

Forget BIT, It is All about TOKEN: Towards Semantic Information Theory for LLMs

Bo Bai

Abstract

Large language models (LLMs) have demonstrated remarkable capabilities in numerous real-world applications. While the vast majority of research conducted from an experimental perspective is progressing rapidly, it demands substantial computational power, data, and other resources. Therefore, how to open the black-box of LLMs from a theoretical standpoint has become a critical challenge. This paper takes the theory of rate-distortion function, directed information, and Granger causality as its starting point to investigate the information-theoretic principles behind LLMs, leading to the development of semantic information theory for LLMs, where the fundamental unit is token, rather than bits that lacks any semantic meaning. By defining the probabilistic model of LLMs, we discuss structure-agnostic information-theoretic measures, such as the directed rate-distortion function in pre-training, the directed rate-reward function in post-training, and the semantic information flow in inference phase. This paper also delves deeply into the theory of token-level semantic embedding and the information-theoretically optimal vectorization method. Thereafter, we propose a general definition of autoregression LLM, where the Transformer architecture and its performance such as ELBO, generalization error bound, memory capacity, and semantic information measures can be derived theoretically. Other architectures, such as Mamba/Mamba2 and LLaDA, are also discussed in our framework. Consequently, this paper provides a theoretical framework for understanding LLMs from the perspective of semantic information theory, which also offers the necessary theoretical tools for further in-depth research.

I. INTRODUCTION

At the end of 2022, ChatGPT emerged and its capabilities stunned the entire world! A few month later, we fortunately invited Prof. Arikan, the inventor of Polar codes, for a panel discussion.¹ My colleague, Dr. Wu, hosted the event, his first question was brilliant: “Prof. Arikan, what do you consider the greatest invention of the information age?” After

Bo Bai, Lab Director and Chief Scientist of Information Theory, is with Theory Lab - Leibniz, Central Research Institute, 2012 Labs, Huawei Technology Co., Ltd., Hong Kong. Email: baibo8@huawei.com

¹The event was broadcast live through the Chaspark website and became the best live event of that year.

a moment of thought, the professor gave a decisive answer: “The **BIT**! I believe that the bit is the greatest invention of the information age.” This answer deeply shook me and has since inspired me to think about a question: What is the most important concept with the same fundamental importance as the bit in AI age, especially after ChatGPT emerged? After deeply involved into the research of LLMs, I finally realized that: the concept I am seeking is none other than the **TOKEN**.

Inspired by Shannon’s seminal 1948 paper [1], I tried to approach the explanation theory of LLMs from inference perspective. Shannon started with the goal of achieving reliable information transmission in a communication system. From that starting point, he laid out a complete set of mathematical concepts and theorems, which is known as information theory. In 1949, Weaver and Shannon co-authored a paper in which they clearly identified three levels of communication problems [2]. They are:

- **Level-A: Technical problem.** How accurately can the symbols of communication be transmitted?
- **Level-B: Semantic problem.** How precisely do the transmitted symbols convey the desired meaning?
- **Level-C: Effectiveness problem.** How effectively does the received meaning affect conduct in the desired way?

Shannon humbly suggested that his theory only solved the problem of reliable communication, i.e., Level-A technical problem. This is because, in Shannon’s theory, information is equivalent to uncertainty. He was not concerned with the meaning or significance of the transmitted message, but only with whether its binary representation was received without error. However, it is shown in our work that by extending Shannon’s theory to center on tokens, the underlying principles of LLMs can be explained from information-theoretic perspective, which will be referred to as semantic information theory.

Early research on semantics can be traced back to the work of Carnap, who had a series of brilliant discussions on this issue from the perspectives of empiricism, ontology, linguistics, and logic [3]–[5]. In the classic book [6], Carnap provides a comprehensive and systematic exposition of semantics and modal logic. The modern developments of these approaches are well summarized in [7], [8]. Deeply influenced by Carnap, Solomonoff proposed the concept of algorithmic probability and integrated it into Bayesian inference framework, thereby providing a formal theory of inductive inference [9]–[11]. In Solomonoff’s theory, the prior probability of a sequence is determined by its complexity. Therefore, the shortest program that can generate the sequence has the highest prior probability, which is referred

to as the universal prior. The length of this shortest program defined on a Turing machine is known as the Kolmogorov complexity of the sequence. In [12], [13], Kolmogorov complexity is introduced as a new logical basis for Shannon’s information theory based on computing complexity on a Turing machine. It can be seen that this is exactly about viewing a sequence from a generative perspective based on Turing machine. Based on the Solomonoff prior and Kolmogorov complexity, a universal reinforcement learning is proposed for sequence decision and AI agent [14]. However, calculating the Kolmogorov complexity of a sequence is a Turing-undecidable problem, which in turn makes the theories of Kolmogorov and Solomonoff difficult to apply in practice.

When we apply Kolmogorov complexity to the sample sequences of a random variable, the expected value is exactly the Shannon entropy [15], [16]. Therefore, it is believed that Shannon’s information theory is a probabilistic special case of Kolmogorov complexity theory. However, the probabilistic approach of information theory is more valuable for modern neural networks and LLMs, the core reason may lie in the computability of information-theoretic measures such as entropy, mutual information, and Kullback-Leibler (KL) divergence (or cross-entropy), and also the fact that they are easy to approximate from data in practice using other more easily computable quantities [17]. This concept is precisely took away from Sutton’s famous short essay [18], specifically the first sentence: “The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin.”

A key question of extending Shannon’s theory to center on tokens is how to represent semantics of a token in a computable form. Unfortunately, source coding in Shannon’s theory only concerns how to represent the original message with the minimum number of binary symbols, but not with the semantics of the source. The idea of representing and retrieving information with vectors can be traced back to the work in [19]. The vector representation became the semantic basis of information-retrieval system [20]. In [21], Bengio et. al was the first to propose simultaneously learning a low-dimensional, distributed representation for words, i.e., a word vector, as part of training a language model. This marked the first time the concept of word vectors was combined with neural networks. In [22], Mikolov et. al introduced two model architectures: CBOW and Skip-gram, which demonstrated that high-quality word vectors can be trained with great efficiency on massive text corpora using a simple neural network. In their following work [23], they showed that the learned word vectors exhibit linear substructures that capture meaningful semantic relationships between words. This finding was groundbreaking and sparked a wave of research on word embeddings,

leading to the development of various models such as GloVe [24], FastText [25], and ELMo [26]. The vector representation of semantics has become the foundation of modern NLP and LLMs [27].

The vector representation, however, is only token-level semantics. How to extend the semantic representation and generation to a sentence, a paragraph, or even an article in a computing efficient way has long been a challenging problem. The advent of the Transformer [28], an architecture founded on the attention mechanism, represented a critical breakthrough, delivering extraordinary potential on NLP tasks. Subsequently, OpenAI introduced a series of GPT models built upon the Transformer architecture, which have exhibited remarkable capabilities in diverse applications [29]–[32]. Based on the classic Transformer architecture, DeepSeek has proposed a suite of enhancements aimed at substantially enhancing training efficiency. Consequently, the published LLMs exhibit remarkable inferential power [33], [34]. However, there still lacks a deep theoretical understanding of the principles behind the Transformer architecture. Therefore, improving the architecture and further enhancing LLM capabilities relies heavily on large-scale experiments on GPUs, which in turn requires an immense investment of resources.

Numerous studies have found that information-theoretic methods have been applied to many aspects of machine learning and have played a significant role [35]. The information bottleneck method, employed to analyze the mechanics of deep learning, has gained significant attention within academia and industry [36]. In [37], the rate-distortion function and information bottleneck method are applied to explain the semantic embedding for LLMs. The language model based textual transform coding is proposed for sharply improving the compression performance of multimedia [38]. To capture both the fidelity and the reality at the same time, the rate-distortion-perception function is surveyed for generative models in our work [39]. The Transformer is modeled as an interacting particle system, with a particular emphasis on long-time clustering behavior [40]. The centrality of data to LLM training underscores the significance of information-theoretic methods in data science, which is comprehensively reviewed in [41]. However, the autoregression LLM (AR-LLM), such as Transformer architecture, have not to be systematically studied from an information-theoretic perspective.

This paper leverages semantic information theory to construct a theoretical framework for understanding LLMs. We first propose a probabilistic model for LLM as a next-token predictor, which reveals it as a discrete-time channel with feedback and state. A significant modification to Shannon’s theory is to treat the channel as a generative model instead

of a media for information transmission. The objective shifts from exactly recovering the original information to ensuring the generated sequence meets specific requirements. This perspective leads us to propose the directed rate-distortion function as a universal measure for LLMs in the pre-training phase [42], [43]. The directed rate-reward function is also introduced for the reinforcement learning based post-training phase [44], which shows that the LLM is approximating Granger causality at a human level for next-token prediction [45]. The semantic information flow is defined and analyzed from the perspective of submartingale for the inference phase. Focusing on the foundations of LLMs, we then delve into the token-level semantic space and its vectorization. The semantic vector compression and the Gromov-Wasserstein distance based semantic distortion metric are discussed [46], [47]. Based on this groundwork, an information-theoretically optimal semantic vectorization method is introduced for next-token prediction. Its connection to contrastive predictive coding (CPC) is also examined [48], [49]. Thereafter, premised on the theory of time-varying vector autoregression (TV-VAR) processes, we formally establish a general mathematical definition for AR-LLMs [50]. It is demonstrated that the Transformer architecture constitutes a specialized case of this general AR-LLM formulation [28]. Based on the variational inference principle, the evidence lower bound (ELBO) of Transformer is derived for both training phase and inference phase [51]. The generalization error bound for Transformer is analyzed by using Rademacher complexity and Talagrand inequality [52]. The memory capacity, referred to as Gardner capacity for Hopfield network, is discussed for Transformer [53]–[55]. The semantic information theoretical measure for LLMs, is discussed from the perspective of directed information estimation. The connection between AR-LLM and other novel architectures, such as Mamba/Mamba2 and LLaDA, are also discussed [56]–[58].

The rest of this paper is organized as follows. Section II presents the key concepts. In Section III, the LLM is studied as a next-token predictor. Section IV discusses the vector representation of token-level semantics. The general definition of AR-LLMs is proposed in Section V, where the Transformer architecture is thoroughly studied. Other LLM architectures are also discussed in this section. Finally, Section VI concludes this paper.

II. PRELIMINARIES

In this section, we will introduce the rate-distortion function, the directed information, and Granger causality, which will play key roles for understanding LLMs in subsequent discussions.

A. Rate-distortion Function

Rate-distortion theory, proposed by Shannon [1] and systematically discussed in [43], addresses the problem of determining the minimum rate R bits/symbol, so that the source symbol can be approximately reconstructed at the receiver without exceeding an expected distortion D .

Definition 1: The rate-distortion function for a source sequence $X_{1:n}$ with a non-negative distortion measure d is defined as

$$R(D) = \lim_{n \rightarrow \infty} \frac{1}{n} \inf_{P(\hat{X}_{1:n}|X_{1:n}): \mathbb{E}\{d(X_{1:n}, \hat{X}_{1:n})\} \leq D} I(X_{1:n}; \hat{X}_{1:n}), \quad (1)$$

where $\hat{X}_{1:n}$ is the output of the lossy source codec.

The rate-distortion function is in general very difficult to compute, where the classical Blahut-Arimoto algorithm is proposed in [59], [60]. Recently, we proposed a communication optimal transport approach and a constrained Blahut-Arimoto algorithm to compute the rate-distortion function and the rate-distortion-perception function [61]–[63].

B. Directed Information

In information theory, the directed information is first defined by Massey in his pioneer work [42] for discussing the channel with feedback. This idea was systematically developed for extensive channels with feedback in [64]. Let $X_{1:n}$ and $Y_{1:n}$ be two random sequences with $n \in \mathbb{N}$, we then have the following definition.

Definition 2: The directed information from $X_{1:n}$ to $Y_{1:n}$ is defined as

$$I(X_{1:n} \rightarrow Y_{1:n}) = \sum_{t=1}^n I(X_{1:t}; Y_t | Y_{1:t-1}). \quad (2)$$

Following this idea, we introduce the backward directed information from $X_{n:1}$ to $Y_{1:n}$ as follows:

Definition 3: The backward directed information from $X_{n:1}$ to $Y_{1:n}$ is defined as

$$I(X_{n:1} \rightarrow Y_{1:n}) = \sum_{t=1}^n I(X_{t+1:n}; Y_t | Y_{1:t-1}). \quad (3)$$

The information density, first proposed by Dobrushin in [65], has been widely used in finite blocklength information theory and machine learning [35]. Similarly, we introduce the directed information density.

Definition 4: The directed information density from $X_{1:n}$ to $Y_{1:n}$ is defined as

$$\imath(X_{1:n} \rightarrow Y_{1:n}) = \sum_{t=1}^n \imath(X_{1:n}; Y_t | Y_{1:t-1}), \quad (4)$$

where

$$i(X_{1:n}; Y_t | Y_{1:t-1}) = \log \frac{P(Y_t | Y_{1:t-1}, X_{1:n})}{P(Y_t | Y_{1:t-1})}. \quad (5)$$

Similar to the rate-distortion function, it is also very difficult to compute directed information in practice. The classical Blahut-Arimoto algorithm has been extended to maximize directed information in [66]. Inspired by the idea of mutual information neural estimator (MINE) [67], the directed information neural estimator (DINE) is proposed in [68]. A seminal work of computing information density is proposed by Strassen in [69].

C. Granger Causality

Granger, the Nobel prize winner of 2003, proposed a general definition of causality in [45], which is referred to as Granger causality afterwards.

Definition 5: Let $\mathcal{U}_t$ be all the knowledge in the universe available at time t with $1 \leq t \leq n$, $\mathcal{U}_t^-$ be the knowledge in the modified universe in which $X_{1:n}$ is excluded, X_t is said to cause Y_{t+1} if

$$P(Y_{t+1} \in \mathcal{A} | \mathcal{U}_t) \neq P(Y_{t+1} \in \mathcal{A} | \mathcal{U}_t^-). \quad (6)$$

This definition is general but not operational. In [70], several version of operational definition have been discussed, where the directed information or transfer entropy are proposed as a strength measure of Granger causality. As a finite length version of directed information, the transfer entropy is first introduced in [71]. In many following works, Granger causality is shown to be equivalent to directed information or transfer entropy for Gaussian vector autoregression (VAR) processes [72]. In fact, Massey also discussed the causality for communication system with feedback in his seminal work [42].

The directed information, transfer entropy, and Granger causality are widely used in physics, neuroscience, social networks, and finance [73]. From the perspective of [74], however, Granger causality is classified as statistical rather than causal.

III. LLM AS A NEXT-TOKEN PREDICTOR

Inspired from information theory, this section will introduce the probabilistic model and architecture irrelevant properties for LLMs.

A. Probabilistic Model of LLMs

The probabilistic model of LLMs is illustrated in Fig. 1. The input token sequence is $X_{1:n}$ with $1 \leq n < t \leq T$ and $n \in \mathbb{N}$, which will be mapped to semantic vector sequence $S_{1:n}$

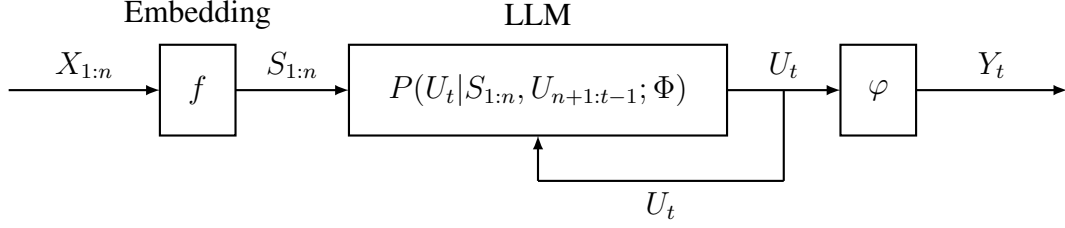


Fig. 1. The probabilistic model of an LLM at time $t \in \mathbb{N}$, where $X_{1:n}$ is the token sequence with $1 \leq n < t \leq T$ whose semantic vector embedding is $S_{1:n}$, $Y_{n+1:T}$ is the output token sequence, whose semantic vector embedding is $U_{n+1:T}$. Φ represents the parameters after training.

by a semantic embedding module f . The LLM is modeled as a transition probability with a parameter Φ , which represents the parameters of the LLM after training. The LLM generates the embedding of next token U_t based on $S_{1:n}$ and the previously generated $U_{n+1:t-1}$, that is

$$P(U_t | U_{n+1:t-1}, S_{1:n}; \Phi). \quad (7)$$

φ is an inverse module of embedding, which maps U_t to the output token Y_t . It should be noticed that the probabilistic model in Fig. 1 is general and architecture irrelevant.

Remark 1 (Kolmogorov Complexity Formulation of LLMs): The Kolmogorov complexity $K(\mathbf{y})$ is defined as the length of the shortest program that generates the output $\mathbf{y}$, formally written as

$$K(\mathbf{y}) = \min_{\mathbf{p}} \{l(\mathbf{p}) : T(\mathbf{p}) = \mathbf{y}\}, \quad (8)$$

where T is a universal Turing machine, $\mathbf{p}$ is the program, and $l(\mathbf{p})$ is the length of $\mathbf{p}$. According to [75], the Kolmogorov complexity can be rewritten as

$$K(\mathbf{y}) = \min_{i, \mathbf{p}} \{K(i) + l(\mathbf{p}) : T_i(\mathbf{p}) = \mathbf{y}\} \quad (9)$$

where $i \in \mathbb{N}$ is the index of a sequence of Turing machines. It can be seen that $K(\mathbf{y})$ is decomposed into two parts: the first part is a Turing machine T_i , i.e., the meaningful information or model in the data, and the second part is the irregular aspects of $\mathbf{y}$, i.e., a program $\mathbf{p}$ to be interpreted by T_i . Following this idea, the LLM is equivalent to T_i , $\mathbf{p}$ is the input $\mathbf{x}$, i.e., the prompt.

B. Directed Rate-distortion Function in Pre-training Phase

From the perspective of information theory, Eq. (7) is a discrete-time channel with feedback and state [42], [64]. The input is $S_{1:n}$, the output is $U_{n+1:T}$, the feedback at time t is U_t , and the channel state is the parameter Φ . In contrast to the reliable communication problem, the

goal here is to ensure the output token sequence aligns with our expectations, rather than a flawless recovery of the input.

As a discrete-time channel with feedback and state, the directed information is a natural choice to measure the information transferred from $S_{1:n}$ to $U_{n+1:T}$ with parameter Φ [42], [64]. According to Definition 2, we have

$$I(S_{1:n} \rightarrow U_{n+1:T}; \Phi) = \sum_{t=n+1}^T I(S_{1:n}; U_t | U_{n+1:t-1}; \Phi). \quad (10)$$

Let $U_{n+1:T}^h$ be the labeled sequence by human being with the input $S_{1:n}$, and $D_{KL}(\cdot \| \cdot)$ be the KL divergence. Denote $P_t^h = P(U_t^h | U_{n+1:t-1}^h, S_{1:n})$ and $Q_t^\Phi = P(U_t | U_{n+1:t-1}, S_{1:n}; \Phi)$ for $t = n+1, \dots, T$, we then have the following definition.

Definition 6: The directed rate-distortion function for LLMs in the pre-training phase is defined as

$$R_{pre}(D) = \frac{1}{T} \inf_{\Phi: \frac{1}{T} \sum_{t=n+1}^T D_{KL}(P_t^h \| Q_t^\Phi) < D} I(S_{1:n} \rightarrow U_{n+1:T}; \Phi). \quad (11)$$

Similar to Shannon capacity, $R_{pre}(D)$ is defined as a universal measure connecting the input sequence $S_{1:n}$ and output sequence $U_{n+1:T}$. Furthermore, $R_{pre}(D)$ is independent of any implementation methods, such as Transformer or novel architectures yet to be conceived. In contrast to the classical rate-distortion function in Definition 1, which governs a lossy source codecs, the output sequence $U_{n+1:T}$ in this context is instead constrained by a condition defined in terms of KL divergence. Therefore, $R_{pre}(D)$ is the minimum information needed from $S_{1:n}$ to generate the expected $U_{n+1:T}$ with an average distortion D . The curve of $R_{pre}(D)$ versus the optimization process of Φ will reveal key properties of the pre-training in practice. Simple derivation will give us the following theorem.

Theorem 1: In the pre-training phase with the cross-entropy loss, we have

$$R_{pre}(0) = \frac{1}{T} I(S_{1:n} \rightarrow U_{n+1:T}^h), \quad (12)$$

when convergence.

Proof: The cross-entropy between P_t^h and Q_t^Φ is given by

$$H(P_t^h, Q_t^\Phi) = H(P_t^h) + D_{KL}(P_t^h \| Q_t^\Phi), \quad t = n+1, \dots, T. \quad (13)$$

Thus, the objective of pre-training can be written as

$$\min_{\Phi} H(P_t^h, Q_t^\Phi) \Leftrightarrow \min_{\Phi} D_{KL}(P_t^h \| Q_t^\Phi), \quad t = n+1, \dots, T. \quad (14)$$

The minimization is achieved by adjusting Φ such that

$$Q_t^{\Phi^h} = P(U_t | U_{n+1:t-1}, S_{1:n}; \Phi^h) = P(U_t^h | U_{n+1:t-1}^h, S_{1:n}) = P_t^h, \quad t = n+1, \dots, T, \quad (15)$$

where Φ^h is the optimal solution of Eq. (14). It implies that

$$D = D_{KL}(P_t^h \| Q_t^{\Phi^h}) = 0, \quad (16)$$

when convergence. Recalling Definition 2, we have

$$\begin{aligned} I(S_{1:n} \rightarrow U_{n+1:T}; \Phi^h) &= \sum_{t=n+1}^T I(S_{1:n}; U_t | U_{n+1:t-1}; \Phi^h) \\ &= \sum_{t=n+1}^T H(U_t | U_{n+1:t-1}; \Phi^h) - \sum_{t=n+1}^T H(U_t | U_{n+1:t-1}, S_{1:n}; \Phi^h) \\ &= \sum_{t=n+1}^T H(U_t^h | U_{n+1:t-1}^h) - \sum_{t=n+1}^T H(U_t^h | U_{n+1:t-1}^h, S_{1:n}) \\ &= I(S_{1:n} \rightarrow U_{n+1:T}^h). \end{aligned} \quad (17)$$

This theorem has been established. ■

The aforementioned definition and theorem show that minimizing the directed information by adjusting Φ filters out information irrelevant to generate the output, which may effectively prevent hallucinations caused by the propagation of extraneous information by LLMs. Therefore, we suggest to use the following loss function for LLM pre-training:

$$\mathcal{L}(\Phi) = I(S_{1:n}; U_t | U_{n+1:t-1}; \Phi) + \lambda H(P_t^h, Q_t^\Phi), \quad t = n+1, \dots, T, \quad (18)$$

where λ is the Lagrangian multiplier.

Remark 2 (Information Geometry and Pre-training): Consider the pre-training phase, the distribution before and after one training step is denoted by $P_t(\Phi) = P(U_t | U_{n+1:t-1}, S_{1:n}; \Phi)$ and $P_t(\Phi') = P(U_t | U_{n+1:t-1}, S_{1:n}; \Phi')$, respectively. According to [76], the entry of the Fisher information matrix at the i -th row and j -th column is given by

$$[\mathcal{I}(\Phi)]_{ij} = \frac{\partial^2}{\partial \Phi'_i \partial \Phi'_j} H(P_t(\Phi), P_t(\Phi')) \Big|_{\Phi'=\Phi}. \quad (19)$$

Thus, the Fisher information matrix represents the curvature of the cross-entropy loss with respect to the parameters Φ . By modifying the gradient with Fisher information matrix, the natural gradient method is then proposed for neural network training. Due to the high computation complexity and storage cost, the Kronecker-factored approximate curvature method is used in practice [77].

C. Directed Rate-reward Function in Post-training Phase

The objective of pre-training is to accurately predict the next-token. The generated token sequence, however, may not follow the human preference. The post-training shifts the focus to

evaluate whether the entire generated sequence aligns with human preferences by fine-tuning with reinforcement learning from human feedback (RLHF) [32].

An evaluation function $w(S_{1:n}, U_{n+1:T})$, the reward function in RLHF, is introduced to assign a score to the generated sequence $U_{n+1:T}$ for the input sequence $S_{1:n}$. We then have the following definition.

Definition 7: The directed rate-reward function for LLMs in the post-training phase is defined as

$$R_{post}(W) = \frac{1}{T} \inf_{\Phi^h: w(S_{1:n}, U_{n+1:T}) > W} I(S_{1:n} \rightarrow U_{n+1:T}; \Phi^h). \quad (20)$$

Therefore, we suggest to use the following loss function for LLM post-training:

$$\mathcal{L}(\Phi^h) = I(S_{1:n} \rightarrow U_{n+1:T}; \Phi^h) - \lambda w(S_{1:n}, U_{n+1:T}), \quad (21)$$

where λ is the Lagrangian multiplier. The optimization solution will be denoted as Φ^{h+} . Recalling the proof of Theorem 1, $\mathcal{L}(\Phi^h)$ is equivalent to the loss function of RL fine-tuning phase in [78].

Theorem 1 shows that the LLM approaches $I(S_{1:n} \rightarrow U_{n+1:T}^h)$ during pre-training, which measures the information transferred from $S_{1:n}$ to $U_{n+1:T}^h$ by human being. The post-training further adjusts the parameter from Φ^h to Φ^{h+} such that the generated sequence $U_{n+1:T}$ meets human preferences. Recalling the discussion in Section II-C, we have the following conclusion.

Corollary 1: The LLM approaches the human-level Granger causality for next-token prediction with human preference after training.

D. Semantic Information Flow in Inference Phase

During the inference phase, the LLM with parameter Φ^{h+} is employed to generate the output token sequence $U_{n+1:T}$ based on the input token sequence $S_{1:n}$. In contrast to the post-training phase, where the focus is on the average performance across all possible output sequences, the inference phase considers the specific output sequence for the given input sequence. Therefore, it is natural to use the directed information density in Definition 4 to analyze the inference process. The semantic information flow can then be defined as follows.

Definition 8: The semantic information flow for LLMs is defined as the directed information density from $S_{1:n}$ to $U_{n+1:t}$ as follows:

$$\imath(S_{1:n} \rightarrow U_{n+1:t}; \Phi^{h+}) = \sum_{\tau=n+1}^t \imath(S_{1:n}; U_{\tau} | U_{n+1:\tau-1}; \Phi^{h+}), \quad t = n+1, \dots, T. \quad (22)$$

In the inference phase, the generation will stop when a special token, denoted by $\triangleleft$, is generated. Thus, T is the stopping time with respect to the event $\{U_T = \mathbf{s}(\triangleleft)\}$, where the vector representation of $\triangleleft$ is $\mathbf{s}(\triangleleft)$. We then have the following theorem.

Theorem 2: The semantic information flow $\imath(S_{1:n} \rightarrow U_{n+1:t}; \Phi^{h+})$ is a Markovian submartingale for $t = n + 1, \dots, T$.

Proof: According to the Definition 4, we have

$$\imath(S_{1:n} \rightarrow U_{n+1:t}; \Phi^{h+}) = \imath(S_{1:n} \rightarrow U_{n+1:t-1}; \Phi^{h+}) + \imath(S_{1:n}; U_t | U_{n+1:t-1}; \Phi^{h+}), \quad (23)$$

and

$$\imath(S_{1:n}; U_t | U_{n+1:t-1}; \Phi^{h+}) = \log \frac{P(U_t | U_{n+1:t-1}, S_{1:n}; \Phi^{h+})}{P(U_t | U_{n+1:t-1}; \Phi^{h+})}. \quad (24)$$

Thus, we consider the conditional expectation as follows:

$$\begin{aligned} & \mathbb{E}\{\imath(S_{1:n} \rightarrow U_{n+1:t}; \Phi^{h+}) | \imath(S_{1:n} \rightarrow U_{n+1:t-1}; \Phi^{h+}), \dots, \imath(S_{1:n} \rightarrow U_{n+1}; \Phi^{h+})\} \\ &= \mathbb{E}\{\imath(S_{1:n} \rightarrow U_{n+1:t}; \Phi^{h+}) | \imath(S_{1:n} \rightarrow U_{n+1:t-1}; \Phi^{h+})\} \\ &= \imath(S_{1:n} \rightarrow U_{n+1:t-1}; \Phi^{h+}) + \mathbb{E}\{\imath(S_{1:n}; U_t | U_{n+1:t-1}; \Phi^{h+})\} \\ &= \imath(S_{1:n} \rightarrow U_{n+1:t-1}; \Phi^{h+}) + D_{KL}(P(U_t | U_{n+1:t-1}, S_{1:n}; \Phi^{h+}) \| P(U_t | U_{n+1:t-1}; \Phi^{h+})) \\ &\geq \imath(S_{1:n} \rightarrow U_{n+1:t-1}; \Phi^{h+}). \end{aligned} \quad (25)$$

The last inequality holds because the KL divergence is non-negative, which establishes this theorem. $\blacksquare$

In the following, we will discuss the properties of semantic information flow as a submartingale. According to Doob decomposition, we have

$$\imath(S_{1:n} \rightarrow U_{n+1:t}; \Phi^{h+}) = M_t + A_t, \quad (26)$$

where A_t is a predictable and non-decreasing process

$$A_t = \sum_{j=n+1}^t \mathbb{E}\{\imath(S_{1:n} \rightarrow U_{n+1:j}; \Phi^{h+}) - \imath(S_{1:n} \rightarrow U_{n+1:j-1}; \Phi^{h+}) | \imath(S_{1:n} \rightarrow U_{n+1:j-1}; \Phi^{h+})\} \quad (27)$$

and M_t is a martingale

$$\begin{aligned} M_t &= \imath(S_{1:n} \rightarrow U_{n+1}; \Phi^{h+}) \\ &+ \sum_{j=n+2}^t (\imath(S_{1:n} \rightarrow U_{n+1:j}; \Phi^{h+}) - \imath(S_{1:n} \rightarrow U_{n+1:j-1}; \Phi^{h+}) - A_j). \end{aligned} \quad (28)$$

Define the sum of the conditional variances of the differences as

$$V_t = \sum_{j=n+1}^t \mathbb{E}\{(M_j - M_{j-1})^2 | M_{j-1}, \dots, M_{n+1}\}. \quad (29)$$

The following corollary can be directly established according to Freedman's inequality [79].

Corollary 2: For all $\alpha, \beta > 0$, we have

$$\Pr\{M_t > \alpha, V_t < \beta\} \leq \exp\left(-\frac{\alpha^2}{2(\alpha + \beta)}\right). \quad (30)$$

According to Doob's optional stopping time theorem [80] for sub-martingale, we have the following corollary directly.

Corollary 3:

$$I(S_{1:n} \rightarrow U_{n+1:T}; \Phi^{h+}) \geq I(S_{1:n} \rightarrow U_{n+1}; \Phi^{h+}). \quad (31)$$

Sharing the same spirit of Shannon capacity, i.e., the maximum mutual information over all input distributions, this corollary inspired us to give the following definition.

Definition 9: The semantic information capacity for LLMs is defined as

$$\max_{P(S_{1:n}):w(S_{1:n}, U_{n+1:T}) > W} I(S_{1:n} \rightarrow U_{n+1:T}; \Phi^{h+}). \quad (32)$$

Eq. (32) can be seen as a theoretical foundation for prompt engineering.

IV. VECTOR REPRESENTATION OF TOKEN-LEVEL SEMANTICS

A prerequisite for the efficient training of LLMs is the effective representation of token-level semantics. This section will first define the token-level semantic space, and then elaborate on the vector representation of semantics, semantic compression/de-dimensionality, and the information-theoretic optimal semantic embedding/vectorization.

A. Token-level Semantic Space

While grammatical and logical rules are central to how human being communicate and think, they are of indirect utility for the automated and computationally efficient processing of natural language by machines. As a starting point, we will disregard the use of intrinsic grammatical and logical structure of a natural language, considering it solely from a probabilistic standpoint.

Definition 10: The token-level semantic space of a language is a probabilistic space $(\Omega, \mathcal{F}, P)$, where $|\Omega| = N \geq 1$ is a set of all tokens, each of which is the atomic unit with specific semantics in this language, $\mathcal{F} \subseteq 2^\Omega$ is the σ -algebra, P is the probability measure defined on $\mathcal{F}$.

The probability measure P , which can be learned from large corpus, encodes semantics of every token in the language with intrinsic grammatical and logical structures. A token sequence generated from P may not be an understandable sentence for human being, because

it may not follow grammatical and logical structures with certain probability. However, computing based directly on the probability measure P is very costly and not practical. Therefore, we need to find a computation efficient representation of token-level semantics.

B. Token-level Semantic Vector Space

It took decades of effort to finally discover that the crucial step was to transition from token-level probabilistic models to semantic models based on vector representations. The shift is favored for its computational efficiency and its remarkable effectiveness in NLP tasks [27]. However, this conclusion is drawn mainly from extensive experiments and lacks a solid theoretical foundation. In this subsection, we will attempt to establish the mathematical foundations of semantic vector spaces.

Definition 11: The token-level semantic vector space of a language is a probabilistic inner product space $S = (\mathbb{S}^{N-1}, \mathcal{F}, \mu, \langle \cdot, \cdot \rangle)$, where $\mathbb{S}^{N-1}$ is a $(N - 1)$ -dimensional unit sphere, each $\mathbf{s} \in \mathbb{S}^{N-1}$ represents a semantic vector, $\mathcal{F}$ is a σ -algebra on $\mathbb{S}^{N-1}$, μ is a probability measure defined on $\mathcal{F}$, $\langle \cdot, \cdot \rangle$ is an inner product.

If we use $\mathbf{s}_1$ and $\mathbf{s}_2$ to denote two column vectors on $\mathbb{S}^{N-1}$, the inner product can be written as $\langle \mathbf{s}_1, \mathbf{s}_2 \rangle = \mathbf{s}_1^T \mathbf{s}_2$. The squared Euclidean distance is defined as $d_e^2(\mathbf{s}_1, \mathbf{s}_2) = \|\mathbf{s}_1 - \mathbf{s}_2\|^2 = (\mathbf{s}_1 - \mathbf{s}_2)^T (\mathbf{s}_1 - \mathbf{s}_2)$. The cosine similarity is defined as $\cos(\mathbf{s}_1, \mathbf{s}_2) = \mathbf{s}_1^T \mathbf{s}_2$. It is noticed that Ω in Definition 10 can only be mapped to N points in $\mathbb{S}^{N-1}$. Let the set of semantic vector of tokens in $\mathcal{A}$ be $\mathcal{S}(\mathcal{A}) \subset \mathbb{S}^{N-1}$ with $\forall \mathcal{A} \subseteq \Omega$. Thus, μ is an extension from P such that $\mu(\mathcal{S}(\mathcal{A})) = P(\mathcal{A})$ if $\mathcal{A} \in \mathcal{F}_S$, otherwise $\mu(\mathcal{S}(\mathcal{A})) = 0$.

Many works suggest that the semantic vector space should be a more complex low dimensional manifold. In practice, however, the Euclidean distance and cosine similarity remain the most widely used metrics, because of its simplicity in computation and adequate performance. Therefore, we argue that defining the semantic vector space directly on $\mathbb{S}^{N-1}$ strikes an effective trade-off between accuracy and computational efficiency.

The essential purpose of representing tokens as vectors is to use the cosine similarity between these high dimension vectors to represent semantic differences. The simple algebraic operations on vectors may not always work, because they do not necessarily reflect semantic relationships. For example, the conceptual illustration in the following may work for some tokens, but not apply to every token [23]:

$$\mathbf{s}(\text{King}) - \mathbf{s}(\text{Men}) + \mathbf{s}(\text{Woman}) \approx \mathbf{s}(\text{Queen}). \quad (33)$$

However, this example effectively demonstrates a projection do exist between the vector representations of “King” and “Men”. Consequently, scalars alone are insufficient to fully

characterize the semantic relations. Moreover, the cosine similarity is invariant to rotation and scaling, and much more robust than Euclidean distance in high dimension space. Thus, the cosine similarity and probability measure in S are of fundamental importance. Following the idea of Gromov-Wasserstein distance [46], [47], we define the distance of two semantic vector spaces as follows:

Definition 12: Let S and S' be two semantic vector spaces with probability measures μ and ν , respectively. The squared distance between S and S' is defined as:

$$d_s^2(S, S') = \min_{\pi \in \Pi(\mu, \nu)} \int_{S \times S'} \int_{S \times S'} |\mathbf{s}_1^T \mathbf{s}'_1 - \mathbf{s}_2^T \mathbf{s}'_2|^2 d\pi(\mathbf{s}_1, \mathbf{s}_2) d\pi(\mathbf{s}'_1, \mathbf{s}'_2), \quad (34)$$

where $\Pi(\mu, \nu)$ is the set of all transportation plans between μ and ν .

The definition seeks to find an optimal transport plan π that minimizes the weighted average of the “internal cosine similarity difference” for all pairs of points, measured before and after the transport. The distance difference imposes a high cost on pairings that distort the intrinsic geometry of two semantic vector spaces. Therefore, if $d_s(S, S') = 0$, S and S' are equivalent in the sense of token-level semantics, which results in an easy translation between these two languages. In fact, the Gromov-Wasserstein distance has already been successfully applied to the alignment of two word embeddings [81].

Remark 3 (Vectorization in Information Theory): The relationship between semantic space and semantic vector space is similar to the relationship between information theory and signal processing. Information theory, based on probability theory, is a framework for understanding the nature and limits of information compression, transmission, and storage. However, it is not particularly concerned with the specific methods of implementation in practice [16]. Signal processing, on the other hand, represents information as vectors in $\mathbb{R}^n$ or $\mathbb{C}^n$, making it suitable for sensing, transmission, and storage in physical media. This representation enables a vast body of mathematical theory to be applied to the design of efficient algorithms for practical sensing, communication, and storage systems [82].

C. Semantic Compression/De-dimensionality

In information theory, the objective of source coding is to use as few bits as possible to represent a source symbol, such that the source message can be exactly recovered for lossless compression or recovered within a given distortion for lossy compression [43]. According to Definition 11, however, $|\Omega| = N$ implies $\mathbb{S}^{N-1}$ is a very high dimension sphere such that the direct computation on S is still not practical. Extensive experimental results suggest that the choice of dimensionality for a semantic vector space involves a crucial trade-off, implying the

existence of an optimal range or “sweet spot” [83]. In this case, the semantic compression is the compression of the entire semantic space, i.e., dimension reduction that preserves cosine similarity.

In practice, the random projection is widely used to reduce the dimensionality of vectors. The distance conservation property is guaranteed by Johnson-Lindenstrauss (JL) lemma [84]. In the following, we introduce the cosine similarity based JL lemma without proof [85].

Lemma 1: Let $\epsilon \in (0, 1)$ and $\{\mathbf{s}_1, \dots, \mathbf{s}_M\} \in \mathbb{S}^{N-1}$, if $m \geq \frac{C}{\epsilon^2} \log M$, there exists a matrix $\mathbf{A} \in \mathbb{R}^{m \times N}$ such that:

$$|\mathbf{s}_i^T \mathbf{s}_j - \mathbf{s}_i^T \mathbf{P} \mathbf{s}_j| \leq \epsilon, \quad \forall i, j \in \{1, \dots, M\}, \quad (35)$$

where $\mathbf{P} = \mathbf{A}^T \mathbf{A}$.

According to JL lemma, the dimensionality of the semantic vector space can be reduced from N to $m \geq \frac{C}{\epsilon^2} \log M$. As aforementioned, each semantic vector can be seen as a real signal vector which should be very sparse in $\mathbb{S}^{N-1}$. Inspired by compressive sensing, the cosine similarity based JL lemma can be improved by applying restricted isometry property (RIP). Let $\mathbf{A}$ be a matrix satisfying (k, δ) -RIP, that is

$$1 - \delta \leq \|\mathbf{A}\mathbf{s}\|^2 \leq 1 + \delta, \quad (36)$$

for all k -sparse $\mathbf{s} \in \mathbb{S}^{N-1}$, i.e., $\|\mathbf{s}\|_0 \leq k$. The following result is established in [86].

Theorem 3: Let $\eta, \epsilon \in (0, 1)$, $\{\mathbf{s}_1, \dots, \mathbf{s}_M\} \in \mathbb{S}^{N-1}$, and $\mathbf{A} \in \mathbb{R}^{m \times N}$ be (k, δ) -RIP with $\delta \leq \epsilon/4$ and $k \geq 40 \log \frac{4M}{\eta}$. Let $\boldsymbol{\sigma}$ a Rademacher sequence, i.e., uniformly distributed on $\{-1, 1\}^N$. Then, with probability exceeding $1 - \eta$,

$$|\mathbf{s}_i^T \mathbf{s}_j - \mathbf{s}_i^T \mathbf{D}_{\boldsymbol{\sigma}} \mathbf{P} \mathbf{D}_{\boldsymbol{\sigma}} \mathbf{s}_j| \leq \epsilon, \quad \forall i, j \in \{1, \dots, M\}, \quad (37)$$

where $\mathbf{D}_{\boldsymbol{\sigma}}$ is a diagonal matrix whose diagonal entries are the elements of the vector $\boldsymbol{\sigma}$ and $\mathbf{P} = \mathbf{A}^T \mathbf{A}$.

According to the theory of compressive sensing, the $m \times N$ partial Gaussian matrix can be used with

$$m \geq \frac{C}{\epsilon^2} \log \frac{M}{\eta} \log N, \quad (38)$$

but the complexity of the matrix-vector multiplication is very high. However, $\mathbf{A}$ can also be obtained by randomly selecting m rows from the discrete Fourier transform (DFT) matrix, discrete cosine transform (DCT) matrix, or Hadamard matrix. In this case, m will be larger than using partial Gaussian matrix, but the complexity is greatly reduced.

Recalling Definition 12, the distortion of semantic compression can be evaluated by the distance of two semantic vector spaces. Let S be the original semantic vector space on $\mathbb{S}^{N-1}$ and S' on $\mathbb{S}^m$ with $1 \leq m < N - 1$, the distortion of semantic compression can be written as

$$d_s^2(S, S') = \min_{\pi \in \Pi(\mu, \mu')} \int_{S \times S'} \int_{S \times S'} |\mathbf{s}^T \mathbf{s}' - \mathbf{s}^T \mathbf{P} \mathbf{s}'|^2 d\pi(\mathbf{s}, \mathbf{A}\mathbf{s}) d\pi(\mathbf{s}', \mathbf{A}\mathbf{s}'), \quad (39)$$

where $\mathbf{A}$ is a $m \times N$ projection matrix and $\mathbf{P} = \mathbf{A}^T \mathbf{A}$. The following theorem can be established by applying Lemma 1 or Theorem 3 directly.

Theorem 4: The distortion of semantic compression can be bounded by ϵ , i.e., $d_s^2(S, S') \leq \epsilon$, with high probability.

The semantic compression/de-dimensionality discussed in this subsection does not consider the distribution on semantic vector space. Therefore, the bound in Theorem 4 is not tight, yet far from optimal in the sense of information theory. Similar to rate-distortion theory, the dimension-distortion theory can be further developed for semantic compression, especially for the case of m smaller than the threshold in Lemma 1 or Theorem 3.

Remark 4 (Approximate Nearest Neighbor Search): Vector databases are regarded as a critical piece of infrastructure for helping LLMs mitigate hallucinations. They can also store vast amounts of private and proprietary data, enhancing the capabilities of LLMs in vertical domains. Consequently, approximate nearest neighbor (ANN) vector search algorithms stand out as a key technology that integrates vector databases with LLMs. From the perspective of information theory, the nearest ANN vector searching is an extension to decoding algorithm, which is to search the nearest codeword for the received symbols. Since 2023, the ANN vector search algorithms proposed by the experts from our lab have been ranked TOP-1 on ANN-Benchmarks leader-board.² Interested researchers can access our code repository.³

D. Semantic Embedding/Vectorization for Next-token Prediction

In practice, we typically select a proper dimension m to directly perform the semantic embedding or vectorization. In the following, we will discuss information-theoretically optimal approach. It is natural to understand that the semantics of an utterance highly depend on the speaker's intended goal, i.e., the downstream task in machine learning. Therefore, for a token sequence with length n , the semantic embedding is a mapping $f : \Omega^n \rightarrow (\mathbb{S}^m)^n$, such that a loss functional $L(f)$, defined by the downstream task, is minimized.

²<https://ann-benchmarks.com>.

³https://github.com/WPJiang/HWTL_SDU-ANNS.

From the perspective of LLMs, the objective is to predict the next token based on the prompt and the parameterized memory. Therefore, $L(f)$ should be designed to best facilitate of achieving this goal. Let $X_{1:n}$ be a token sequence, $S_{1:n}$ be the corresponding semantic vector representation of $X_{1:n}$. For the task of the next token prediction, S_t should contain all the information in $X_{1:t}$ which is useful to predict $X_{t+1:n}$. From the perspective of information theory, the optimal semantic encoder for next token prediction should be the solution of the following problem:

$$\max_{S_t=f(X_{1:t})} I(X_{t+1:n}; S_t | S_{1:t-1}), \quad 1 \leq t \leq n \in \mathbb{N}. \quad (40)$$

The condition means S_t only contains new information for predicting $X_{t+1:n}$ which is not contained in $S_{1:t-1}$.

The solution of Eq. (40) maximizes the backward directed information $I(X_{n:1} \rightarrow S_{1:n})$ as follows:

$$I^*(X_{n:1} \rightarrow S_{1:n}) = \sum_{t=1}^n \max_{S_t=f(X_{1:t})} I(X_{t+1:n}; S_t | S_{1:t-1}). \quad (41)$$

Following the inequalities of directed information in [64], we have

$$I^*(X_{n:1} \rightarrow S_{1:n}) \leq \sum_{t=1}^n \max_{S_t=f(X_{1:t})} I(X_{t+1:n}; S_t) \leq \sum_{t=1}^n \sum_{k=1}^{n-t} \max_{S_t=f(X_{1:t})} I(X_{t+k}; S_t). \quad (42)$$

Inspired by the idea of predictive coding in information theory [87], [88], the CPC is proposed for semantic embedding in [48], which is also adopt in OpenAI [49]. Let $Z_{1:n}$ be the latent representation of $X_{1:n}$ with $Z_t = g_{\text{ENC}}(X_t)$, $S_{1:n}$ be the semantic vector obtained by CPC, which is defined as $S_t = g_{\text{AR}}(Z_{1:t-1})$. The training process of CPC is to solve the following optimization problem:

$$\sum_{k=1}^{n-t} \max_{S_t=f(X_{1:t})} I(X_{t+k}; S_t). \quad (43)$$

Therefore, the CPC maximizes the upper-bound of $I^*(X_{n:1} \rightarrow S_{1:n})$, which is a sub-optimal semantic encoder from the perspective of information theory. In this context, the information theoretical optimal semantic embedding can be achieved, if we can optimize the backward directed information Eq. (41) or its tighter upper bound.

V. AUTOREGRESSION LLMs

In this section, we focus on LLMs with a special architecture, i.e., AR-LLMs. The Transformer architecture and its performance can be derived from our general definition. Other LLM architectures, such as Mamba/Mamba2 and LLaDA, are also discussed.

A. TV-VAR based AR-LLMs

Let $\mathbf{s}_t$ with $t = 1, \dots, n$ and $\mathbf{u}_t$ with $t = n + 1, \dots, T$ be sample vectors of random variables S_t and U_t . To simplify the notation, we let $\mathbf{u}_t = \mathbf{s}_t$ for $t = 1, \dots, n$. We then have the following definition.

Definition 13: The TV-VAR based AR-LLM is defined as

$$\mathbf{u}_t = \arg \text{softmax} \left(\frac{1}{\Xi} \tilde{\mathbf{u}}_{1:N}^T \left(\sum_{j=1}^{t-1} \mathbf{A}_{tj} \mathbf{u}_j \right) \right), \quad t = n + 1, \dots, T, \quad (44)$$

where $\mathbf{A}_{tj}$ is the coefficient matrix, $\tilde{\mathbf{u}}_{1:N}$ are all possible token vectors in $\mathcal{S}(\Omega)$, and Ξ is the sampling temperature.

In contrast to the standard VAR model [50], $\mathbf{A}_{tj}$ is time-variant, which is very difficult to estimate in practice.

B. Transformer Architecture

Consider a decomposition of $\mathbf{A}_{tj}$ as follows:

$$\mathbf{A}_{tj} = \pi_{tj} \mathbf{A}, \quad (45)$$

where $\mathbf{A}$ is a time-invariant parameter matrix, and π_{tj} is the only time-variant scalar weight satisfying $\sum_{j=1}^{t-1} \pi_{tj} = 1$ and $\pi_{tj} \geq 0$. Simple derivation yields the following theorem.

Theorem 5: The Transformer is an AR-LLM with the following form

$$\mathbf{u}_t = \arg \text{softmax} \left(\frac{1}{\Xi} \tilde{\mathbf{u}}_{1:N}^T \left(\sum_{j=1}^{t-1} \pi_{tj} \mathbf{A} \mathbf{u}_j \right) \right), \quad t = n + 1, \dots, T, \quad (46)$$

where π_{tj} is the output of the softmax, that is

$$\pi_{tj} = \frac{\exp(\mathbf{u}_{t-1}^T \mathbf{B} \mathbf{u}_j)}{\sum_{i=1}^{t-1} \exp(\mathbf{u}_{t-1}^T \mathbf{B} \mathbf{u}_i)}, \quad j = 1, \dots, t-1. \quad (47)$$

Proof: Let $\mathbf{q}_t$, $\mathbf{k}_t$, and $\mathbf{v}_t$ be sample vectors of random variables Q_t , K_t , and V_t . The attention scheme in [28] implies

$$\begin{cases} \mathbf{q}_t = \mathbf{W}_q \mathbf{u}_t, \\ \mathbf{k}_t = \mathbf{W}_k \mathbf{u}_t, \\ \mathbf{v}_t = \mathbf{W}_v \mathbf{u}_t, \end{cases} \quad (48)$$

for $t = 1, \dots, T$. The output of the Transformer is

$$\mathbf{u}_t = \arg \text{softmax} \left(\frac{1}{\Xi} \tilde{\mathbf{u}}_{1:N}^T \left(\sum_{j=1}^{t-1} \pi_{tj} \mathbf{v}_j \right) \right), \quad t = n + 1, \dots, T, \quad (49)$$

where

$$\pi_{tj} = \frac{\exp(\mathbf{q}_{t-1}^T \mathbf{k}_j)}{\sum_{i=1}^{t-1} \exp(\mathbf{q}_{t-1}^T \mathbf{k}_i)}, \quad j = 1, \dots, t-1 \quad (50)$$

is the attention score. This theorem is established by letting $\mathbf{A} = \mathbf{W}_v$ and

$$\mathbf{B} = \mathbf{W}_q^T \mathbf{W}_k. \quad (51)$$

■

This theorem shows that the Transformer is equivalent to a decomposition of $\mathbf{A}_{tj}$ as follows:

$$\mathbf{A}_{tj} = \pi_{tj} \mathbf{A}, \quad (52)$$

where π_{tj} measures the semantic relevance from $\mathbf{u}_j$ with $j = 1, \dots, t-1$ for predicting $\mathbf{u}_t$. In an utterance, the semantic relevance is asymmetric between different tokens. Recalling Section IV, the inner product is used to measure the correlations of token-level semantic. For the asymmetric semantic relevance in an utterance, the inner-product based bilinear form for predicting $\mathbf{u}_t$ is introduced as follows:

$$B(\mathbf{u}_{t-1}, \mathbf{u}_j) = \mathbf{u}_{t-1}^T \mathbf{B} \mathbf{u}_j, \quad j = 1, \dots, t-1, \text{ and } t = n+1, \dots, T, \quad (53)$$

where $\mathbf{B} \neq \mathbf{B}^T$ in general. π_{tj} can then be assigned by using softmax as Eq. (47). According to Jaynes' maximum entropy principle [89], the softmax is a probability assignment on discrete sample space that maximize the entropy with the constraint on the first order moment. Therefore, the obtained estimation of the semantic relevance is the one with the maximum uncertainty, i.e., the best achievable estimation in the worst case.

C. ELBO of the Transformer

The performance of AR-LLM can be analyzed from variational inference perspective. Similar to [90], J is introduced as a latent variable defined on $\{1, \dots, T\}$. π_{tj} can then be seen as the probability that choosing the position $J = j$. Thus, the prediction of U_t in Eq. (46) is the expectation over J as follows:

$$\mathbf{u}_t = \arg \text{softmax} \left(\frac{1}{\Xi} \tilde{\mathbf{u}}_{1:N}^T \mathbb{E}_{J \sim Q(\cdot | U_{n+1:t-1}, S_{1:n}; \{\mathbf{A}, \mathbf{B}\})} \{\mathbf{A} \mathbf{u}_J\} \right), \quad t = n+1, \dots, T, \quad (54)$$

where

$$Q(j | U_{n+1:t-1}, S_{1:n}; \{\mathbf{A}, \mathbf{B}\}) = \pi_{tj}, \quad j = 1, \dots, t-1. \quad (55)$$

By applying the principle of variational inference [51], we then have the following theorems.

Theorem 6: The pre-training phase of Transformer is equivalent to

$$\max_{\mathbf{A}, \mathbf{B}} \text{ELBO}(Q(J | U_{n+1:t-1}^h, S_{1:n}; \{\mathbf{A}, \mathbf{B}\})), \quad t = n+1, \dots, T. \quad (56)$$

Proof: In the pre-training phase, we will maximize the following cross-entropy loss:

$$\max_{\Phi} H(P_t^h, Q_t^\Phi) = \min_{\Phi} \mathbb{E}_{P_t^h} \{\log Q_t^\Phi\}, \quad t = n+1, \dots, T. \quad (57)$$

In the optimum, we have

$$Q_t^{\Phi^h} = P(U_t | U_{n+1:t-1}, S_{1:n}; \Phi^h) = P(U_t^h | U_{n+1:t-1}^h, S_{1:n}) = P_t^h. \quad (58)$$

Therefore, the pre-training phase is equivalent to solve the following optimization problem:

$$\max_{\Phi} \log P(U_t^h | U_{n+1:t-1}^h, S_{1:n}; \Phi). \quad (59)$$

According to the principle of variational inference, we have

$$\begin{aligned} & \log P(U_t^h | U_{n+1:t-1}^h, S_{1:n}; \Phi) \\ &= \log \sum_{j=1}^{t-1} P(U_t^h, j | U_{n+1:t-1}^h, S_{1:n}; \Phi) \\ &= \log \sum_{j=1}^{t-1} P(U_t^h, j | U_{n+1:t-1}^h, S_{1:n}; \Phi) \frac{Q(j | U_{n+1:t-1}^h, S_{1:n}; \{\mathbf{A}, \mathbf{B}\})}{Q(j | U_{n+1:t-1}^h, S_{1:n}; \{\mathbf{A}, \mathbf{B}\})} \\ &= \log \mathbb{E}_{J \sim Q(\cdot | U_{n+1:t-1}^h, S_{1:n}; \{\mathbf{A}, \mathbf{B}\})} \left\{ \frac{P(U_t^h, J | U_{n+1:t-1}^h, S_{1:n})}{Q(J | U_{n+1:t-1}^h, S_{1:n}; \{\mathbf{A}, \mathbf{B}\})} \right\} \\ &\geq \mathbb{E}_{J \sim Q(\cdot | U_{n+1:t-1}^h, S_{1:n}; \{\mathbf{A}, \mathbf{B}\})} \left\{ \log \frac{P(U_t^h, J | U_{n+1:t-1}^h, S_{1:n})}{Q(J | U_{n+1:t-1}^h, S_{1:n}; \{\mathbf{A}, \mathbf{B}\})} \right\}. \end{aligned} \quad (60)$$

The last term is exactly the ELBO, which can be rewritten as

$$\begin{aligned} & \text{ELBO}(Q(J | U_{n+1:t-1}^h, S_{1:n}; \{\mathbf{A}, \mathbf{B}\})) \\ &= \mathbb{E}_{J \sim Q(\cdot | U_{n+1:t-1}^h, S_{1:n}; \{\mathbf{A}, \mathbf{B}\})} \{\log P(U_t^h, J | U_{n+1:T}^h, S_{1:n})\} \\ & \quad - D_{KL}(Q(J | U_{n+1:t-1}^h, S_{1:n}; \{\mathbf{A}, \mathbf{B}\}) \| P(J | U_{n+1:t-1}^h, S_{1:n})). \end{aligned} \quad (61)$$

As a result, the training phase is equivalent to

$$\max_{\mathbf{A}, \mathbf{B}} \text{ELBO}(Q(J | U_{n+1:t-1}^h, S_{1:n}; \{\mathbf{A}, \mathbf{B}\})), \quad t = n+1, \dots, T. \quad (62)$$

■

Theorem 7: The inference phase of Transformer is equivalent to

$$\max_{U_t \in \mathcal{S}(\Omega)} \text{ELBO}(Q_t(J | U_{n+1:t-1}, S_{1:n}; \{\mathbf{A}^{h+}, \mathbf{B}^{h+}\})), \quad t = n+1, \dots, T, \quad (63)$$

where $\mathbf{A}^{h+}$ and $\mathbf{B}^{h+}$ are the parameter matrices after training.

Proof: In the inference phase, U_t is chosen from $\mathcal{S}(\Omega)$ such that

$$\log P(U_t | U_{n+1:t-1}, S_{1:n}; \Phi^{h+}) \quad (64)$$

is maximized. According to the principle of variational inference, we have

$$\begin{aligned}
& \log P(U_t | U_{n+1:t-1}, S_{1:n}; \Phi^{h+}) \\
&= \log \sum_{j=1}^{t-1} P(U_t, j | U_{n+1:t-1}, S_{1:n}; \Phi^{h+}) \\
&= \log \sum_{j=1}^{t-1} P(U_t, j | U_{n+1:t-1}, S_{1:n}; \Phi^{h+}) \frac{Q(j | U_{n+1:t-1}, S_{1:n}; \{\mathbf{A}^{h+}, \mathbf{B}^{h+}\})}{Q(j | U_{n+1:t-1}, S_{1:n}; \{\mathbf{A}^{h+}, \mathbf{B}^{h+}\})} \\
&= \log \mathbb{E}_{J \sim Q(\cdot | U_{n+1:t-1}, S_{1:n}; \{\mathbf{A}^{h+}, \mathbf{B}^{h+}\})} \left\{ \frac{P(U_t, J | U_{n+1:t-1}, S_{1:n}; \Phi^{h+})}{Q(J | U_{n+1:t-1}, S_{1:n}; \{\mathbf{A}^{h+}, \mathbf{B}^{h+}\})} \right\} \\
&\geq \mathbb{E}_{J \sim Q(\cdot | U_{n+1:t-1}, S_{1:n}; \{\mathbf{A}^{h+}, \mathbf{B}^{h+}\})} \left\{ \log \frac{P(U_t, J | U_{n+1:t-1}, S_{1:n}; \Phi^{h+})}{Q(J | U_{n+1:t-1}, S_{1:n}; \{\mathbf{A}^{h+}, \mathbf{B}^{h+}\})} \right\}.
\end{aligned} \tag{65}$$

The last term is exactly the ELBO, which can be rewritten as

$$\begin{aligned}
& \text{ELBO}(Q(J | U_{n+1:t-1}, S_{1:n}; \{\mathbf{A}^{h+}, \mathbf{B}^{h+}\})) \\
&= \mathbb{E}_{J \sim Q(\cdot | U_{n+1:t-1}, S_{1:n}; \{\mathbf{A}^{h+}, \mathbf{B}^{h+}\})} \{ \log P(U_t | J, U_{n+1:t-1}, S_{1:n}; \Phi^{h+}) \} \\
&\quad - D_{KL}(Q(J | U_{n+1:t-1}, S_{1:n}; \{\mathbf{A}^{h+}, \mathbf{B}^{h+}\}) \| P(J | U_{n+1:t-1}, S_{1:n}; \Phi^{h+})).
\end{aligned} \tag{66}$$

As a result, the inference phase is equivalent to

$$\max_{U_t \in \mathcal{S}(\Omega)} \text{ELBO}(Q(J | U_{n+1:t-1}, S_{1:n}; \{\mathbf{A}^{h+}, \mathbf{B}^{h+}\})), \quad t = n+1, \dots, T. \tag{67}$$

■

D. Generalization Error Bound of the Transformer

Rademacher complexity and Talagrand's concentration inequalities are fundamental tools in statistical learning theory for analyzing the generalization error bounds of machine learning algorithms [52]. This section applies these tools to study the generalization error bound of the Transformer.

Let $\mathbf{u}_t^h$ be the ground-truth output vector at time t for $t = n+1, \dots, T$, where the corresponding random variable is U_t^h . Therefore, the generalization error is given By

$$H(P(U_t^h), Q(U_t)), \tag{68}$$

where $P(U_t^h)$ is the one-shot coding, $Q(U_t)$ is the output of the softmax function. Given t , we take M samples from the Transformer output U_t , each of which is denoted as $\mathbf{u}_{mt}$ for $m = 1, \dots, M$. Recalling Theorem 5, the i -th entry of the logits $\mathbf{z}_m$ is defined by

$$z_m^i = \frac{1}{\Xi} \tilde{\mathbf{u}}_i^T \left(\sum_{j=1}^{t-1} \pi_{tj} \mathbf{A} \mathbf{u}_{mj} \right), \quad i = 1, \dots, N. \tag{69}$$

The empirical generalization error over a sample set with size M is given by

$$\hat{\mathcal{L}}(\mathbf{A}, \mathbf{B}) = \frac{1}{M} \sum_{m=1}^M \mathbf{1}^T(\mathbf{u}_{mt}^h) \log \frac{1}{\mathbf{q}(\mathbf{z}_m)} = \frac{1}{M} \sum_{m=1}^M \log \frac{1}{q(z_m^h)}, \quad (70)$$

where $\mathbf{q}(\mathbf{z}_m)$ is the output of the softmax function, and

$$q(z_m^h) = \mathbf{1}^T(\mathbf{u}_{mt}^h) \mathbf{q}(\mathbf{z}_m). \quad (71)$$

Theorem 8: For any $\delta > 0$, the generalization error of the Transformer is upper bounded by

$$H(P(U_t^h), Q(U_t)) \leq \hat{\mathcal{L}}(\mathbf{A}, \mathbf{B}) + \frac{2\sqrt{2}}{M} \sum_{m=1}^M |z_m^h| + 3\sqrt{\frac{\log \frac{2}{\delta}}{2M}}, \quad t = n+1, \dots, T. \quad (72)$$

with probability at least $1 - \delta$ over the choice of M samples.

Proof: The empirical Rademacher complexity of the Transformer is given by

$$\hat{\mathcal{R}}(\mathbf{A}, \mathbf{B}) = \mathbb{E}_{\sigma} \left\{ \sup_{\mathbf{A}, \mathbf{B}} \frac{1}{M} \sum_{m=1}^M \sigma_m \log \frac{1}{q(z_m^h)} \right\}, \quad (73)$$

where σ is a Rademacher sequence. According to Theorem 3.3 in [52], we have

$$H(P(U_t^h), Q(U_t)) \leq \hat{\mathcal{L}}(\mathbf{A}, \mathbf{B}) + 2\hat{\mathcal{R}}(\mathbf{A}, \mathbf{B}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2M}}. \quad (74)$$

Because $\mathbf{q}(\mathbf{z}_m)$ is the output of the softmax function, $\hat{\mathcal{L}}(\mathbf{A}, \mathbf{B})$ is $\sqrt{2}$ -Lipschitz over z_m^h for l^2 -norm. According to Talagrand's Lemma in [52], we have

$$\hat{\mathcal{R}}(\mathbf{A}, \mathbf{B}) \leq \mathbb{E}_{\sigma} \left\{ \sup_{\mathbf{A}, \mathbf{B}} \frac{1}{M} \sum_{m=1}^M \sigma_m z_m^h \right\} \leq \frac{\sqrt{2}}{M} \sum_{m=1}^M |z_m^h|. \quad (75)$$

This theorem has been established. ■

This result shows that the logits determines the accuracy during the inference phase. Therefore, when using quantization for inference acceleration, it is crucial to ensure that the quantization algorithm has a minimal impact on the logits.

E. Memory Capacity of the Transformer

The statistical physics approaches, such as spin glass model and replica method, have been widely used to analyze the performance of signal processing, coding, and satisfiability (SAT) problems [91]. In a series of landmark papers [53]–[55], Gardner investigated the memory capacity of the classical Hopfield network [92] by applying the replica method, which is referred to as Gardner capacity afterwards.

Definition 14: Let N_P be the maximum number of random patterns which can be memorized in a classical Hopfield network with n neurons. The generalized Gardner capacity is defined as

$$C_G = \frac{\alpha(N_P)}{n}, \quad (76)$$

where $\alpha(\cdot)$ is chosen to scale with n . It is an identity function in the original definition.

As a matter of fact, generalized Gardner capacity has a deep connection with Shannon capacity. If the pattern here is not a binary n -sequence but a binary n -sphere, the Gardner capacity is equivalent to Shannon capacity, where $\alpha(\cdot)$ is chosen as a logarithm function. The transformation from n -sequence to n -sphere is critical, which explains the error correction capability of modern neural networks.

Recent work in [93] focused on the modern continuous Hopfield network, which is shown to be equivalent to the attention scheme. It is also proved that the memory capacity is exponential in the dimension of the space of the query and key-value patterns. Therefore, it is not surprising that a large amount of patterns can be memorized by a small LLM. Following this idea, we model the behavior of Transformers with associative memories using modern continuous Hopfield networks, which is used to explain the scaling law from theoretic perspective [94].

F. Semantic Information Theoretic Measure for the Transformer

In Section III, we introduce semantic information theoretic measures for LLMs, such as the directed rate-distortion function in the pre-training phase, the directed rate-reward function in the post-training phase, and the semantic information flow in inference phase, where the key is to estimate the directed information.

The directed information $I(S_{1:n} \rightarrow U_{n+1:t}; \Phi)$ can be represented by KL divergence as follows

$$I(S_{1:n} \rightarrow U_{n+1:t}; \Phi) = D_{KL} \left(P(S_{1:n}, U_{n+1:t}) \| P(S_{1:n}) \prod_{j=n+1}^t P(U_j | U_{n+1:j}) \right). \quad (77)$$

Therefore, the Donsker-Varadhan representation can be used for directed information estimation [95]. This idea is proposed and thoroughly analyzed in [96] for transfer entropy estimation, where the transformer itself is used as the estimator.

G. Other Architectures

To simplify the computation complexity in both training and inference phases. Various LLM architectures, such as Mamba/Mamba2 [56], [57] and LLaDA [58], have been proposed. We will discuss the relation between these new architectures and Definition 13.

1) *Mamba/Mamba2*: To save the computation of softmax in attention scheme, Mamba/Mamba2 architectures are proposed and thoroughly analyzed in [56], [57]. Inspired by control theory, the discrete state space model (SSM) used in Mamba/Mamba2 is

$$\begin{cases} \mathbf{u}_t = \mathbf{A}_t \mathbf{u}_{t-1} + \mathbf{B}_t \mathbf{s}_t; \\ \mathbf{y}_t = \mathbf{C} \mathbf{u}_t. \end{cases} \quad (78)$$

Clearly, the SSM is a special case of the AR-LLM in Definition 13, which exactly belongs to linear TV-VAR models [50]. The linear TV-VAR model is widely used in time series analysis for economics and finance [97], [98]. Therefore, the developed parameter estimation method may be applicable to improve the performance of Mamba/Mamba2. Because there lacks the bilinear model of semantic relevance, it is not difficult to understand that the performance of Mamba/Mamba2 could be worse than Transformer. However, the Mamba/Mamba2 architectures inspire us to consider other forms of AR-LLM which may have a similar performance as Transformer but much lower computation complexity. Based on the improved Mamba2 [99], Qwen3-Next is the first LLM which implements the hybrid attention scheme.⁴ The Transformer, however, is different from linear TV-VAR model because π_{tj} introduces a non-linear relation, i.e., the softmax function over a bilinear form of $\mathbf{u}_t$ and $\mathbf{u}_j$.

2) *LLaDA*: As a diffusion LLM, LLaDA constitutes a groundbreaking attempt to transcend the Transformer paradigm [58]. In LLaDA, it assumes many tokens in an utterance are masked, which will be predicted based on the unmasked ones. The loss function for training LLaDA is a cross-entropy computed only on the masked tokens:

$$\mathcal{L}(\Phi) = -\mathbb{E}_{\tau, U_{1:T}^\tau, U_{1:T}^0} \left\{ \frac{1}{\tau} \sum_{t=1}^T \mathbf{1}(U_t^\tau = M) \log P(U_t^0 | U_{1:T}^\tau; \Phi) \right\}, \quad (79)$$

where M denote the masked token. The transformer without causal mask is used as the core component to predict the masked tokens. Evidently, while LLaDA is fundamentally built upon a diffusion framework, the AR-LLM remains central to the task of masked token prediction in LLaDA.

⁴<https://qwen.ai/blog?from=research.latest-advancements-list&id=4074cca80393150c248e508aa62983f9cb7d27cd&>

VI. CONCLUSIONS

Drawing from the theory of rate-distortion function, directed information, and Granger causality, this paper aims to uncover the semantic information-theoretic principles underlying LLMs. We discussed the structure-agnostic information-theoretic measures, the token-level semantic embedding, and the general definition of AR-LLM, from which the Transformer architecture and its performance have been derived theoretically. Our theory indicates that the capabilities of current LLMs remain within the scope of Granger causality. How to achieve the counterfactual reasoning and system 2 reasoning abilities [100], [101], remains a formidable challenge. Consequently, our semantic information theory framework provides a lens through which many experimental observations can be explained, which also paves the way for unlocking the full potential of LLMs.

ACKNOWLEDGMENT

I am grateful to T. Wu, X. Niu, K. Zhang, C. Zhang, Y. Lan, Z. Zhong, B. Chen, and Q. Zhang for productive discussions.

REFERENCES

- [1] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 7, pp. 379-423, Oct. 1948.
- [2] W. Weaver, "Recent contributions to the mathematical theory of communications," *The Rockefeller Foundation*, Sep. 1949.
- [3] R. Carnap, "Empiricism, semantics, and ontology," *Revue Internationale de Philosophie*, no. 4, pp. 20-40, Apr. 1950.
- [4] R. Carnap and Y. Bar-Hillel, "An outline of a theory of semantic information," Massachusetts Institute of Technology, Cambridge, MA, USA, Research Laboratory of Electronics Technical Report No. 247, Oct. 1952.
- [5] Y. Bar-Hillel and R. Carnap, "Semantic information," *The British Journal for the Philosophy of Science*, vol. 4, no. 14, pp. 147-157, Aug. 1953.
- [6] R. Carnap, *Meaning and Necessity: A Study in Semantics and Modal Logic*, 2nd ed. Chicago, IL, USA: University of Chicago Press, 1988.
- [7] M. Burgin, *Theory of Information: Fundamentality, Diversity and Unification*. Singapore: World Scientific Publishing, 2009.
- [8] L. Floridi, Ed., *The Routledge Handbook of Philosophy of Information*. London, UK: Routledge, 2016.
- [9] R. Solomonoff, "A formal theory of inductive inference - Part 1," *Information and Control*, vol. 7, no. 1, pp. 1-22, Mar. 1964.
- [10] R. Solomonoff, "A formal theory of inductive inference - Part 2," *Information and Control*, vol. 7, no. 2, pp. 224-254, Jun. 1964.
- [11] R. Solomonoff, "The discovery of algorithmic probability," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 73-88, Aug. 1997.
- [12] A. Kolmogorov, "Three approaches to the quantitative definition of information," *International Journal of Computer Mathematics*, vol. 2, no. 1-4, pp. 157-168, Jan. 1968.

- [13] A. Kolmogorov, “Logical basis for information theory and probability theory,” *IEEE Trans. Inf. Theory*, vol. 14, no. 5, pp. 662-664, Sep. 1968.
- [14] M. Hutter, *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Berlin, Germany: Springer, 2004.
- [15] A. Shen, V. Uspensky, and N. Vereshchagin, *Kolmogorov Complexity and Algorithmic Randomness*. Providence, RI, USA: American Mathematical Society, 2022.
- [16] T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: John Wiley & Sons, 2006.
- [17] B. Poole, S. Ozair, A. Oord, A. Alemi, and G. Tucker, “On variational bounds of mutual information,” in *Proc. 36th ICML '19*, Long Beach, CA, USA: ICML, Jun. 2019.
- [18] R. Sutton, “The bitter lesson,” University of Alberta, Edmonton, Canada, Mar. 2019.
- [19] H. Luhn, “A new method of recording and searching information,” *American Documentation*, vol. 4, no. 1, pp. 14-16, Jan. 1953.
- [20] G. Salton, A. Wong, and C. Yang, “A vector space model for automatic indexing,” *Commun. ACM*, vol. 18, no. 11, pp. 613-620, Nov. 1975.
- [21] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, “A neural probabilistic language model,” *J. Machine Learn. Res.*, vol. 3, pp. 1137-1155, 2003.
- [22] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv: 1301.3781*, Sep. 2013.
- [23] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Proc. 27th NIPS '13*, Lake Tahoe, NV, USA, Dec. 2013.
- [24] J. Pennington, R. Socher, and C. Manning, “GloVe: Global vectors for word representation,” in *Proc. ACL EMNLP '14*, Doha, Qatar, Oct. 2014.
- [25] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135-146, 2017.
- [26] M. Peters et al., “Deep contextualized word representations,” in *Proc. ACL NAACL-HLT '18*, New Orleans, LA, USA, Jun. 2018.
- [27] D. Jurafsky and J. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd ed. Draft, 2025.
- [28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proc. 31st NIPS '17*, Long Beach, CA, USA, 4-9 Dec. 2017.
- [29] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” *OpenAI*, Jun. 2018.
- [30] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” *OpenAI*, Feb. 2019.
- [31] T. Brown et al., “Language models are few-shot learners,” in *Proc. 34th NeurIPS '20*, Virtual Conference, 6-12 Dec. 2020.
- [32] L. Ouyang et al., “Training language models to follow instructions with human feedback,” *arXiv: 2203.02155*, Mar. 2022.
- [33] D. Guo et al., “DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning,” *Nature*, vol. 645, no. 8081, pp. 633-638, Sep. 2025.
- [34] “DeepSeek-V3.2-Exp: Boosting long-context efficiency with DeepSeek sparse attention,” *DeepSeek*, Hangzhou, China, Sep. 2025.
- [35] Y. Polyanskiy and Y. Wu, *Information Theory: From Coding to Learning*. Cambridge, UK: Cambridge University Press, 2025.

- [36] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *arXiv: 1703.00810*, Apr. 2017.
- [37] C. Shani, D. Jurafsky, Y. LeCun, and R. Shwartz-Ziv, "From tokens to thoughts: How LLMs and humans trade compression for meaning," *arXiv: 2505.17117*, Jun. 2025.
- [38] T. Weissman, "Toward textual transform coding," *IEEE BITS Inform. Theory Mag.*, vol. 3, no. 2, pp. 32-40, Jun. 2023.
- [39] X. Niu, B. Bai, N. Guo, W. Zhang, and W. Han, "Rate-distortion-perception trade-off in information theory, generative models, and intelligent communications," *Entropy*, vol. 27, no. 4, Apr. 2025.
- [40] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet, "A mathematical perspective on transformers," *arXiv: 2312.10794*, Aug. 2025.
- [41] M. Rodrigues and Y. Eldar, *Information-Theoretic Methods in Data Science*. Cambridge, UK: Cambridge University Press, 2021.
- [42] J. Massey, "Causality, feedback and directed information," in *Proc. IEEE ISIT '90*, Waikiki, HI, USA, Nov. 1990.
- [43] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ, USA: Prentice Hall PTR, 1971.
- [44] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: The MIT Press, 2018.
- [45] C. Granger, "Testing for causality: A personal viewpoint," *Journal of Economic Dynamics and Control*, vol. 2, no. 1, pp. 329-352, Jan. 1980.
- [46] M. Gromov, *Metric Structures for Riemannian and Non-Riemannian Spaces*. Boston, MA, USA: Birkhäuser, 2007.
- [47] C. Villani, *Optimal Transport: Old and New*. New York, NY, USA: Springer, 2009.
- [48] A. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv: 1807.03748*, Jan. 2019.
- [49] A. Neelakantan et al., "Text and code embeddings by contrastive pre-training," *arXiv: 2201.10005*, Jan. 2022.
- [50] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*. Berlin, Germany: Springer, 2007.
- [51] M. Wainwright and M. Jordan, "Graphical models, exponential families, and variational inference," *Foundation and Trends in Machine Learning*, vol. 1, no. 1-2, pp. 1-305, Nov. 2008.
- [52] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, 2nd ed. Cambridge, MA, USA: The MIT Press, 2018.
- [53] E. Gardner, "The space of interactions in neural network models," *J. Phys. A: Math. Gen.*, vol. 21, no. 1, pp. 257-270, Jan. 1988.
- [54] E. Gardner and B. Derrida, "Optimal storage properties of neural network models," *J. Phys. A: Math. Gen.*, vol. 21, no. 1, pp. 271-284, Jan. 1988.
- [55] E. Gardner and B. Derrida, "Three unfinished works on the optimal storage capacity of networks," *J. Phys. A: Math. Gen.*, vol. 22, no. 12, pp. 1983-1994, Jun. 1989.
- [56] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv: 2312.00752*, May 2024.
- [57] T. Dao and A. Gu, "Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality," *arXiv: 2405.21060*, May 2024.
- [58] S. Nie et al., "Large language diffusion models," *arXiv: 2502.09992*, Feb. 2025.
- [59] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. 18, no. 4, pp. 460-473, Jul. 1972.
- [60] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 18, no. 1, pp. 14-20, Jan. 1972.

- [61] S. Wu, W. Ye, H. Wu, H. Wu, W. Zhang, and B. Bai, “A communication optimal transport approach to the computation of rate distortion functions,” *arXiv: 2212.10098*, Dec. 2022.
- [62] L. Chen et al., “A constrained BA algorithm for rate-distortion and distortion-rate functions,” *arXiv: 2305.02650*, Jan. 2024.
- [63] C. Chen et al., “Computation of rate-distortion-perception functions with Wasserstein barycenter,” in *Proc. IEEE ISIT '23*, Taipei, Taiwan, Jun. 2023.
- [64] G. Kramer, “Directed information for channels with feedback,” Ph. D Dissertation, ETH Zurich, Zurich, Switzerland, 1998.
- [65] R. Dobrushin, “General formulation of Shannon’s main theorem in information theory,” *American Mathematical Society Translations: Series 2*, vol. 33, no. 2, pp. 323-438, 1963.
- [66] I. Naiss and H. Permuter, “Extension of the Blahut-Arimoto algorithm for maximizing directed information,” *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 204-222, Jan. 2013.
- [67] M. Belghazi et al., “MINE: Mutual information neural estimation,” *arXiv: 1801.04062*, Aug. 2021.
- [68] D. Tsur, Z. Aharoni, Z. Goldfeld, and H. Permuter, “Neural estimation and optimization of directed information over continuous spaces,” *IEEE Trans. on Inf. Theory*, vol. 69, no. 8, pp. 4777-4798, Aug. 2023.
- [69] V. Strassen, “Asymptotische abschätzungen in Shannon’s informationstheorie,” in *Proc. Trans. 3rd Prague Conf. Inf. Theory '62*, Prague, Czech Republic, 1962.
- [70] P. Amblard and O. Michel, “The relation between Granger causality and directed information theory: A review,” *Entropy*, vol. 15, no. 1, pp. 113-143, Jan. 2013.
- [71] T. Schreiber, “Measuring information transfer,” *Phys. Rev. Lett.*, vol. 85, no. 2, pp. 461-464, Jul. 2000.
- [72] L. Barnett, A. Barrett, and A. Seth, “Granger causality and transfer entropy are equivalent for Gaussian variables,” *Phys. Rev. Lett.*, vol. 103, no. 23, p. 238701, Dec. 2009.
- [73] D. Genççağa, Ed., “Transfer entropy,” *Entropy*, vol. 20, no. 4, p. 288, Apr. 2018.
- [74] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. New York, NY, USA: Cambridge University Press, 2009.
- [75] P. Grünwald and P. Vitányi, “Shannon information and Kolmogorov complexity,” *arXiv: cs/0410002*, Jul. 2010.
- [76] S. Amari, *Information Geometry and Its Applications*, Tokyo, Japan: Springer, 2016.
- [77] J. Martens and R. Grosse, “Optimizing neural networks with Kronecker-factored approximate curvature,” in *Proc. 32nd ICML '15*, Lille, France: ICML, Jul. 2015.
- [78] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. Manning, and C. Finn, “Direct preference optimization: Your language model is secretly a reward model,” *arXiv: 2305.18290*, Jul. 2024.
- [79] D. Freedman, “On tail probabilities for martingales,” *The Annals of Probability*, vol. 3, no. 1, pp. 100-118, Feb. 1975.
- [80] D. Williams, *Probability with Martingales*. Cambridge, UK: Cambridge University Press, 1991.
- [81] D. Alvarez-Melis and T. Jaakkola, “Gromov-Wasserstein alignment of word embedding spaces,” in *Proc. ACM EMNLP '18*, Brussels, Belgium, Oct. 2018.
- [82] A. Lapidoth, *A Foundation in Digital Communication*. New York, NY, USA: Cambridge University Press, 2009.
- [83] T. Landauer, P. Foltz, and D. Laham, “An introduction to latent semantic analysis,” *Discourse Processes*, vol. 25, no. 2-3, pp. 259-284, Jan. 1998.
- [84] W. Johnson, J. Lindenstrauss, and G. Schechtman, “Extensions of Lipschitz maps into Banach spaces,” *Israel J. Math.*, vol. 54, no. 2, pp. 129-138, Jun. 1986.
- [85] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. New York, NY, USA: Birkhäuser, 2013.
- [86] F. Krahmer and R. Ward, “New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property,” *SIAM J. Math. Anal.*, vol. 43, no. 3, pp. 1269-1281, Jan. 2011.

- [87] P. Elias, "Predictive coding - Part 1," *IRE Trans. Inf. Theory*, vol. 1, no. 1, pp. 16-24, Mar. 1955.
- [88] P. Elias, "Predictive coding - Part 2," *IRE Trans. Inf. Theory*, vol. 1, no. 1, pp. 24-33, Mar. 1955.
- [89] E. Jaynes, *Probability Theory: The Logic of Science*. New York, NY, USA: Cambridge University Press, 2003.
- [90] Y. Kim, C. Denton, L. Hoang, and A. Rush, "Structured attention networks," *arXiv: 1702.00887*, Feb. 2017.
- [91] N. Macris and R. Urbanke, *Statistical Physics for Communications, Signal Processing, and Computer Science*. Lausanne, Swiss: École Polytechnique Fédérale de Lausanne, 2017.
- [92] J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences*, vol. 79, no. 8, pp. 2554-2558, Apr. 1982.
- [93] H. Ramsauer et al., "Hopfield networks is all you need," *arXiv: 2008.02217*, Apr. 2021.
- [94] X. Niu, B. Bai, L. Deng, and W. Han, "Beyond scaling laws: Understanding transformer performance with associative memory," *arXiv: 2405.08707*, 14 May 2024.
- [95] M. Donsker and S. Varadhan, "Asymptotic evaluation of certain markov process expectations for large time, IV," *Comm. Pure Appl. Math.*, vol. 36, no. 2, pp. 183-212, Mar. 1983.
- [96] O. Luxembourg, D. Tsur, and H. Permuter, "TREET: Transfer entropy estimation via transformers," *arXiv:2402.06919*, Jul. 2025.
- [97] T. Lubik and C. Matthes, "Time-varying parameter vector autoregressions: Specification, estimation, and an application," *Economic Quarterly*, vol. 101, no. 4, pp. 323-352, Q4 2015.
- [98] J. Haslbeck, L. Bringmann, and L. Waldorp, "A tutorial on estimating time-varying vector autoregressive models," *Multivariate Behavioral Research*, vol. 56, no. 1, pp. 120-149, Jan. 2021.
- [99] S. Yang, J. Kautz, and A. Hatamizadeh, "Gated delta networks: Improving Mamba2 with delta rule," *arXiv: 2412.06464*, Mar. 2025.
- [100] J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*. New York, NY, USA: Basic Books, 2018.
- [101] D. Kahneman, *Thinking, Fast and Slow*. New York, NY, USA: Farrar, Straus and Giroux, 2013.