathbb{E}_{p_{\mathbf{X}\mathbf{Y}}} \log q(\mathbf{Y}|\mathbf{X}) - \mathbb{E}_{p_{\mathbf{X}} \otimes p_{\mathbf{Y}}} \log q(\mathbf{Y}|\mathbf{X}), \quad (7)$$

where the second term can be calculated by shuffling the second halves of samples in the dataset. Analysis in (Cheng et al., 2020) shows that vCLUB provides an upper bound on the true bipartite mutual information when q closely approximates p . Even when q deviates moderately from p , though the upper bound property may not hold, vCLUB continues to provide reliable estimates of the true bipartite mutual information.

Our empirical analysis in Fig.2(b, c) examines equal-length partitions of \mathbf{X} and \mathbf{Y} ($\ell = L/2$), which maximizes bipartite mutual information for fixed L . Using both the bias-corrected direct estimator [Eq.(6)] and vCLUB estimator [Eq. (7)], we measure scaling on the PG19 dataset⁵ (Rae et al., 2020) (a collection of books before 1919) and WIKIPEDIA (Foundation), employing the LLaMA 3.1 405B model (Meta) as q . All measurements reveal clear power-law scaling across thousands of tokens. Additional measurements using LLaMA 3.1 70B (Meta) and DeepSeek V3 Base models (DeepSeek-AI et al., 2024), along with varying ℓ/L ratios, can be found in Appx. A.V). We note that both estimators likely underestimate the true exponent β (see Appx. A.VI for discussion).

4.3.2. TWO-POINT MUTUAL INFORMATION

Two-point mutual information estimation is more straightforward, requiring only 1- and 2-gram statistics without LLM approximations. We estimate this quantity using entropy calculations for individual tokens and token pairs separated by distance d . Following (Grassberger, 2008), we employ their bias-reduced entropy estimator:

$$\hat{H}(X) = \hat{H}(Y) = \log N - \frac{1}{N} \sum_{m=1}^M n_m G(n_m), \quad (8)$$

where N is the total number of tokens, M is the vocabulary size, n_m is the number of tokens whose token ID equals n_m , and $G(\cdot)$ is defined as

$$G(n) = \psi(n) + \frac{(-1)^n}{2} \left(\psi\left(\frac{n+1}{2}\right) - \psi\left(\frac{n}{2}\right) \right) \quad (9)$$

with $\psi(\cdot)$ the digamma function.

The entropy of pairs of tokens is estimated analogously, with the summation running over all ordered pairs of tokens (m_i, m_j) , resulting in the total number of terms quadratic in the vocabulary size. The mutual information is then estimated as

$$\hat{I}_d^{\text{TP}}(X; Y) = \hat{H}(X) + \hat{H}(Y) - \hat{H}(XY). \quad (10)$$

⁵We couldn't use the BOOKS3 dataset because it was removed due to copyright infringement concerns.

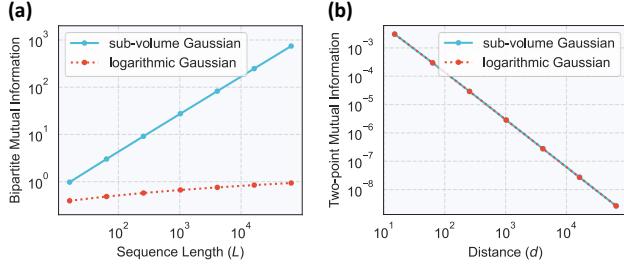


Figure 3. Bipartite and two-point mutual information scalings for two families of multivariate Gaussian distributions with varying sequence lengths (see Appx. B). The two families of distributions (a) have significantly different scaling for bipartite mutual information (L^β vs $\log L$), but (b) identical two-point mutual information scaling ($d^{-\alpha}$).

We note that this mutual information estimator exhibits systematic bias. The entropy estimator has a negative bias that increases (in absolute value) with dimension of the sample space $|\Omega|$. Since $|\Omega_X| = |\Omega_Y| = M$ where as $|\Omega_{XY}| = M^2$, the bias in $\hat{H}(XY)$ exceeds that in $\hat{H}(X) = \hat{H}(Y)$, leading to a positive bias in \hat{I}_d . This bias becomes particularly problematic at large distances d , where $H(XY) \approx H(X) + H(Y)$ and I_d approaches zero. To mitigate this issue, we perform additional bias correction by fitting the systematic bias from the data (see Appx. A.VII for detail).

We apply this methodology to measure two-point mutual information on both the PG19 dataset and WIKIPEDIA, confirming power-law decay in both cases [Fig. 2 (e, f)].

4.4. Distinction between Bipartite and Two-point Mutual Information

Bipartite and two-point mutual information differ fundamentally in their characterization of dependencies in multivariate systems. We demonstrate through two examples that two-point mutual information can fail to capture the true nature of long-range dependencies, leading to potentially misleading conclusions.

Consider a simple distribution where all tokens must be identical: $p(x_1, x_2, \dots, x_L) = \mathbb{1}(x_1 = x_2 = \dots = x_L)/M$, with $\mathbb{1}(\cdot)$ being the indicator function that evaluates to 1 when the condition is satisfied and 0 otherwise, and M the vocabulary size. Notice that this distribution permits a Markov chain construction as $p(x_1) = 1/M$ and $p(x_i|x_{1:i-1}) = \mathbb{1}(x_i = x_{i-1})$, and therefore has a simple token-to-token dependency structure. Despite this simplicity, the two-point mutual information of the distribution suggests a misleading strong “long-range” dependency—it maintains a large value of $I_d^{\text{TP}} = \log M$ without any decay as d increases, far exceeding that of natural languages. In contrast, the bipartite mutual information correctly identifies

the simple dependency structure as $I_{\ell,L}^{\text{BP}} = \log M$ remains constant for any choice of ℓ and L —two segments share exactly the same amount of information ($\log M$) regardless of their lengths, which is no more than the information shared between just two adjacent tokens.

For a more realistic setting, in Fig. 3, we consider two families of multivariate Gaussian distributions of different lengths (see Appx. B for details in constructing these distributions). In particular, both families show identical power-law decay in their two-point mutual information between variables at maximum separation. However, they exhibit dramatically different bipartite mutual information scaling—one scales as L^β , similar to natural language, while the other scales as $\log L$, similar to critical physical systems.

These examples illustrate that two-point mutual information alone is insufficient, and can be misleading: it may suggest strong long-range dependencies where none exist, or fail to distinguish between systems with fundamentally different dependency structures.

5. Long-Context Language Modeling (L²M) Condition

Having established bipartite mutual information as a more reliable measure of dependencies, we analyze how a model’s capacity to handle long contexts fundamentally depends on its ability to store past information, using bipartite mutual information scaling as our theoretical framework. Intuitively, to model natural language effectively, a model must be able to capture all dependencies between past and future tokens. Since these dependencies (measured by bipartite mutual information) grow with sequence length, the model’s state capacity for storing past information must necessarily grow as well. We formalize this intuition through the Long-context Language Modeling (L²M) condition and explore its implications in detail throughout this section.

5.1. Theoretical Derivations

To analyze how models handle long-range dependencies, we first formalize the notion of “model state for storing past information”, which we call the *history state*.

Definition 5.1. Consider a sequence of tokens $w_{1:L}$. Denote $x_{1:\ell} := w_{1:\ell}$ and $y_{1:L-\ell} := w_{\ell+1:L}$. For autoregressive neural networks that parameterize conditional probabilities $q(y_{1:L-\ell}|x_{1:\ell})$, we define the *history state* at ℓ as the intermediate variables $z_\ell = f(x_{1:\ell-1})$ of smallest size that, together with x_ℓ , fully characterizes the model’s behavior through $q(y|x_\ell, z_\ell)$ ⁶ [Fig. 1(c)].

⁶We separate x_ℓ from z_ℓ to accurately reflect its distinct role as current input token in autoregressive models, though including it in z_ℓ would not affect the main results of this paper.

As illuminating examples, the history state corresponds to the latent state in RNNs and state space models (SSMs) after processing token $w_{\ell-1}$, and to the key-value pairs up to token $w_{\ell-1}$ for transformers (see Appx. C). Generally, the history state \mathbf{z}_ℓ should be chosen as the smallest hidden state responsible for caching historical information.

We now establish that the size of this history state upper bounds a model’s capacity to capture bipartite mutual information:

Theorem 5.2. *The maximum bipartite mutual information that a model can capture is bounded by the size of its history state:*

$$I_{L/2;L}^{\text{BP},q} \leq C \cdot \dim(\mathbf{z}_{L/2}) + \log(M) \quad (11)$$

where C is some constant and M denotes the vocabulary size.

Here, we focus on $\ell = L/2$ where the bipartite mutual information reaches the maximum for a given L .

We present an intuitive proof below assuming discrete history states $\mathbf{z}_{L/2}$. This assumption naturally aligns with float point or quantized numerical representations in neural networks, and does not affect our theoretical conclusions. A complete proof with relaxed assumptions is provided in Appx. D.

Proof. By the data processing inequality: $I^q(X_{1:L/2}; Y_{1:L/2}) \leq I^q(\mathbf{Z}_{L/2}, X_{L/2}; Y_{1:L/2})$. This is upper bounded by the entropy: $H^q(\mathbf{Z}_{L/2}, X_{L/2})$, which is further upper bounded by $H^q(\mathbf{Z}_{L/2}) + H^q(X_{L/2}) \leq C \cdot \dim(\mathbf{z}_{L/2}) + \log(M)$, where the last inequality follows from the bound on entropies of discrete variables. \square

Having established that a model’s history state dimension bounds its ability to capture bipartite mutual information, we now formalize how models must scale to maintain performance as sequence length increases. To systematically analyze this scaling behavior, we consider both increasing sequence lengths and model sizes, examining whether models can continue to capture the mutual information present in the data as both scales grow. We formalize this notion through the following definition of MI-capable model scaling.

Definition 5.3. Let $\{W_{1:L}\}_{L=1}^\infty$ be a series of natural language datasets of different lengths, which we obtain by truncating an ideal infinite-length dataset. Define a scaling of a model architecture (or simply a scaling of models) as a series of models $\{q_L\}_{L=1}^\infty$ of the same architecture, with possibly increasing model sizes as L grows. We say a scaling of models is *MI-capable* if and only if each model q_L can achieve the desired bipartite mutual information for its corresponding sequence length L : $I_{L/2;L}^{\text{BP},q_L} = I_{L/2;L}^{\text{BP}}$.

Since each model’s mutual information is bounded by its history state dimension, we can now characterize the necessary scaling of history states for MI-capability.

Theorem 5.4 (L²M Condition). *For a scaling of models $\{q_L\}_{L=1}^\infty$ to be MI-capable, their history states $\mathbf{z}_{L/2}^{q_L}$ just before token $w_{L/2}$ must satisfy the scaling*

$$\dim(\mathbf{z}_{L/2}^{q_L}) \gtrsim I_{L/2;L}^{\text{BP}} \sim L^\beta.$$

In other words, to capture the full bipartite mutual information at each sequence length, the history state dimension must grow at least as fast as the power-law scaling of mutual information in the data.

Proof. We prove by contrapositive. By Thm. 5.2, each model’s mutual information is bounded by its history state dimension: $I_{L/2;L}^{\text{BP},q_L} \lesssim \dim(\mathbf{z}_{L/2}^{q_L})$. If $\dim(\mathbf{z}_{L/2}^{q_L}) \prec I_{L/2;L}^{\text{BP}}$, then $I_{L/2;L}^{\text{BP},q_L} \prec I_{L/2;L}^{\text{BP}}$ ⁷, implying there exists some L where $I_{L/2;L}^{\text{BP},q_L} < I_{L/2;L}^{\text{BP}}$, violating MI-capability. \square

This is the central result of the paper and establishes the minimal scaling required for a model’s history state dimension to capture the power-law growth of bipartite mutual information in natural language.

5.2. Implications to Common LLM Architectures

Our theoretical framework allows us to analyze how different architectures’ history states scale with sequence length, and what this implies for their ability to capture long-range dependencies.

In transformer-based models (excluding sparse attention and linear attention variants), the history state consists of stored key-value pairs for all previous tokens. With fixed model size, the size of key-value pairs already grows linearly with sequence length: $\dim(\mathbf{z}_{L/2}^{q_L}) \sim L \gtrsim L^\beta$. This means transformer models naturally satisfy the L²M condition without model size scaling, as the history state dimension automatically grows with sequence length, despite the quadratic computational cost.

On the other hand, although often celebrated for their “infinite” context length and linear complexity, SSMs, RNNs, and linear attention models do not satisfy the L²M condition without scaling up model sizes as sequence length increases. When the model size is fixed, the history state dimension remains constant regardless of the sequence length. Our theory shows this constant-size state cannot capture the growing mutual information. To actually satisfy the L²M condition, these architectures require increasingly larger models as sequence length grows, so that their history state

⁷We use “ \lesssim ” to denote scaling at the same or lower rate, and “ \prec ” to denote scaling at a strictly lower rate.

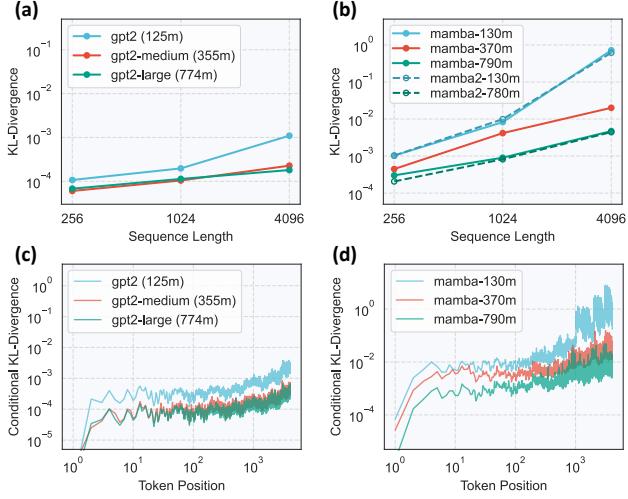


Figure 4. Evaluation of KL-divergence across model architectures trained on sub-volume Gaussian distributions. (a, b) Average KL-divergence per token for models trained on different sequence lengths. (c, d) Position-wise conditional KL-divergence for models trained on sequence length 4096. Lower values indicate better performance.

dimensions can increase accordingly. This requirement effectively offsets their computational efficiency advantage for long sequence length modeling.

For other architectures like sparse attention models, we can similarly analyze their history state scaling to understand when they can or cannot satisfy the L^2M condition.

We note that the L^2M condition only addresses a model’s ability to capture long-range dependencies, not its overall language modeling capability. Therefore, the L^2M condition remains a necessary but not sufficient condition—failing to satisfy it implies inherent limitations at longer sequences, while satisfying it does not guarantee effective language modeling. It is also distinct from neural scaling laws, which typically study how model performance scales with model size, dataset size, and compute budget at a given sequence length.

6. Empirical Verification

Having established the L^2M condition and its implications for various LLM architectures, we now empirically validate our theoretical predictions.

6.1. Experimental Setup

We first validate our theory using the sub-volume Gaussian distributions introduced in Fig. 3. This distribution family closely resembles both the bipartite and two-point mutual information scaling observed in natural language,

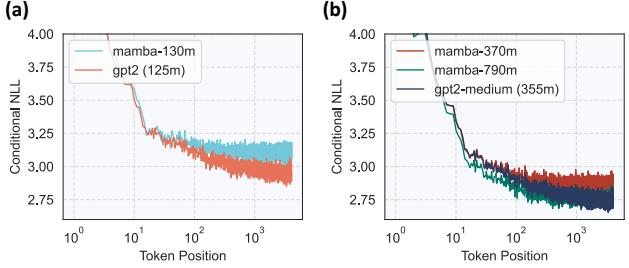


Figure 5. Position-wise conditional negative log likelihood (NLL) evaluation for models trained on 4096-token sequences on the PG19 dataset (Rae et al., 2020). Lower values indicate better performance, with performance differences at later positions highlighting varying capabilities in modeling long-range dependencies.

while allowing for efficient calculation of conditional probabilities and KL-divergences (see Appx. B), which would be intractable with real natural language datasets. To thoroughly stress the models, we stack 64 sub-volume Gaussian distributions together and group the random variables at the same position into a single token (see Appx. E for details).

We then extend our analysis to the PG19 dataset (Rae et al., 2020), a high-quality collection of pre-1919 books with long contextual dependencies.

We evaluate GPT2 (Brown et al., 2020), Mamba (Gu & Dao, 2024), and Mamba2 (Dao & Gu, 2024) models of varying sizes as representative transformer and state space model architectures. Additional experimental details can be found in Appx. E.

6.2. Results

In Fig. 4, we present the average per-token KL-divergence (a, b) and position-wise conditional KL-divergence (c, d) for both architectures (see Appx. E for definition). We specifically use KL-divergence rather than negative log likelihood (NLL) for more interpretable results. NLL conflates model performance (KL-divergence) with the inherent entropy of the data, which can obscure differences in model capabilities, particularly at longer sequence lengths where both the sub-volume Gaussian distribution and natural language tend to have lower conditional and average entropies.

Our results reveal that GPT2 maintains consistent KL-divergence across varying sequence lengths and positions. In contrast, smaller Mamba and Mamba2 models demonstrate increasing difficulty with longer contexts, requiring substantially larger model sizes to achieve comparable performance at sequence length 4096. These findings directly align with our L^2M condition and remain consistent across both Mamba variants, providing strong empirical validation of our theoretical framework.

In Fig. 5, we show the position-wise conditional NLL of models trained on the PG19 dataset (Rae et al., 2020) with 4096-token sequences, where calculating KL-divergence is not feasible. In this experiment, we observe two important patterns. In (a), GPT2-125M and Mamba-130M achieve comparable performance early in sequences, but GPT2’s advantage becomes evident at later positions. In (b), we see similar behavior with GPT2-Medium (355M) and Mamba-370M. Furthermore, GPT2-Medium (355M) initially underperforms Mamba-790M at early token positions, yet their performance converges at later positions despite the Mamba model being more than twice the size. These observations confirm transformers’ stronger capacity for modeling long-range dependencies, consistent with predictions by our L^2M condition.

We emphasize that Mamba’s linear computational complexity can make larger Mamba models practically more efficient than smaller transformers. Our results should not be interpreted as suggesting Mamba’s architectural inferiority. They simply demonstrate that a model’s capacity for capturing long-range dependencies aligns with our theoretical L^2M framework, regardless of the architecture.

Additional experimental results can be found in Appx. F.

7. Conclusion

In this work, we establish a bipartite mutual information scaling law, based on the relaxed Hilberg conjecture, to characterize long-range dependencies in natural language, fundamentally distinct from conventional two-point mutual information scaling. Through evaluations on diverse datasets with state-of-the-art LLMs, we validate this law and observe a consistent power-law growth pattern. We introduce the L^2M condition, proving that a model’s state size must scale faster than bipartite mutual information for effective long-context modeling, which we verify empirically using transformers and state space models. These findings suggest promising directions, such as designing synthetic language datasets with proper mutual information scaling and developing memory- and computation-efficient LLM architectures. Two intriguing questions emerge: is it possible to design architectures that just meet this theoretical minimum requirement, and can there exist architectures with linear computational complexity that satisfy the L^2M condition? By bridging theoretical scaling laws with empirical verification, our work provides a solid foundation for advancing long-context language modeling.

Limitations

Our theoretical framework specifically examines models’ capacity to capture long-range dependencies, without addressing other aspects of language modeling, such as reason-

ing or world knowledge. While our empirical validation on the Gaussian distribution dataset captures the essential mutual information scaling effects, it may not fully reflect the complexities of natural language. Our analysis is currently limited to English text, primarily because the LLMs used for approximation were predominantly trained on English data, and verification of the scaling law in other languages would provide valuable insights into potential universal properties across different linguistic structures. Additionally, our theory focuses on autoregressive language models, and extending it to other architectures such as discrete diffusion models or even vision models could reveal similar information scaling behaviors in different domains. Our evaluations rely on open-source models like LLaMA and DeepSeek, and further verification using closed-source state-of-the-art models would provide additional validation of our findings.

Impact Statement

This work advances our theoretical understanding of how language models process long-range dependencies, with implications for the design and deployment of more efficient LLM architectures. By establishing the L^2M condition, we provide a principled framework for evaluating architectures’ fundamental capacity for long-context modeling. This could lead to more efficient models that maintain effectiveness while reducing computational resources, potentially decreasing the environmental footprint of training and inference. Our findings may influence the development of specialized architectures for tasks requiring long-context understanding, such as legal document analysis, scientific research, and complex reasoning. However, improved long-context modeling could also amplify existing challenges in LLMs, including the propagation of biases over longer contexts and enhanced capabilities for generating persuasive misinformation. Research applying the L^2M framework should consider these ethical dimensions, particularly how improvements in long-range dependency modeling might affect model safety, fairness, and the verifiability of model outputs.

Acknowledgements

The authors acknowledge support from the National Science Foundation under Cooperative Agreement PHY-2019786 ([The NSF AI Institute for Artificial Intelligence and Fundamental Interactions](#)). Z.C. acknowledges support from the Mathworks Fellowship. Z.C. thanks Rumen Dangovski for helpful discussions. The research was sponsored by the United States Air Force Research Laboratory and the Department of the Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The computations in this paper were partly run on the FASRC cluster supported by the FAS

Division of Science Research Computing Group at Harvard University.

References

- Alcaraz, F. C. and Rajabpour, M. Universal behavior of the shannon mutual information of critical quantum chains., 2013.
- Bahri, Y., Dyer, E., Kaplan, J., Lee, J., and Sharma, U. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024. doi: 10.1073/pnas.2311878121. URL <https://www.pnas.org/doi/abs/10.1073/pnas.2311878121>.
- Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., Klambauer, G., Brandstetter, J., and Hochreiter, S. xlstm: Extended long short-term memory, 2024. URL <https://arxiv.org/abs/2405.04517>.
- Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, R. D. Mine: Mutual information neural estimation, 2021. URL <https://arxiv.org/abs/1801.04062>.
- Beltagy, I., Peters, M. E., and Cohan, A. Longformer: The long-document transformer, 2020. URL <https://arxiv.org/abs/2004.05150>.
- Bentz, C., Alikaniotis, D., Cysouw, M., and i Cancho, R. F. The entropy of words - learnability and expressivity across more than 1000 languages, 2017.
- Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., and van der Wal, O. Pythia: A suite for analyzing large language models across training and scaling, 2023. URL <https://arxiv.org/abs/2304.01373>.
- Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., Pieler, M., Prashanth, U. S., Purohit, S., Reynolds, L., Tow, J., Wang, B., and Weinbach, S. Gpt-neox-20b: An open-source autoregressive language model, 2022. URL <https://arxiv.org/abs/2204.06745>.
- Bordelon, B., Atanasov, A., and Pehlevan, C. A dynamical model of neural scaling laws, 2024. URL <https://arxiv.org/abs/2402.01092>.
- Brown, G., Pocock, A., Zhao, M.-J., and Luján, M. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13(2):27–66, 2012. URL <http://jmlr.org/papers/v13/brown12a.html>.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y., Li, Y.-F., Lundberg, S. M., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.
- Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., Vogt-Maranto, L., and Zdeborová, L. Machine learning and the physical sciences. *Rev. Mod. Phys.*, 91:045002, Dec 2019. doi: 10.1103/RevModPhys.91.045002. URL <https://link.aps.org/doi/10.1103/RevModPhys.91.045002>.
- Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/7c9d0b1f96aebd7b5eca8c3eda19ebb-Paper.pdf.
- Chen, Z. and Luo, D. Entangling intelligence: Ai-quantum crossovers and perspectives. In *2024 IEEE 6th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA)*, pp. 516–519, 2024. doi: 10.1109/TPS-ISA62245.2024.00069.
- Chen, Z., Luo, D., Hu, K., and Clark, B. K. Simulating 2+1d lattice quantum electrodynamics at finite density with neural flow wavefunctions, 2022. URL <https://arxiv.org/abs/2212.06835>.
- Chen, Z., Newhouse, L., Chen, E., Luo, D., and Soljacic, M. Antn: Bridging autoregressive neural networks

- and tensor networks for quantum many-body simulation. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 450–476. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/01772a8b0420baec00c4d59fe2fbace6-Paper.pdf.
- Chen, Z., Dangovski, R., Loh, C., Dugan, O., Luo, D., and Soljačić, M. Quanta: Efficient high-rank fine-tuning of llms with quantum-informed tensor adaptation, 2024a. URL <https://arxiv.org/abs/2406.00132>.
- Chen, Z., McCarran, J., Vizcaino, E., Soljacic, M., and Luo, D. TENG: Time-evolving natural gradient for solving PDEs with deep neural nets toward machine precision. In *Forty-first International Conference on Machine Learning*, 2024b. URL <https://openreview.net/forum?id=v1I4zRAjMb>.
- Cheng, P., Hao, W., Dai, S., Liu, J., Gan, Z., and Carin, L. Club: A contrastive log-ratio upper bound of mutual information, 2020. URL <https://arxiv.org/abs/2006.12013>.
- Child, R., Gray, S., Radford, A., and Sutskever, I. Generating long sequences with sparse transformers, 2019. URL <https://arxiv.org/abs/1904.10509>.
- Choi, S., Salamin, Y., Roques-Carmes, C., Dangovski, R., Luo, D., Chen, Z., Horodynski, M., Sloan, J., Uddin, S. Z., and Soljačić, M. Photonic probabilistic machine learning using quantum vacuum noise. *Nature Communications*, 15(1):7760, Sep 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-51509-0. URL <https://doi.org/10.1038/s41467-024-51509-0>.
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. Palm: Scaling language modeling with pathways, 2022. URL <https://arxiv.org/abs/2204.02311>.
- Cohn, H. and Zhao, Y. Sphere packing bounds via spherical codes. *Duke Mathematical Journal*, 163(10), July 2014. ISSN 0012-7094. doi: 10.1215/00127094-2738857. URL <http://dx.doi.org/10.1215/00127094-2738857>.
- Dai, Y.-W., Chen, X.-H., Cho, S. Y., and Zhou, H.-Q. Critical exponents of block-block mutual information in one-dimensional infinite lattice systems., 2020.
- Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., and Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context, 2019. URL <https://arxiv.org/abs/1901.02860>.
- Dao, T. Flashattention-2: Faster attention with better parallelism and work partitioning, 2023. URL <https://arxiv.org/abs/2307.08691>.
- Dao, T. and Gu, A. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality, 2024. URL <https://arxiv.org/abs/2405.21060>.
- Dao, T., Fu, D. Y., Ermon, S., Rudra, A., and Ré, C. Flashattention: Fast and memory-efficient exact attention with io-awareness, 2022. URL <https://arxiv.org/abs/2205.14135>.
- De, S., Smith, S. L., Fernando, A., Botev, A., Cristian-Muraru, G., Gu, A., Haroun, R., Berrada, L., Chen, Y., Srinivasan, S., Desjardins, G., Doucet, A., Budden, D., Teh, Y. W., Pascanu, R., Freitas, N. D., and Gulcehre, C. Griffin: Mixing gated linear recurrences with local attention for efficient language models, 2024. URL <https://arxiv.org/abs/2402.19427>.
- Debowski, L. Excess entropy in natural language: present state and perspectives, 2011.
- DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Zhang, H., Ding, H., Xin, H., Gao, H., Li, H., Qu, H., Cai, J. L., Liang, J., Guo, J., Ni, J., Li, J., Wang, J., Chen, J., Chen, J., Yuan, J., Qiu, J., Li, J., Song, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Xu, L., Xia, L., Zhao, L., Wang, L., Zhang, L., Li, M., Wang, M., Zhang, M., Zhang, M., Tang, M., Li, M., Tian, N., Huang, P., Wang, P., Zhang, P., Wang, Q., Zhu, Q., Chen, Q., Du, Q., Chen, R. J., Jin, R. L., Ge, R., Zhang, R., Pan, R., Wang, R., Xu, R., Zhang, R., Chen, R., Li, S. S., Lu, S., Zhou, S., Chen, S., Wu, S., Ye, S., Ye, S., Ma, S., Wang, S., Zhou, S., Yu, S., Zhou, S., Pan, S., Wang, T., Yun, T., Pei, T., Sun, T., Xiao, W. L., Zeng, W., Zhao, W., An, W., Liu,

- W., Liang, W., Gao, W., Yu, W., Zhang, W., Li, X. Q., Jin, X., Wang, X., Bi, X., Liu, X., Wang, X., Shen, X., Chen, X., Zhang, X., Chen, X., Nie, X., Sun, X., Wang, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yu, X., Song, X., Shan, X., Zhou, X., Yang, X., Li, X., Su, X., Lin, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhu, Y. X., Zhang, Y., Xu, Y., Xu, Y., Huang, Y., Li, Y., Zhao, Y., Sun, Y., Li, Y., Wang, Y., Yu, Y., Zheng, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Tang, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Wu, Y., Ou, Y., Zhu, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Zha, Y., Xiong, Y., Ma, Y., Yan, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Wu, Z. F., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Huang, Z., Zhang, Z., Xie, Z., Zhang, Z., Hao, Z., Gou, Z., Ma, Z., Yan, Z., Shao, Z., Xu, Z., Wu, Z., Zhang, Z., Li, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Gao, Z., and Pan, Z. Deepseek-v3 technical report, 2024. URL <https://arxiv.org/abs/2412.19437>.
- Ding, J., Ma, S., Dong, L., Zhang, X., Huang, S., Wang, W., Zheng, N., and Wei, F. Longnet: Scaling transformers to 1,000,000,000 tokens, 2023. URL <https://arxiv.org/abs/2307.02486>.
- Dugan, O., Beneto, D. M. J., Loh, C., Chen, Z., Dangovski, R., and Soljačić, M. Occamllm: Fast and exact language model arithmetic in a single step, 2024. URL <https://arxiv.org/abs/2406.06576>.
- Ebeling, W. and Neiman, A. Long-range correlations between letters and sentences in texts. *Physica A: Statistical Mechanics and its Applications*, 215(3):233–241, 1995. ISSN 0378-4371. doi: [https://doi.org/10.1016/0378-4371\(95\)00025-3](https://doi.org/10.1016/0378-4371(95)00025-3). URL <https://www.sciencedirect.com/science/article/pii/0378437195000253>.
- Ebeling, W. and Pöschel, T. Entropy and long-range correlations in literary english, 1994a.
- Ebeling, W. and Pöschel, T. Entropy and long-range correlations in literary english. *Europhysics Letters*, 26(4):241, may 1994b. doi: 10.1209/0295-5075/26/4/001. URL <https://dx.doi.org/10.1209/0295-5075/26/4/001>.
- Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy models of superposition, 2022. URL <https://arxiv.org/abs/2209.10652>.
- Foundation, W. Wikimedia downloads. URL <https://dumps.wikimedia.org>.
- Futrell, R., Qian, P., Gibson, E., Fedorenko, E., and Blank, I. Syntactic dependencies correspond to word pairs with high mutual information. In Gerdes, K. and Kahane, S. (eds.), *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pp. 3–13, Paris, France, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-7703. URL <https://aclanthology.org/W19-7703/>.
- Gao, S., Steeg, G. V., and Galstyan, A. Efficient estimation of mutual information for strongly dependent variables, 2015. URL <https://arxiv.org/abs/1411.2003>.
- Ge, Y., Hua, W., Ji, J., Tan, J., Xu, S., and Zhang, Y. Opennagi: When llm meets domain experts, 2023.
- Gelbukh, A. and Sidorov, G. Zipf and heaps laws' coefficients depend on language. In *Conference on Intelligent Text Processing and Computational Linguistics*, 2001. URL <https://api.semanticscholar.org/CorpusID:20344161>.
- Goldfeld, Z. and Polyanskiy, Y. The information bottleneck problem and its applications in machine learning. *IEEE Journal on Selected Areas in Information Theory*, 1(1):19–38, 2020. doi: 10.1109/JSAIT.2020.2991561.
- Gou, Z., Shao, Z., Gong, Y., Shen, Y., Yang, Y., Huang, M., Duan, N., and Chen, W. Tora: A tool-integrated reasoning agent for mathematical problem solving, 2023.
- Grassberger, P. Entropy estimates from insufficient samplings, 2008. URL <https://arxiv.org/abs/physics/0307138>.
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esibov, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton,

- J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces, 2024. URL <https://openreview.net/forum?id=AL1fq05o7H>.
- Gu, A., Goel, K., and Re, C. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*,

- 2022a. URL <https://openreview.net/forum?id=uYLFozlvlAC>.
- Gu, A., Goel, K., and Ré, C. Efficiently modeling long sequences with structured state spaces, 2022b. URL <https://arxiv.org/abs/2111.00396>.
- He, Z., Qin, Z., Prakriya, N., Sun, Y., and Cong, J. Hmt: Hierarchical memory transformer for long context language processing, 2024. URL <https://arxiv.org/abs/2405.06067>.
- Hilberg, W. Der bekannte grenzwert der redundanzfreien information in texten - eine fehlinterpretation der shannonschen experimete? *Frequenz*, 44:243 – 248, 1990. URL <https://api.semanticscholar.org/CorpusID:124878549>.
- Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization, 2019. URL <https://arxiv.org/abs/1808.06670>.
- Inc., W. R. Mathematica, Version 14.2. URL <https://www.wolfram.com/mathematica>. Champaign, IL, 2024.
- Jiang, Y., Rajendran, G., Ravikumar, P., Aragam, B., and Veitch, V. On the origins of linear representations in large language models, 2024.
- Kabatiansky, G. A. and Levenshtein, V. I. On bounds for packings on a sphere and in space. *Problemy Perekhodchi Informatsii*, 14(1):3–25, 1978. URL <http://mi.mathnet.ru/ppi1518>. English translation: *Problems Inform. Transmission*, vol. 14, no. 1, pp. 1–17, 1978.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models, 2020. URL <https://arxiv.org/abs/2001.08361>.
- Katharopoulos, A., Vyas, A., Pappas, N., and Fleuret, F. Transformers are rnns: Fast autoregressive transformers with linear attention, 2020. URL <https://arxiv.org/abs/2006.16236>.
- Kawabata, T. and Dembo, A. The rate-distortion dimension of sets and measures. *IEEE Transactions on Information Theory*, 40(5):1564–1572, 1994. doi: 10.1109/18.333868.
- Kosinski, M. Theory of mind may have spontaneously emerged in large language models, 2023.
- Kraskov, A., Stögbauer, H., and Grassberger, P. Estimating mutual information. *Phys. Rev. E*, 69:066138, Jun 2004. doi: 10.1103/PhysRevE.69.066138. URL <https://link.aps.org/doi/10.1103/PhysRevE.69.066138>.
- Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention, 2023. URL <https://arxiv.org/abs/2309.06180>.
- Lami, G., Carleo, G., and Collura, M. Matrix product states with backflow correlations. *Phys. Rev. B*, 106:L081111, Aug 2022. doi: 10.1103/PhysRevB.106.L081111. URL <https://link.aps.org/doi/10.1103/PhysRevB.106.L081111>.
- Lee, C.-H., Khan, A., Luo, D., Santos, T. P., Shi, C., Janicek, B. E., Kang, S., Zhu, W., Sobh, N. A., Schleife, A., Clark, B. K., and Huang, P. Y. Deep learning enabled strain mapping of single-atom defects in two-dimensional transition metal dichalcogenides with sub-picometer precision. *Nano Letters*, 20(5):3369–3377, May 2020. ISSN 1530-6984. doi: 10.1021/acs.nanolett.0c00269. URL <https://doi.org/10.1021/acs.nanolett.0c00269>.
- Li, J., Chen, D., Cai, T., Chen, P., Hong, Y., Chen, Z., Shen, Y., and Gan, C. Flexattention for efficient high-resolution vision-language models, 2024. URL <https://arxiv.org/abs/2407.20228>.
- Lin, H. W. and Tegmark, M. Critical behavior in physics and probabilistic formal languages. *Entropy*, 19(7), 2017. ISSN 1099-4300. doi: 10.3390/e19070299. URL <https://www.mdpi.com/1099-4300/19/7/299>.
- Lu, S., Kanász-Nagy, M., Kukuljan, I., and Cirac, J. I. Tensor networks and efficient descriptions of classical data, 2024. URL <https://arxiv.org/abs/2103.06872>.
- Luo, D. and Clark, B. K. Backflow transformations via neural networks for quantum many-body wave functions. *Phys. Rev. Lett.*, 122:226401, Jun 2019. doi: 10.1103/PhysRevLett.122.226401. URL <https://link.aps.org/doi/10.1103/PhysRevLett.122.226401>.
- Luo, D., Chen, Z., Carrasquilla, J., and Clark, B. K. Autoregressive neural network for simulating open quantum systems via a probabilistic formulation. *Phys. Rev. Lett.*, 128:090501, Feb 2022. doi: 10.1103/PhysRevLett.128.090501. URL <https://link.aps.org/doi/10.1103/PhysRevLett.128.090501>.

- Luo, D., Chen, Z., Hu, K., Zhao, Z., Hur, V. M., and Clark, B. K. Gauge-invariant and anyonic-symmetric autoregressive neural network for quantum lattice models. *Phys. Rev. Res.*, 5:013216, Mar 2023. doi: 10.1103/PhysRevResearch.5.013216. URL <https://link.aps.org/doi/10.1103/PhysRevResearch.5.013216>.
- Meta. URL <https://ai.meta.com/blog/meta-llama-3-1/>.
- MONTEMURRO, M. A. and PURY, P. A. Long-range fractal correlations in literary corpora. *Fractals*, 10(04):451–461, 2002. doi: 10.1142/S0218348X02001257. URL <https://doi.org/10.1142/S0218348X02001257>.
- Moro, V., Loh, C., Dangovski, R., Ghorashi, A., Ma, A., Chen, Z., Kim, S., Lu, P. Y., Christensen, T., and Soljačić, M. Multimodal learning for materials, 2024. URL <https://arxiv.org/abs/2312.00111>.
- Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., Sutton, C., and Odena, A. Show your work: Scratchpads for intermediate computation with language models, 2021. URL <https://arxiv.org/abs/2112.00114>.
- OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgium, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O’Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotstetd, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024. URL <https://arxiv.org/abs/2303.08774>.
- Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models, 2023.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library, 2019. URL <https://arxiv.org/abs/1912.01703>.
- Peng, B., Alcaide, E., Anthony, Q. G., Albalak, A., Arcadinho, S., Biderman, S., Cao, H., Cheng, X., Chung, M. N., Derczynski, L., Du, X., Grella, M., GV, K. K., He, X., Hou, H., Kazienko, P., Kocon, J., Kong, J., Koptyra, B., Lau, H., Lin, J., Mantri, K. S. I., Mom, F., Saito, A., Song, G., Tang, X., Wind, J. S., Woźniak, S., Zhang, Z., Zhou, Q., Zhu, J., and Zhu, R.-J. RWKV: Reinventing RNNs for the transformer era. In *The 2023 Conference*

- on Empirical Methods in Natural Language Processing*, 2023. URL <https://openreview.net/forum?id=7SaXczaBpG>.
- Poole, B., Ozair, S., van den Oord, A., Alemi, A. A., and Tucker, G. On variational bounds of mutual information, 2019. URL <https://arxiv.org/abs/1905.06922>.
- Qin, Z., Sun, W., Li, D., Shen, X., Sun, W., and Zhong, Y. Lightning attention-2: A free lunch for handling unlimited sequence lengths in large language models, 2024. URL <https://arxiv.org/abs/2401.04658>.
- Rae, J. W., Potapenko, A., Jayakumar, S. M., and Lillicrap, T. P. Compressive transformers for long-range sequence modelling, 2019. URL <https://arxiv.org/abs/1911.05507>.
- Rae, J. W., Potapenko, A., Jayakumar, S. M., Hillier, C., and Lillicrap, T. P. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=SylKikSYDH>.
- Shah, J., Bikshandi, G., Zhang, Y., Thakkar, V., Ramani, P., and Dao, T. Flashattention-3: Fast and accurate attention with asynchrony and low-precision, 2024. URL <https://arxiv.org/abs/2407.08608>.
- Shen, H. Mutual information scaling and expressive power of sequence models, 2019. URL <https://arxiv.org/abs/1905.04271>.
- Stanley, H. E. Power laws and universality. *Nature*, 378: 554–555, 1995. doi: 10.1038/378554a0.
- Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R. J., Voss, C., Radford, A., Amodei, D., and Christiano, P. Learning to summarize from human feedback, 2020.
- Stokes, J., Izaac, J., Killoran, N., and Carleo, G. Quantum Natural Gradient. *Quantum*, 4:269, May 2020. ISSN 2521-327X. doi: 10.22331/q-2020-05-25-269. URL <https://doi.org/10.22331/q-2020-05-25-269>.
- Sukhbaatar, S., Grave, E., Bojanowski, P., and Joulin, A. Adaptive attention span in transformers, 2019. URL <https://arxiv.org/abs/1905.07799>.
- Sun, Y., Li, X., Dalal, K., Xu, J., Vikram, A., Zhang, G., Dubois, Y., Chen, X., Wang, X., Koyejo, S., Hashimoto, T., and Guestrin, C. Learning to (learn at test time): Rnns with expressive hidden states, 2024. URL <https://arxiv.org/abs/2407.04620>.
- Tishby, N. and Zaslavsky, N. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pp. 1–5, 2015. doi: 10.1109/ITW.2015.7133169.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lamplie, G. Llama: Open and efficient foundation language models, 2023a. URL <https://arxiv.org/abs/2302.13971>.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models, 2023b. URL <https://arxiv.org/abs/2307.09288>.
- Tschannen, M., Djolonga, J., Rubenstein, P. K., Gelly, S., and Lucic, M. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=rkxoh24FPH>.
- Wang, J., Chen, Z., Luo, D., Zhao, Z., Hur, V. M., and Clark, B. K. Spacetime neural network for high dimensional quantum dynamics, 2021. URL <https://arxiv.org/abs/2108.02200>.
- Wang, J., Meng, F., Liang, Y., and Zhou, J. Drt-o1: Optimized deep reasoning translation via long chain-of-thought, 2024. URL <https://arxiv.org/abs/2412.17498>.
- Wang, Y., Le, H., Gotmare, A. D., Bui, N. D. Q., Li, J., and Hoi, S. C. H. Codet5+: Open code large language models for code understanding and generation, 2023.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., brian ichter, Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain of thought prompting elicits reasoning in large language models. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Huggingface’s transformers: State-of-the-art natural language processing, 2020. URL <https://arxiv.org/abs/1910.03771>.

Wu, D., Rossi, R., Vicentini, F., and Carleo, G. From tensor-network quantum states to tensorial recurrent neural networks. *Physical Review Research*, 5(3), July 2023. ISSN 2643-1564. doi: 10.1103/physrevresearch.5.1032001. URL <http://dx.doi.org/10.1103/PhysRevResearch.5.L032001>.

Yuan, A., Coenen, A., Reif, E., and Ippolito, D. Wordcraft: Story writing with large language models, 2022.

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., and Ahmed, A. Big bird: Transformers for longer sequences. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17283–17297. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf.

Zhu, C., Ping, W., Xiao, C., Shoeybi, M., Goldstein, T., Anandkumar, A., and Catanzaro, B. Long-short transformer: Efficient transformers for language and vision, 2021. URL <https://arxiv.org/abs/2107.02192>.

Łukasz Debowski. The relaxed hilberg conjecture: A review and new experimental support. *Journal of Quantitative Linguistics*, 22(4):311–337, 2015. doi: 10.1080/09296174.2015.1106268. URL <https://doi.org/10.1080/09296174.2015.1106268>.

A. Additional Details on Mutual Information Scalings

A.I. Hilberg Conjecture

In the main text, we mainly discussed the bipartite mutual information with equal splitting of tokens

$$I_{L/2;L}^{\text{BP}} := I(X_{1:L/2}, Y_{1:L/2}) \quad (\text{A.1})$$

with $X_{1:L/2} := W_{1:L/2}$ and $Y_{1:L/2} := W_{L/2+1:L}$ where $W_{1:L}$ is the sequence of tokens. The relaxed Hilberg conjecture states that this bipartite mutual information follows a power-law growth

$$I_{L/2;L}^{\text{BP}} \sim L^\beta. \quad (\text{A.2})$$

In Hilberg's original work (Hilberg, 1990), however, he considered a more strict conjecture that says the entropy of natural language should also follow a power law

$$H(W_{1:L}) \sim L^\beta. \quad (\text{A.3})$$

It is easy to see that the Hilberg conjecture leads to the relaxed Hilberg conjecture.

$$I_{L/2;L}^{\text{BP}} = H(W_{1:L/2}) + H(W_{L/2+1:L}) - H(W_{1:L}) = A \left(\frac{L}{2} \right)^\beta + A \left(\frac{L}{2} \right)^\beta - AL^\beta = A(2^{1-\beta} - 1)L^\beta \sim L^\beta. \quad (\text{A.4})$$

However, the converse is not true—allowing the entropy of natural language to have an additional linear term does not alter the bipartite mutual information scaling.

A.II. Universal Compression Code Approximation of Mutual Information

One important result in information theory is the relation between the entropy of a random variable, to its average compression rate. In particular, The average number of bits needed to store the random variable, often referred to as the Kolmogorov complexity, is upper bounded by its entropy

$$K(X) = \mathbb{E}[|\mathcal{C}(X)|] \leq H_2(X), \quad (\text{A.5})$$

where $\mathcal{C}(\cdot)$ denotes a compression algorithm, $|\cdot|$ is the length of the bitstring, and $H_2(\cdot) = \log 2H(\cdot)$ is the entropy measured in bits, to distinguish from entropy measured in nats used in the main text. Because this difference in unit only results in a difference in multiplicative constant and does not alter the scaling, we will drop this log 2 constant and make no distinction between the units in later discussions.

The Kolmogorov complexity allows the bipartite mutual information to be estimated as

$$I_{L/2;L}^{\text{BP}} \approx I_{L/2;L}^{\text{BP,Kol}} := K(W_{1:L/2}) + K(W_{L/2+1:L}) - K(W_{1:L}). \quad (\text{A.6})$$

Since no compression code perfectly compresses natural language, it is reasonable to believe that the Kolmogorov complexity grows faster than the entropy of natural language. Assuming the original Hilberg conjecture [Eq. (A.3)] is true, it is reasonable to believe that the Kolmogorov complexity scales as

$$K(W_{1:L}) \sim L^{\beta+\varepsilon}, \quad (\text{A.7})$$

with ε due to the inefficiency of compression code. Therefore, the Kolmogorov estimation of the mutual information would also grow as

$$I_{L/2;L}^{\text{BP,Kol}} \sim L^{\beta+\varepsilon}, \quad (\text{A.8})$$

resulting in an overestimation of the scaling exponent.

A.III. Direct Estimation of Bipartite Mutual Information Using LLMs

In the main text, our direct estimator for the bipartite mutual information is

$$I_{\ell;L}^{\text{BP,direct}} = \mathbb{E}_{p_{\mathbf{X}\mathbf{Y}}} [\log q(\mathbf{Y}|\mathbf{X}) - \log q(\mathbf{Y})] = H(p_{\mathbf{Y}}, q_{\mathbf{Y}}) - H(p_{\mathbf{Y}|\mathbf{X}}, q_{\mathbf{Y}|\mathbf{X}}) = I^p(\mathbf{X}; \mathbf{Y}) + \varepsilon(p, q). \quad (\text{A.9})$$

where as usual, $\mathbf{X} := X_{1:\ell} := W_{1:\ell}$ and $\mathbf{Y} := Y_{1:L-\ell} := W_{\ell+1:L}$ with $W_{1:L}$ being a sequence of tokens. However, as discussed in the main text, the $H(p_{\mathbf{Y}}, q_{\mathbf{Y}})$ term suffers from an additional bias—we cannot guarantee that \mathbf{Y} starts at the beginning of a sentence, but LLMs model distributions conditioned on BOS token. To mitigate this issue, we use n -gram calculations to correct the entropy of the first two tokens as explained below.

We first rewrite the (marginal) cross entropy as

$$H(p_{\mathbf{Y}}, q_{\mathbf{Y}}) = -\mathbb{E}_p[\log q(\mathbf{Y})] = -\sum_{i=1}^{L-\ell} \mathbb{E}_p[\log q(Y_i | Y_{1:i-1})], \quad (\text{A.10})$$

where as usual, the omitted expectation over the conditional variable is implied in the cross entropy calculation.

In modern LLMs, we can only compute $q(y_i | y_{1:i-1}, w_{BOS}) \neq q(y_i | y_{1:i-1})$, resulting in an additional error in the bipartite mutual information estimation. In practice, this difference becomes less pronounced for larger i , because it matters less if the sequence starts at the beginning of a sequence or not if there are many $y_{1:i-1}$ prior tokens to conditional on. Therefore, we focusing on reducing the bias for small i . In addition, if i is small, we can iterate over the dataset and construct a histogram for the i -gram distribution $p(y_{1:i})$.

We denote the count for each i -tuple of tokens with $n_{y_{1:i}}$ and the total number of samples with N . Then, the entropy of the distribution can be estimated naively as

$$\hat{H}^{\text{naive}}(Y_{1:i}) = -\sum_{y_{1:i}} \frac{n_{y_{1:i}}}{N} \log \frac{n_{y_{1:i}}}{N} = \log N - \frac{1}{N} \sum_{y_{1:i}} n_{y_{1:i}} \log n_{y_{1:i}}, \quad (\text{A.11})$$

where the summation runs over all possible combination of tokens $y_{1:i} := (y_1, y_2, \dots, y_i)$.

However, this estimation is severely biased and underestimates the true entropy, due to the concavity of logarithm function. In (Grassberger, 2008), a bias-corrected estimator is proposed by replacing the logarithm function with a new function

$$\hat{H}^G(Y_{1:i}) = \log N - \frac{1}{N} \sum_{y_{1:i}} n_{y_{1:i}} G(n_{y_{1:i}}), \quad (\text{A.12})$$

where

$$G(n) = \psi(n) + \frac{(-1)^n}{2} \left(\psi\left(\frac{n+1}{2}\right) - \psi\left(\frac{n}{2}\right) \right), \quad (\text{A.13})$$

with $\psi(\cdot)$ the digamma function. We note that Ref. (Grassberger, 2008) was not able to obtain the closed form expression for $G(\cdot)$, which we derived with the help of Wolfram Mathematica (Inc.).

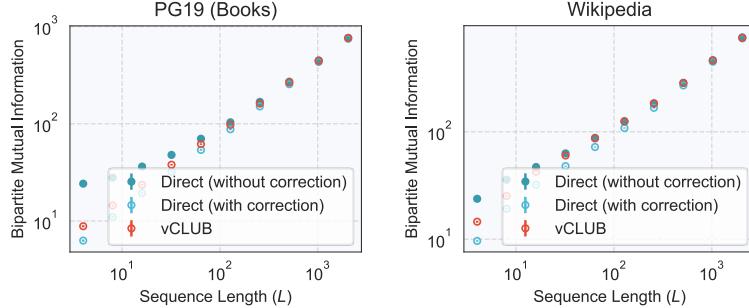


Figure A.1. Effect of bias correction method in the direct estimator. The bias only affects the estimation at small sequence lengths, and all methods converge at large sequence lengths.

This bias-corrected estimator still underestimates the true entropy, but much less compared to the original naïve estimator. In the main text, we estimate the (marginal) cross entropy with 2-gram correction in the following way. Breaking up the cross entropy as

$$H(p_{\mathbf{Y}}, q_{\mathbf{Y}}) = -\sum_{i=3}^{L-\ell} \mathbb{E}_p[\log q(Y_i | Y_{1:i-1})] + H(p_{Y_1 Y_2}, q_{Y_1 Y_2}). \quad (\text{A.14})$$

For the first term, we use LLM generated $q(y_i|y_{i:i-1}, w_{BOS})$ as approximation. For the second term, we mitigate the bias from LLM estimation by combining it with Eq. (A.12) as $H(p_{Y_1 Y_2}, q_{Y_1 Y_2|w_{BOS}})/5 + 4\hat{H}_p^G(Y_1, Y_2)/5$. In Fig. A.1, we also present the result without this correction and show that this bias correction mostly affects the estimation at small lengths L , and does not alter the generalizing scaling behavior. In addition, since the result from this bias-corrected direct estimator agrees with the vCLUB (Cheng et al., 2020) estimator, we believe this correction is reasonable.

A.IV. Bipartite Mutual Information Scaling Using Various LLMs

In the main text, we use the LLaMA 3.1 405B model to approximate the underlying distribution of natural language. In this section, we provide additional estimations of the bipartite mutual information scaling using the DeepSeek V3 Base model and LLaMA 3.1 70B model. We note that because we are merely measuring the conditional probabilities of the input tokens without interactions with the agent, we believe the non-instruction-finetuned model better suits our tasks.

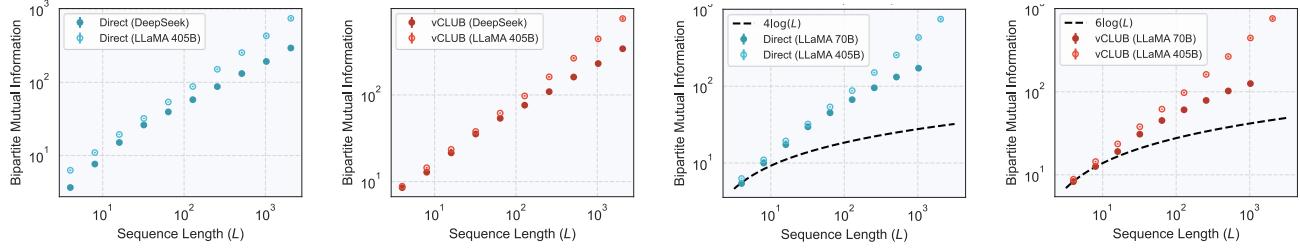


Figure A.2. Bipartite mutual information estimation using Deepseek V3 Base model and LLaMA 3.1 70B model compared to LLaMA 3.1 405B model on the PG19 dataset. All direct measurements include the bias correction described in Appx. A.III. Both models appear to produce worse estimations of bipartite mutual information at long sequence lengths, but still suggest a power-law scaling.

In Fig. A.2, we report the results on PG19 dataset and compare them to LLaMA 3.1 405B model’s results. We find that the DeepSeek V3 Base model estimates consistently lower bipartite mutual information compared to the LLaMA 3.1 405B model, especially at long sequence lengths. Based on our argument in Appx. A.VI, this suggests that the DeepSeek V3 Base model does not approximate the probability distribution at long sequence lengths as well as the LLaMA 3.1 405B model. We believe this could be attributed to the following reasons. (1) The DeepSeek V3 Base model is a mixture-of-expert (MOE) model. When we draw samples randomly from the dataset, the topic may not be immediately clear, whereas an MOE model usually needs to know the topic to correctly predict the conditional probabilities. (2) The DeepSeek model is trained without using the cross-sample attention masking (DeepSeek-AI et al., 2024) whereas the LLaMA model is (Grattafiori et al., 2024). Without the mask, concatenated training samples affect the model’s learned long-range dependencies. In spite of this, the DeepSeek model still largely suggests the existence of power-law growth of bipartite mutual information.

LLaMA 3.1 70B model, on the other hand, being a smaller model, estimates lower bipartite mutual information at large sequence lengths. It is unsurprising because a smaller model is expected to be a worse approximation to the true underlying distribution of natural language. In addition, since its predecessor, the LLaMA 3.1 70B model, only has a context window of 8192, this model, despite having a larger claimed context window, is likely not trained on as much high-quality long context length data as LLaMA 3.1 405B. Nevertheless, even the small LLaMA 3.1 70B model estimates the bipartite mutual information growth closer to power law than logarithmic.

A.V. Bipartite Mutual Information Scaling Under Various Ratios of ℓ/L

In the main text, we focused on the bipartite mutual information with equal splits. However, the bipartite mutual information scaling is not limited to equal bipartition. In this section, we provide additional results for various ratios of ℓ/L .

In Fig. A.3, we provide estimation of the bipartite mutual information scaling for $\ell/L = 3$ and $\ell/L = 4$. All results show clear power-law relations, and are consistent with Fig. 2 in the main text.

A.VI. Why The Estimated Exponent β Is Likely An Underestimation?

In the main text, we mentioned that our measured exponent β using LLMs likely underestimates the true β . Here, we discuss the reasons.

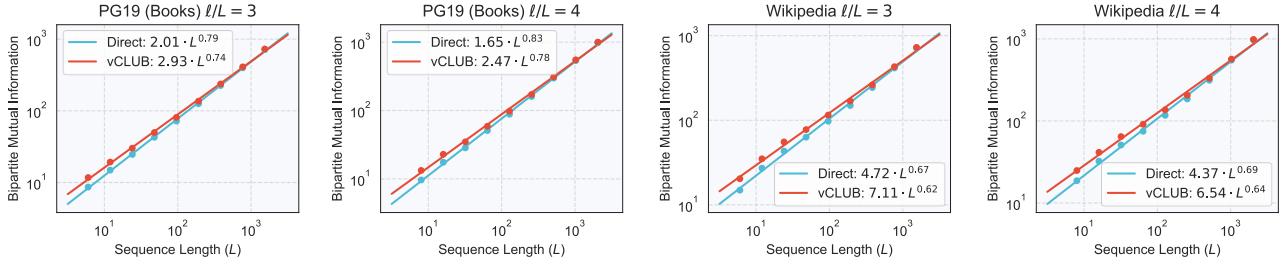


Figure A.3. Bipartite mutual information estimation using different ratios of ℓ/L . All results suggest the existence of power-law scaling, with various fitted exponents.

For the direct estimator,

$$I_{\ell;L}^{\text{BP,direct}} = H(p_{\mathbf{Y}}, q_{\mathbf{Y}}) - H(p_{\mathbf{Y}|\mathbf{X}}, q_{\mathbf{Y}|\mathbf{X}}), \quad (\text{A.15})$$

both terms (without the minus sign) overestimates the true (conditional) entropy, but for different extent and at different scales.

At small L , the first term suffers from the bias from the BOS token as discussed in Appx. A.III. The second term, despite also an overestimation, does not suffer from the BOS token issue. Therefore, at small L , the direct estimator tends to overestimate the true entropy.

At large L , the bias from the BOS token is less severe. However, modeling $p(\mathbf{Y}|\mathbf{X})$ requires the model to correctly capture all the dependencies between \mathbf{X} and \mathbf{Y} , making it significantly harder than modeling $p(\mathbf{Y})$ alone. Therefore, $q(\mathbf{Y}|\mathbf{X})$ is likely a worse estimation of the true distribution than $q(\mathbf{Y})$, resulting in more overestimation in the second term, and an underestimation of the bipartite mutual information.

This means that the direct estimator tends to overestimate the true bipartite mutual information at small L and underestimate it at large L , resulting in an underestimation of the fitted exponent.

The vCLUB estimator, as pointed out in (Cheng et al., 2020), is an upper bound to the true mutual information if q is close to p , but fails to maintain the property when the KL-divergence between them increases. Therefore, it is likely that this estimator also overestimates the true bipartite mutual information at small L and underestimates it at large L , resulting in a similar underestimation of the fitted exponent as our direct estimator. As our fitted exponent for the vCLUB estimator is smaller than that of the direct estimator, we conclude that the vCLUB estimator has a larger bias in this case, and it is reasonable to believe that the true exponent is even larger.

A.VII. Estimation of Two-Point Mutual Information

As discussed in the main text, the estimation of two-point mutual information can be calculated directly using the n -gram approximation [Eq. (A.12)], and compute the two-point mutual information as

$$\hat{I}_d^{\text{TP}}(X; Y) = \hat{H}^G(X) + \hat{H}^G(Y) - \hat{H}^G(XY). \quad (\text{A.16})$$

As discussed in Appx. A.III, this entropy estimator has a negative bias, whose magnitude depends on the ratio $|\Omega|/N$, with $|\Omega|$ the size of the corresponding sample space. Since the sample space for the joint distribution is larger, it has a larger negative bias, resulting in a positive bias in \hat{I} . When d is small, this bias is relatively small compared to the mutual information itself. However, as d becomes larger, X and Y become less correlated, and $H(XY) \rightarrow H(X) + H(Y)$. In this case, the estimator can be dominated by this bias, and fitting for the power-law exponent becomes impossible.

To mitigate this issue, we propose a bias-corrected estimator.

$$\hat{I}_d^{\text{TP,corrected}}(X; Y) = \hat{H}^G(X) + \hat{H}^G(Y) - \hat{H}^G(XY) - C, \quad (\text{A.17})$$

where C is an unknown positive constant that does not depend on the distance d , which accounts for the bias of the original estimator.

To obtain this bias correction term and fit the power-law exponent, we minimize the following loss function

$$\sum_d (\log(\hat{I}_d^{\text{TP}} - C) - (\log A - \alpha \log d))^2, \quad (\text{A.18})$$

which is just $\hat{I}_d^{\text{TP}} = Ad^{-\alpha} + C$ fitted in log-log space. Then, we take the fitted C as the systematic bias and α as the fitted power-law exponent.

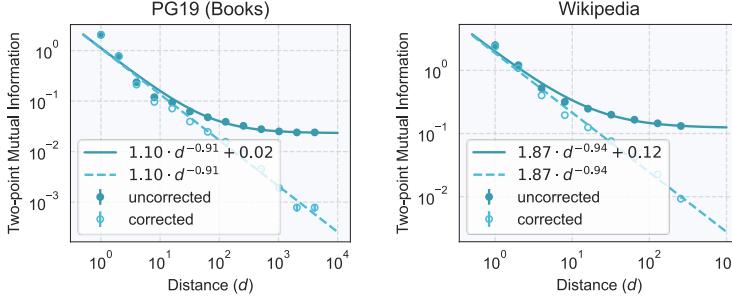


Figure A.4. Effect of bias correction for two-point mutual information. The bias causes a plateau at large distances.

In Fig. A.4, we compare the corrected and uncorrected two-point mutual information as a function of d (only the corrected version is shown in the main text). Without the bias correction, the data appear to have larger long-range dependencies, but after the bias correction, all points lie on a straight line in a log-log plot. The bias correction constant is much smaller than the entropies involved in the calculation, even the smallest two-token entropy measured is 12.5, at least two orders of magnitude larger than the fitted bias correction. In addition, a single variable added to the fitting function can fit the data so well. These suggest the bias correction is reasonable and highly effective.

We note that on WIKIPEDIA, we were only able to measure the two-point mutual information up to $d = 256$, due to limited long-context length data in WIKIPEDIA.

B. Multivariate Gaussian Distributions

In the main text, we considered two families of multivariate Gaussian distributions of different sequence lengths to demonstrate the distinction between bipartite and two-point mutual information scalings. In particular, one is designed to mimic natural language, both in terms of the sub-volume law growth of the bipartite mutual information and the power-law decay of two-point mutual information. This family of distributions is also used to empirically verify our theory on L²M condition for different LLM architectures. The other is designed to have the same two-point mutual information scaling, but very different bipartite mutual information scaling, showcasing that one can have distributions with the same two-point mutual information scaling, but drastically different bipartite mutual information scalings.

B.I. Construction

Let's start by considering the family of distributions with sub-volume law growth. The distributions are constructed in a hierarchical manner.

We start at the first layer, with four independent standard Gaussian random variables (X_1, X_2, X_3, X_4) . Then, define the change-of-coordinate matrix

$$\mathcal{M} = \begin{pmatrix} \gamma & \gamma & \gamma & \rho \\ -\gamma & \gamma & -\gamma & \rho \\ -\gamma & -\gamma & \gamma & \rho \\ \gamma & -\gamma & -\gamma & \rho \end{pmatrix}, \quad (\text{B.19})$$

where we choose $\gamma = \sqrt{5}/4$ and $\rho = 1/4$. The output of the first layer is defined as

$$\mathbf{Y} = \mathcal{M}\mathbf{X}, \quad (\text{B.20})$$

where the independent random variables are correlated. It is easy to verify that this operation only changes the off-diagonal elements in the covariance matrix, and leaves the diagonal elements unaffected.

For the second layer and up, we first stack three independently sampled copies from the previous layer and attach additional independent standard Gaussian random variables as the fourth elements as

$$\mathcal{X} = \begin{pmatrix} Y_{1,1} & Y_{1,2} & Y_{1,3} & W_1 \\ Y_{2,1} & Y_{2,2} & Y_{2,3} & W_2 \\ Y_{3,1} & Y_{3,2} & Y_{3,3} & W_3 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}, \quad (\text{B.21})$$

where \mathbf{X} refers to the input to the new layer, $Y_{i,j}$ refers to the i th output from previous layer of the j th copy, and W_i refers to the i th newly sampled standard Gaussian random variable.

Note that at this point, all rows are independent from each other; therefore we apply the change of coordinate matrix at each row to correlate them. The matrix is then flattened to obtain \mathbf{Z}

$$\mathbf{Z} = (Z_{1,1}, Z_{1,2}, Z_{1,3}, Z_{1,4}, Z_{2,1}, Z_{2,2}, Z_{2,3}, Z_{2,4}, Z_{3,1}, Z_{3,2}, Z_{3,3}, Z_{3,4}, \dots), \quad (\text{B.22})$$

where the subscripts denote the variables' original position in the matrix. Before outputting from this layer, we perform an addition operation to each pair of random variables $(Z_{i,4}, Z_{i+1,1})$, by applying a coordinate transformation that modifies their correlations as

$$\text{corr}(Z_{i,4}, Z_{i+1,1}) \rightarrow \frac{2}{5}(\text{corr}(Z_{i,3}, Z_{i,4}) + \text{corr}(Z_{i+1,1}, Z_{i+1,2})) + \frac{1}{5}. \quad (\text{B.23})$$

This operation may seem arbitrary, but it is crucial to introduce correlations that give a linear ordering of the random variables. Without this operation, the distribution simply forms a tree structure.

Now, we can truncate the construction at different layers l and form a family of distributions with different sequence lengths $L = 4^l$. In Fig. 3 of the main text, we consider up to 8 layers, and in Fig. 4 and 5, we consider $l = 4, 5$ and, 6.

The second family of distributions is constructed analogously. The only difference is that we replace Eq. (B.24) with a single copy of \mathbf{Y} and three independent copies of \mathbf{W} as

$$\mathcal{X} = \begin{pmatrix} Y_1 & W_{1,1} & W_{1,2} & W_{1,3} \\ Y_2 & W_{2,1} & W_{2,2} & W_{2,3} \\ Y_3 & W_{3,1} & W_{3,2} & W_{3,3} \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}, \quad (\text{B.24})$$

B.II. Properties

These two series of distributions have many nice properties, in addition to their bipartite and two-point mutual information scalings. Due to the analytic construction, their bipartite and two-point mutual information can be calculated exactly, without relying on LLM approximations. Moreover, they permit efficient sampling and evaluation of conditional probabilities. In addition, infinitely many samples can be drawn from the distribution, giving us an infinite dataset size.

C. Model State for Storing Past Information

In Definition 5.1 in the main text, we give a concrete definition of “model state for storing past information” as history state, and claim that it is the past key-value pairs for transformers and latent state for SSMs and RNNs. Here, we explain them in more detail.

C.I. Transformers

In transformers, only the attention block mixes information among different tokens, therefore we only need to analyze the behavior of the attention block. We will be assuming the existence of the causal mask, as our theory only applies to generative LLMs. Denoting the input and output of the attention layer as \mathbf{x} and \mathbf{y} (notice they are no longer two parts of the

same sequence), the self-attention mechanism is defined as

$$\mathbf{y} = \text{softmax}((W_q \mathbf{x})(W_k \mathbf{x})^T) W_v \mathbf{x}, \quad (\text{C.25})$$

where W_q , W_k and W_v are the weight matrices. For simplicity, we drop the usual $\sqrt{h_{\text{dim}}}$ normalization and the output weight matrix, as they are irrelevant to our discussion.

Separating the calculation for each token, the mechanism can be rewritten as

$$y_i = \frac{\sum_{j=1}^i e^{(W_q x_i)(W_k x_j)} W_v x_j}{\sum_{j'=1}^i e^{(W_q x_i)(W_k x_{j'})}} = \frac{e^{(W_q x_i)(W_k x_i)} W_v x_i + \sum_{j=1}^{i-1} e^{(W_q x_i)(W_k x_j)} k_j v_j}{e^{(W_q x_i)(W_k x_i)} + \sum_{j'=1}^{i-1} e^{(W_q x_i)(W_k x_{j'})}}, \quad (\text{C.26})$$

where $k_j = W_k x_j$ and $v_j = W_v x_j$ are keys and values which we sum over past tokens. Clearly, the attention output only depends on the current token x_i and the past key-value pairs $k_{1:i-1}$ and $v_{1:i-1}$. This argument extends to all y_k with $k \geq i$, where all y_k 's dependency on $x_{1:i-1}$ is via $k_{1:i-1}$ and $v_{1:i-1}$. Therefore, key-value pairs are the model state for storing past information, and their sizes grow linearly with input sequence length. We note that Eq. (C.26) also describes how key-value caching works.

C.II. State Space Models and RNNs

State space models (SSMs) and RNNs, on the other hand, are easier to analyze. These models in general all have some latent state, or hidden state, with a fixed size, and some mechanism to update the state when a new token is observed. The output depends only on the past latent state, and the current token. They can in general be written in the following way.

$$\begin{aligned} h_i &= f(h_{i-1}, x_i), \\ y_i &= g(h_{i-1}, x_i), \end{aligned} \quad (\text{C.27})$$

for some update function f and output function g . It is obvious that the “model state for storing past information” is exactly this latent state, which does not grow with the input sequence.

We note that this discussion also applies to linear attention models, whose key-value pairs can be merged into a latent state with fixed size, due to the replacement of softmax function. Test time training (TTT) models can also be included in this discussion. They can be viewed as SSMs with inner model parameters as latent state, and test time training process as update function.

C.III. Other Architectures

For other models, such as sparse transformers or some compression-based models, the analysis has to be performed separately. Nevertheless, in general, one wants to identify the smallest hidden state used to cache past information, as it is the bottleneck for modeling long sequences with large bipartite mutual information.

D. Additional Proofs for Theorem 5.2

In the main text, we only provided an intuitive proof based on the discreteness of model states. Here, we relax this assumption in different ways and provide additional proofs to the theorem. We start by restating the theorem.

Theorem D.1 (Restatement of Theorem 5.2). *A model's capacity to capture bipartite mutual information is bounded by the size of its history state as*

$$I_{\ell;L}^{\text{BP},q} \leq C \cdot \dim(\mathbf{z}_\ell) + \log(M) \quad (\text{D.28})$$

where C is some constant and M denotes the vocabulary size.

We start with a proof assuming the following observed fact of neural networks—neural networks store distinct information in almost orthogonal directions (AODs) of the hidden state (Elhage et al., 2022; Park et al., 2023; Jiang et al., 2024).

Proof. An autoregressive neural network's dependency on past tokens is through the intermediate variable $\mathbf{z}_\ell = \mathbf{f}(x_{1:\ell-1})$ such that $q(\mathbf{y}|\mathbf{x}) := q(\mathbf{y}|x_\ell, \mathbf{z}_\ell)$. This can be viewed as the process $\mathbf{X} \rightarrow (\mathbf{Z}_\ell, X_\ell) \rightarrow \mathbf{Y}$. According to the data processing inequality,

$$I^q(X_{1:\ell}; Y_{1:L-\ell}) \leq I^q(\mathbf{Z}_\ell X_\ell; Y_{1:L-\ell}) \leq \mathcal{H}^q(\mathbf{Z}_\ell X_\ell) \leq \mathcal{H}^q(\mathbf{Z}_\ell) + H^q(X_\ell), \quad (\text{D.29})$$

where we use \mathcal{H} to denote a generalized notion of entropy which measures the amount of information that can be stored in \mathbf{Z}_ℓ , because the standard definition of entropy fails for continuous random variables. Notice that because we only care about the general scaling of \mathcal{H} , its exact definition does not matter in this proof. Since neural networks store distinct information in AODs of the hidden state, \mathcal{H} should scale at most logarithmically with respect to the number of AODs as the state size increases. According to the Kabatjanskii–Levenshtein bound (Kabatiansky & Levenshtein, 1978; Cohn & Zhao, 2014), given an error tolerance ε , the number of AODs is upper bounded by $\exp(f(\varepsilon) \cdot \dim(\mathbf{z}_\ell))$ for some function f that purely depends on the error threshold. Therefore, the generalized entropy at most scales as $\mathcal{H}^q(\mathbf{Z}_\ell) \lesssim \log \exp(f(\varepsilon) \cdot \dim(\mathbf{z}_\ell)) \sim \dim(\mathbf{z}_\ell)$. It is easy to see that $H^q(X_\ell) \leq \log(M)$ with M being the vocabulary size, so we conclude

$$I_{\ell;L}^{\text{BP},q} \leq C \cdot \dim(\mathbf{z}_\ell) + \log(M). \quad (\text{D.30})$$

□

Alternatively, it is possible to prove the theorem only assuming certain Lipschitz continuous conditions.

Proof. This time, we again start with the data processing inequality, but we rewrite the bound as

$$I^q(X_{1:\ell}; Y_{1:L-\ell}) \leq I^q(\mathbf{Z}_\ell X_\ell; Y_{1:L-\ell}) \leq I^q(\mathbf{Z}_\ell; Y_{1:L-\ell}) + I^q(X_\ell; Y_{1:L-\ell}) \leq I^q(\mathbf{Z}_\ell; Y_{1:L-\ell}) + \log(M), \quad (\text{D.31})$$

where the last inequality used the fact that $I^q(X_\ell; Y_{1:L-\ell}) \leq H^q(X_\ell) \leq \log(M)$, with M again being the vocabulary size. Notice that the history state is a function of the input tokens $\mathbf{z}_\ell = \mathbf{f}(x_{1:\ell-1})$, with $x_{1:\ell-1} \in \{1, 2, \dots, M\}^{\ell-1}$. Assuming \mathbf{f} satisfies $\|\mathbf{f}(x_{1:\ell-1}) - \mathbf{f}(x'_{1:\ell-1})\|_2 \leq K_f \mathbb{1}(x_{1:\ell-1} \neq x'_{1:\ell-1})$, then $\mathbf{z}_\ell = \mathbf{f}(x_{1:\ell-1})$ lives in a d -dimensional ball of radius K_f , where d is the size of \mathbf{z}_ℓ . Now, consider a quantization of \mathbf{z}_ℓ as $\mathbf{Q}(\mathbf{z}_\ell)$, which maps each \mathbf{z}_ℓ to a nearest ε -ball that covers the full space of \mathbf{z}_ℓ . Then, we rewrite

$$I^q(\mathbf{Z}_\ell, \mathbf{Y}) = I^q(\mathbf{Q}(\mathbf{Z}_\ell), \mathbf{Y}) + H^q(\mathbf{Y}|\mathbf{Q}(\mathbf{Z}_\ell)) - H^q(\mathbf{Y}|\mathbf{Z}_\ell). \quad (\text{D.32})$$

□

We assume the neural network is entropy-Lipschitz, satisfying $|H^q(\mathbf{Y}|\mathbf{z}_\ell) - H^q(\mathbf{Y}|\mathbf{z}'_\ell)| \leq K_H \|\mathbf{z}_\ell - \mathbf{z}'_\ell\|_2$. Therefore $H^q(\mathbf{Y}|\mathbf{Q}(\mathbf{Z}_\ell)) - H^q(\mathbf{Y}|\mathbf{Z}_\ell) \leq K_H \varepsilon$. For the first term, notice that $\mathbf{Q}(\mathbf{Z}_\ell)$ is discrete, which takes a maximum of $(2K_f/\varepsilon)^d$ possible values due to covering number argument. Therefore, $I^q(\mathbf{Q}(\mathbf{Z}_\ell), \mathbf{Y}) \leq H^q(\mathbf{Q}) \leq d \log(2K_f/\varepsilon)$, which leads to $I^q(\mathbf{Z}_\ell, \mathbf{Y}) \leq d \log(2K_f/\varepsilon) + K_H \varepsilon \leq C \cdot d$ for some constant C . This concludes our proof

$$I_{\ell;L}^{\text{BP},q} \leq C \cdot d + \log(M) = C \cdot \dim(\mathbf{z}_\ell) + \log(M) \quad (\text{D.33})$$

We believe this theorem is more universal and can be proved in additional ways, such as by connecting it to the channel capacity and potentially showing $I_{\ell;L}^{\text{BP},q} \leq d \log(1 + \text{SNR}) + \log(M)$. We also believe the theorem can be established with more relaxed assumptions, similar to how the information dimension is proved to be the upper bound of lossless compression of continuous random variables (Kawabata & Dembo, 1994). However, additional proofs are beyond the scope of this work, and current proofs should be already applicable in any practical settings.

E. Additional Details on Experimental Setup

In this section, we provide additional details on the experimental setup.

For experiments on the multivariate Gaussian distribution, as mentioned in the main text, we stack 64 copies of the distribution and group the 64 Gaussian variables at each position to form a single token. More specifically, an example sample looks like

$$\mathbf{W} = (W_1, W_2, \dots, W_L) := ((Z_{1,1}, Z_{1,2}, \dots, Z_{1,64}), (Z_{2,1}, Z_{2,2}, \dots, Z_{2,64}), \dots, (Z_{L,1}, Z_{L,2}, \dots, Z_{L,64})), \quad (\text{E.34})$$

where again, the two subscripts (i, j) refers to the i th random variable from j th copy. In this way, the bipartite mutual information matches better with natural language, not only in scaling, but also in the coefficient (multiplicative constant). We in addition prepend an all zero token W_0 in each sample to mimic the effect of BOS token.

In order to process continuous random variables, we replace the embedding layers of GPT2 and Mamba(2) models with two layer MLPs. For output, since all the conditional distributions are also Gaussian, we use a different two layer MLP to outputs the 64 conditional means $\mu_{q_{Z_{i,j}|Z_{0:i-1,j}}}$ and standard deviations $\sigma_{q_{Z_{i,j}|Z_{0:i-1,j}}}$.

As discussed in Appx. B.II, due to the analytical construction, the Gaussian distribution permits efficient calculation of conditional probabilities. Therefore, instead of simply training the neural networks with negative log likelihood on the samples alone, we use the average conditional KL-divergence estimated as

$$D_{KL}(p||q_\theta) = \mathbb{E}_{p_Z} \left[\frac{1}{L} \sum_{i=1}^L \frac{1}{64} \sum_{j=1}^{64} \log \frac{\sigma_{q_{Z_{i,j}|Z_{0:i-1,j}}} \sigma_{p_{Z_{i,j}|Z_{0:i-1,j}}} + (\mu_{q_{Z_{i,j}|Z_{0:i-1,j}}} - \mu_{p_{Z_{i,j}|Z_{0:i-1,j}}})^2}{2\sigma_{q_{Z_{i,j}|Z_{0:i-1,j}}}^2} \right] - \frac{1}{2} \quad (\text{E.35})$$

instead to reduce the sampling variance.

For the Gaussian distribution training, during each iteration, we choose a batch size of 4 (4 times sequence length number of tokens) with freshly generated samples, meaning we never reuse any sample, and therefore only has a single epoch, thanks to the “infinite” dataset size. We train all neural networks using AdamW optimizer and a cosine decay scheduler with warmup. We choose a peak learning rate of 0.00005, a weight decay of 0.01, 2000 warmup steps, and 500000 training steps in total. The results reported are at the end of the training.

For the PG19 dataset, we train on standard average negative log likelihood. We first split the dataset into samples of around 5000 tokens, where we ensure each sample starts at the beginning of some sentence. Then we train the models for 5 epochs (around 235000 iterations) with a batch size of 8 (32768 tokens for sequence length of 4096). Similar to the other experiments, we train all neural networks using AdamW optimizer and a cosine decay scheduler with warmup. We choose a peak learning rate of 0.00005, a weight decay of 0.01 2000 warmup steps, and 500000 steps in total. The results reported are at the end of the training using a separate evaluation dataset containing 10000 samples. To keep consistency across different models, we always use the same tokenizer from GPT-Neo-X ([Black et al., 2022](#)).

In the main text, we report results on the position-wise conditional KL-divergence

$$D_{KL}^{(i)}(p||q_\theta) = \mathbb{E}_{p_W} [\log p(W_i|W_{1:i-1}) - \log q_\theta(W_i|W_{1:i-1})], \quad (\text{E.36})$$

average KL-divergence

$$D_{KL}^{Avg}(p||q_\theta) = \frac{1}{L} \sum_{i=1}^L D_{KL}^{(i)}(p||q_\theta), \quad (\text{E.37})$$

and position-wise conditional NLL

$$\text{NLL}^{(i)}(p||q_\theta) = -\mathbb{E}_{p_W} [\log q_\theta(W_i|W_{1:i-1})]. \quad (\text{E.38})$$

One can also define a average NLL as

$$\text{NLL}^{Avg}(p||q_\theta) = \frac{1}{L} \sum_{i=1}^L \text{NLL}^{(i)}(p||q_\theta), \quad (\text{E.39})$$

which we will use in Appx. F.

Our experiments are performed mostly on H100 GPUs, with varying VRAM size between 80GB and 94GB. Some experiments are performed on A100 GPUs with 80GB VRAM instead. We use the vLLM library when running inference to estimate the mutual information scaling. For DeepSeek V3 and LLaMA 3.1 405B models, we run the FP8 version of the models using 8 H100 GPUs (with 94GB VRAM); and for LLaMA 3.1 70B model, we run the FP16 version of the model using 4 H100 GPU (with 94GB VRAM). The model weights and configurations are downloaded from HuggingFace ([Wolf et al., 2020](#)).

When training GPT and Mamba(2) models on the Gaussian distribution, we use our custom library developed in PyTorch, ([Paszke et al., 2019](#)); when training GPT and Mamba models on the PG19 dataset, we use the trainer from HuggingFace `transformers` library. All models are initialized from scratch, with model configurations taken from HuggingFace. All training experiments are performed on individual H100 and A100 GPUs, with FP32 precision to avoid possible training

failures. Although training with FP16 would make the experiments run faster, it should not affect the actual results. We note that for Mamba2, we use the official implementation instead of the HuggingFace version, and for the GPT2 experiments on PG19, we re-implement the attention mechanism with FlexAttention (Li et al., 2024) to save memory, as the official FlashAttention (Dao et al., 2022; Dao, 2023; Shah et al., 2024) does not support FP32 precision.

The code for reproducing our mutual information estimation and the PG19 results is available at https://github.com/LSquaredM/mutual_info_scaling_law.

F. Additional Experimental Results

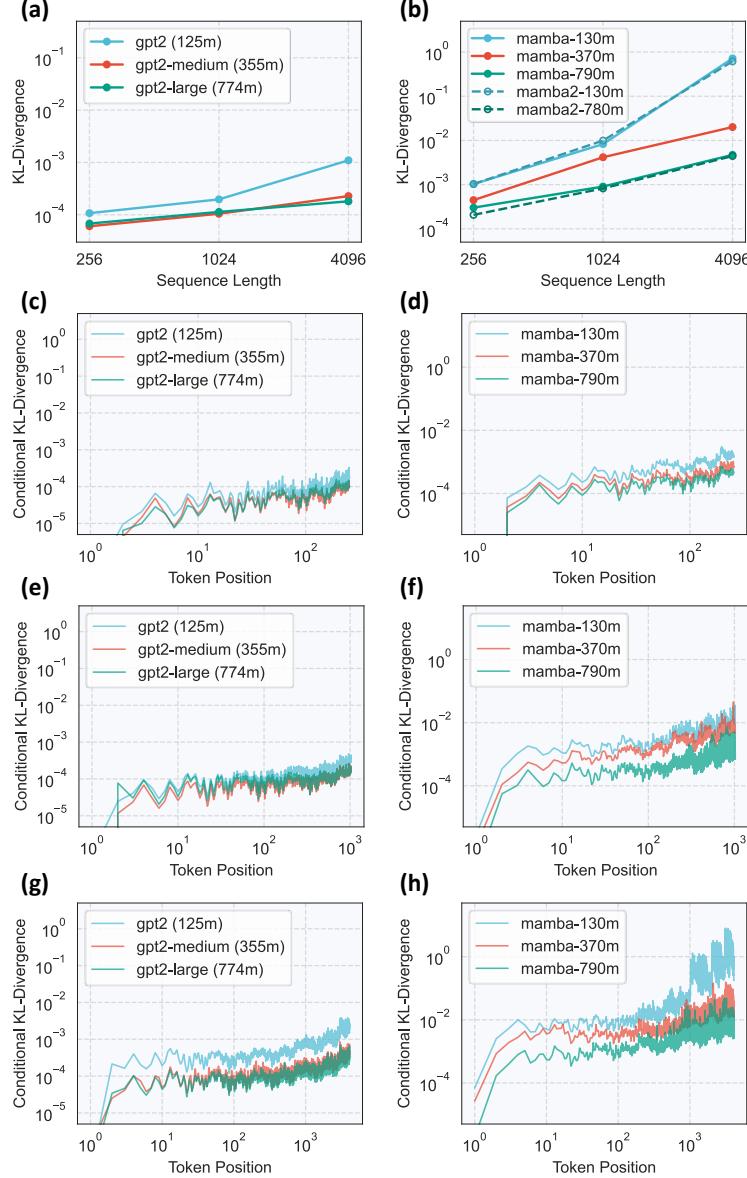


Figure F.5. Evaluation of KL-divergence across model architectures trained on sub-volume Gaussian distributions. (a, b) Average KL-divergence per token for models trained on different sequence lengths [same as Fig. 4 (a, b)]. (c, d) Position-wise conditional KL-divergence for models trained on sequence length 256. (e, f) Position-wise conditional KL-divergence for models trained on sequence length 1024. (g, h) Position-wise conditional KL-divergence for models trained on sequence length 4096 [same as Fig. 4 (c, d)]. Lower values indicate better performance.

In this section, we show additional experimental results. In Fig. F.5, we include additional positional-wise conditional KL-divergences of models trained on sub-volume Gaussian distributions with sequence length 256 (c, d) and sequence length 1024 (e, f). As clearly demonstrated in the figure, for short sequence lengths, Mamba maintains similar performances to GPT2; Mamba models of different sizes also appear to have a smaller performance gap. However, as we go to longer sequence lengths, smaller Mamba models start to fail, while GPT2 always maintain relatively stable performances, consistent with our theory.

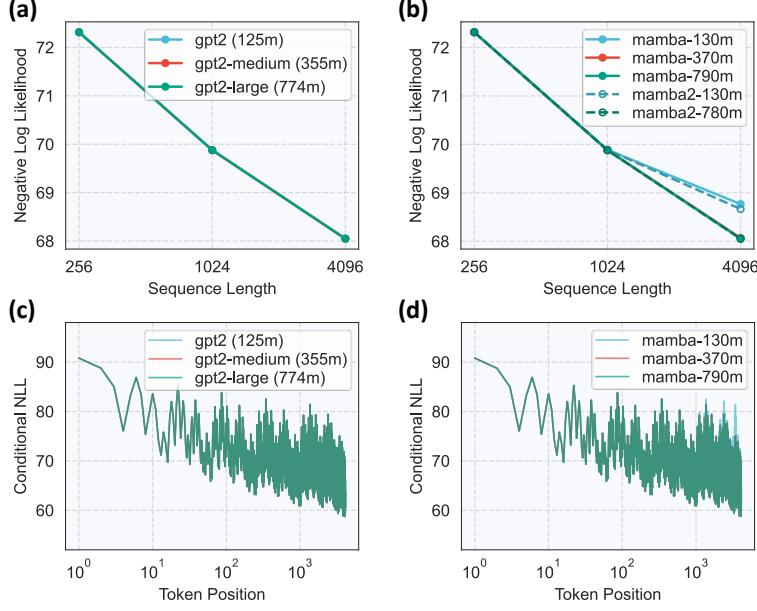


Figure F.6. Negative log likelihood (NLL) across model architectures trained on sub-volume Gaussian distributions (a, b) Average NLL per token for models trained on different sequence lengths. (c, d) Position-wise conditional NLL for models trained on sequence length 4096. Lower values indicate better performance.

In Fig. F.6, we show the negative log likelihood (NLL) of models trained on sub-volume Gaussian distributions. As mentioned in the main text, because NLL combines the KL-divergence with the intrinsic entropy of the underlying distribution (the average and position-wise conditional of which decays as sequence lengths), the differences between model performances are less visible. It's worth noting that, since Gaussian random variables are continuous, NLL values can differ by an arbitrary additive constant by rescaling the distribution. Therefore, the exact values of conditional NLL do not carry intrinsic meaning, though relative comparisons (which is exactly the same as the KL-divergence) between models remain valid.

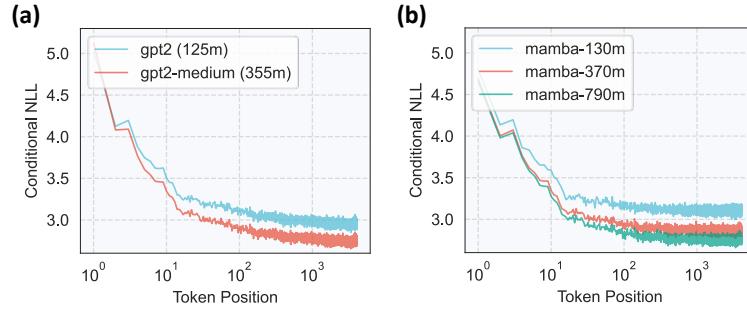


Figure F.7. Position-wise conditional negative log likelihood (NLL) evaluation for models trained on 4096-token sequences of the PG19 dataset.

In Fig. F.7, we show additional comparisons of the conditional NLL of models trained on PG19 dataset. Ignoring the decay of conditional NLL with token position, which is likely due to the decay in inherent conditional entropy of natural language, the relative differences between models exhibits similar trend as the sub-volume Gaussian results. Since the models have to

learn additional semantic meanings of natural language in addition to the long-range dependencies, we find it reasonable for smaller models to perform even worse on PG19 dataset than on the sub-volume Gaussian distribution. The results here are consistent with our theory.

Model	Average Negative Log Likelihood ↓
GPT2 (125M)	2.977
GPT2-medium (255M)	2.771
Mamba-130M	3.111
Mamba-370M	2.875
Mamba-790M	2.766

Table F.1. Average conditional negative log likelihood evaluation for models trained on 4096-token sequences of the PG19 dataset.

For completeness, we further show the average NLL of the models trained on PG19 dataset with a sequence length of 4096 in Table F.1, where results similar to Fig. F.7 is observed. Notice that, in Fig. F.7, the token position is plotted in log-scale; therefore, the average NLL is very close to the conditional NLLs at later token positions.