foundations

MDPI

Article

Mathematical Formalism and Physical Models for Generative Artificial Intelligence

Zeqian Chen

Wuhan Institute of Physics and Mathematics, IAPM, Chinese Academy of Sciences, 30 West District, Xiao-Hong-Shan, Wuhan 430071, China; chenzeqian@hotmail.com

Abstract

This paper presents a mathematical formalism for generative artificial intelligence (GAI). Our starting point is an observation that a "histories" approach to physical systems agrees with the compositional nature of deep neural networks. Mathematically, we define a GAI system as a family of sequential joint probabilities associated with input texts and temporal sequences of tokens (as physical event histories). From a physical perspective on modern chips, we then construct physical models realizing GAI systems as open quantum systems. Finally, as an illustration, we construct physical models realizing large language models based on a transformer architecture as open quantum systems in the Fock space over the Hilbert space of tokens. Our physical models underlie the transformer architecture for large language models.

Keywords: generative artificial intelligence; attention mechanism; large language model; sequential joint probability; event histories; open quantum system; Kraus operator; Fock space

1. Introduction

Generative artificial intelligence (AI) models are important for modeling intelligent machines as physically described in [1,2]. Generative AI is based on deep neural networks (DNNs for short), and a common characteristic of DNNs is their compositional nature (cf. [3]): data is processed sequentially, layer by layer, resulting in a discrete-time dynamical system. The introduction of the transformer architecture for generative AI in 2017 marked the most striking advancement in terms of DNNs (cf. [4]). Indeed, the transformer is a model architecture eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output. At each step, the model is auto-regressive, consuming the previously generated symbols as additional input when generating the next. The transformer has achieved great success in natural language processing (cf. [5]).

The transformer has a modularization framework and is constructed by two main building blocks: self-attention and feed-forward neural networks. Self-attention is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence. However, despite its meteoric rise within deep learning, we believe there is a gap in our theoretical understanding of what the transformer is and why it works physically (cf. [6]).

We think that there are two origins for the modularization framework of generative AI models. One is a mathematical origin in which a joint probability distribution can be computed by sequentially conditional probabilities. For instance, the probability of

check for updates

Received: 13 May 2025
Revised: 9 June 2025
Accepted: 16 June 2025
Published: 24 June 2025

Citation: Chen, Z. Mathematical Formalism and Physical Models for Generative Artificial Intelligence. *Foundations* **2025**, *5*, 23. https://doi.org/10.3390/foundations5030023

generating a text $t_1t_2\cdots t_N$ given an input X in a transformer architecture is equal to the joint probability distribution $P_X(t_1,\cdots,t_N)$ such that

$$P_X(t_1,\cdots,t_N) = P_X(t_1)P_X(t_2|t_1)\cdots P_X(t_N|t_1,\ldots,t_{N-1}), \tag{1}$$

where the conditional probability $P_X(t_\ell|t_1,\ldots,t_{\ell-1})$ is given by the ℓ-th attention block in the transformer. Another is a physical origin, in which a physical process is considered to be a sequence of events as a history. As such, generating a text $t_1t_2\cdots t_N$ given an input X in a physical machine is a process in which, given an input X at time τ_0, an event $|t_1\rangle\langle t_1|$ occurs at time τ_1, an event $|t_2\rangle\langle t_2|$ occurs at time τ_2, ..., and last, an event $|t_N\rangle\langle t_N|$ occurs at time τ_N. A theory of the "histories" approach to physical systems was established by Isham [2], and the mathematical theory of it associated with joint probability distributions was then developed by Gudder [1]. Based on their theory, in this paper, we present a mathematical formalism for generative AI and describe the associated physical models.

To the best of our knowledge, physical models for generative AI are usually described by using systems of mean-field interacting particles (cf. [7,8] and references therein); i.e., generative AI models are regarded as classical statistical systems. However, since modern chips process data by controlling the flow of electric current, i.e., the dynamics of many electrons, they should be regarded as quantum statistical ensembles and open quantum systems from a physical perspective (cf. [9,10]). Consequently, based on our mathematical formalism for generative AI, we construct physical models realizing generative AI systems as open quantum systems. As an illustration, we construct physical models realizing large language models based on a transformer architecture as open quantum systems in the Fock space over the Hilbert space of tokens.

The paper is organized as follows. In Section 2, we include some notation and definitions on the attention mechanism, the transformer, and the effect algebras. In Section 3, we give the definition of a generative AI system as a family of sequential joint probabilities associated with input texts and temporal sequences of tokens. This is based on the mathematical theory developed by Gudder (cf. [1]) for a historical approach to physical evolution processes. Those joint probabilities characterize the attention mechanisms as well as the mathematical structure of the transformer architecture. In Section 4, we present the construction of physical models realizing generative AI systems as open quantum systems. Our physical models are given by an event-history approach to physical systems; we refer to [2] for the background of physics for this formulation. In Section 5, we construct physical models realizing large language models based on a transformer architecture as open quantum systems in the Fock space over the Hilbert space of tokens. Finally, in Section 6, we give a summary of our innovative points listed item by item and conclude the contributions of the paper.

2. Preliminaries

In this section, we present a mathematical description of the attention mechanism and transformer architecture for generative AI and include some notations and basic properties of σ-effect algebras (cf. [11]). For the sake of convenience, we collect some notations and definitions. Denote by $\mathbb{N}$ the natural number set $\{1,2,\cdots\}$, and for $n\in\mathbb{N}$, we use the notation $[n]$ to represent the set $\{1,\ldots,n\}$. For $d\in\mathbb{N}$, we denote by $\mathbb{R}^d$ the d-dimensional Euclid space with the usual inner product $\langle -,-\rangle$. For two sets X,Y, we denote by $\mathrm{Hom}(X,Y)$ the set of all maps from X into Y. For a set S, we denote $S^* = \bigcup_{n\in\mathbb{N}} S^{(n)}$, where $S^{(n)}$ is the set of all sequences $(s_1,\ldots,s_n)$ of n elements in S,; i.e., S^* is the set of all finite sequences of elements in S.

2.1. Deep Neural Networks

A DNN is constructed by connecting multiple neurons. Recall that a (feed-forward) neural network of depth L consists of some number of neurons arranged in $L+1$ layers. Layer $\ell = 0$ is the input layer, where data is presented to the network, while layer $\ell = L$ is where the output is read out. All layers in between are referred to as the hidden layers, and each hidden layer has an activation that is a map in the same layer. Specifically, let $\{X_\ell\}_{\ell=0}^{L}$ be a sequence of sets where X_ℓ indexes the neurons in layer ℓ, and let $\{V_\ell\}_{\ell=0}^{L}$ be a sequence of vector spaces. A mapping $\Phi : \mathrm{Hom}(X_0, V_0) \mapsto \mathrm{Hom}(X_L, V_L)$ is called a feed-forward neural network of depth L if there exists a sequence $\{W_\ell\}_{\ell=1}^{L}$ of maps $W_\ell : \mathrm{Hom}(X_{\ell-1}, V_{\ell-1}) \mapsto \mathrm{Hom}(X_\ell, V_\ell)$ and a sequence $\{\sigma_\ell\}_{\ell=1}^{L-1}$ of maps $\sigma_\ell : V_\ell \mapsto V_\ell$, which is called the activation function at the layer ℓ, such that

$$\Phi(f_0) = W_L(\sigma_{L-1}(W_{L-1} \cdots \sigma_1(W_1(f_0)) \cdots)), \tag{2}$$

for $f_0 \in \mathrm{Hom}(X_0, V_0)$, where f_0 is called the input and $f_L = \Phi(f_0) \in \mathrm{Hom}(X_L, V_L)$ is the output. We call $(\{W_\ell\}_{\ell=1}^{L}, \{\sigma_\ell\}_{\ell=1}^{L-1})$ the architecture of the neural network Φ. Of course, Φ is determined by its architecture, and there exist different choices of architectures yielding the same Φ.

In their most basic form, X_ℓ is a finite set of n_ℓ elements and $V_\ell = \mathbb{R}$, a feed-forward neural network $\Phi : \mathbb{R}^{n_0} \mapsto \mathbb{R}^{n_L}$ is a function of the following form: the input is $x^{(0)} = x \in \mathbb{R}^{n_0}$, $x^{(\ell)} = \sigma_\ell(W_\ell(x^{(\ell-1)}))$ for $\ell = 1, \ldots, L-1$, and

$$\Phi(x) = x^{(L)} = W_L(\sigma_{L-1}(W_{L-1} \cdots \sigma_1(W_1(x^{(0)})) \cdots)), \tag{3}$$

where $x^{(L)} \in \mathbb{R}^{n_L}$ is the output. This can be illustrated as follows:

$$\begin{array}{ccccccccc} \mathbb{R}^{n_0} & \xrightarrow{W_1} & \mathbb{R}^{n_1} & \xrightarrow{\sigma_1} & \mathbb{R}^{n_1} & \cdots & \mathbb{R}^{n_{L-2}} & \xrightarrow{W_{L-1}} & \mathbb{R}^{n_{L-1}} \xrightarrow{\sigma_{L-1}} \mathbb{R}^{n_{L-1}} \\ & & & & & & & & \Big\downarrow W_L \\ & & & & & & & \Phi & \mathbb{R}^{n_L}. \end{array}$$

Here, the map $W_\ell : \mathbb{R}^{n_{\ell-1}} \mapsto \mathbb{R}^{n_\ell}$ is usually of the form

$$W_\ell x^{(\ell-1)} = A_\ell x^{(\ell-1)} + b_\ell, \quad \ell = 1, \ldots, L, \tag{4}$$

where A_ℓ is an $n_\ell \times n_{\ell-1}$ matrix called a weight matrix and $b_\ell \in \mathbb{R}^{n_\ell}$ is called a bias vector for each ℓ, and the function $\sigma_\ell : \mathbb{R}^{n_\ell} \mapsto \mathbb{R}^{n_\ell}$ represents the activation function at the ℓ-th layer. The set of all entries of the weight matrices and bias vectors of a neural network Φ are called the parameters of Φ. These parameters are adjustable and learned during the training process, determining the specific function realized by the network. Also, the depth L, the number of neurons in each layer, and the activation functions of a neural network Φ are called the hyperparameters of Φ. They define the network's architecture (and training process) and are typically set before training begins. For a fixed architecture, every choice of network parameters as in (3) defines a specific function Φ, and this function is often referred to as a model.

In a feed-forward neural network, the inputs to neurons in the ℓ-th layer are usually exclusively neurons from the $(\ell-1)$-th layer. However, residual neural networks (ResNets for short) allow skip connections; that is, information is allowed to skip layers in the sense that the neurons in layer ℓ may have $x^{(0)}, \ldots, x^{(\ell-1)}$ as their input (and not just $x^{(\ell-1)}$). In their most basic form, $x^{(0)} = x \in \mathbb{R}^d$, and

$$x^{(\ell)} = x^{(\ell-1)} + Q_\ell \sigma(A_\ell x^{(\ell-1)} + b_\ell), \quad \ell = 1, \ldots, L-1, \tag{5}$$

where $\sigma : \mathbb{R}^d \mapsto \mathbb{R}^d$ is a vector function, Q_ℓ, A_ℓ's are $d \times d$ matrices, and b_ℓ's are vectors in $\mathbb{R}^d$. In contrast to feed-forward neural networks, recurrent neural networks (RNNs for short) allow information to flow backward in the sense that $x^{(\ell-1)}, x^{(\ell+1)}, \ldots, x^{(L)}$ may serve as input for the neurons in layer ℓ and not just $x^{(\ell-1)}$. We refer to [12] for more details, such as training for a neural network.

2.2. Attention

The fundamental definition of attention was given by Bahdanau et al. in 2014. To describe the mathematical definition of attention, we denote by $\mathbb{Q} \subset \mathbb{R}^{d_q}$ the query space, $\mathbb{K} \subset \mathbb{R}^{d_k}$ the key space, and $\mathbb{V} \subset \mathbb{R}^{d_v}$ the value space. We call an element $q \in \mathbb{Q}$ a query, $k \in \mathbb{K}$ a key, $v \in \mathbb{V}$, and so on.

Definition 1 (cf. [13]). *Let $a : \mathbb{Q} \times \mathbb{K} \mapsto \mathbb{R}$ be a function. Let $K = \{k_1, \ldots, k_N\} \subset \mathbb{K}$ be a set of keys and $V \subset \mathbb{V}$ a set of values. Given a $q \in \mathbb{Q}$, the attention* $\mathrm{Att} : (q, K, V) \mapsto \mathbb{R}$ *is defined by*

$$\mathrm{Att}(q, K, V) = \sum_{n=1}^{N} \mathrm{softmatch}_a(q, K)_n \cdot v_n, \tag{6}$$

where $\mathrm{softmatch}_a(q, K)$ *is a probability distribution over $K = \{k_1, \ldots, k_N\}$ defined by*

$$\mathrm{softmatch}_a(q, K)_n = \frac{e^{a(q,k_n)}}{\sum_{j=1}^{N} e^{a(q,k_j)}}, \quad n = 1, \ldots, N. \tag{7}$$

This means that a value v_n in (6) occurs with probability $\mathrm{softmatch}_a(q, K)_n$ *for $n \in [N]$.*

For $Q = \{q_1, \ldots, q_M\} \subset \mathbb{Q}$, we define

$$\mathrm{Att}(Q, K, V) = \{\mathrm{Att}(q_m, K, V)\}_{m=1}^{M}. \tag{8}$$

In particular, when $Q = K = V$, $\mathrm{Att}(Q, Q, Q)$ is said to be self-attention at Q, and the mapping SelfAtt *defined by*

$$\mathrm{SelfAtt}(Q) = \mathrm{Att}(Q, Q, Q), \tag{9}$$

is called the self-attention map.

We remark that

(1) For a finite sequence $\{x_n\}_{n=1}^N$ of real numbers, define

$$\mathrm{softmax}(\{x_j\}_{j=1}^N)_n = \frac{e^{x_n}}{\sum_{j=1}^{N} e^{x_j}}, \quad n \in [N]. \tag{10}$$

Then,

$$\mathrm{softmatch}_a(q, K)_n = \mathrm{softmax}(\{a(q, k_j)\}_{j=1}^N)_n, \tag{11}$$

as usual in the literature.

(2) We have $|K| = |V| = N$, but $|Q| = M \neq N$ in general.

(3) The function $a : \mathbb{Q} \times \mathbb{K} \mapsto \mathbb{R}$ is called a similarity function, usually given by

$$a(q, k) = \frac{1}{\sqrt{d'}} \langle W^Q q, W^K k \rangle, \tag{12}$$

where W^Q is a $d' \times d_q$ real matrix called a query matrix and W^K is a $d' \times d_k$ real matrix called a key matrix. For $q \in \mathbb{Q}, k \in \mathbb{K}$, the real number $a(q, k)$ is interpreted as the similarity between the query q and the key k.

(4) In the representation learning framework of attention, we usually assume the finite set **T** of tokens has been embedded in $\mathbb{R}^d$, where d is called the embedding dimension, so we identify each $t \in \mathbf{T}$ with one of finitely-many vectors x in $\mathbb{R}^d$. We assume that the structure (positional information, adjacency information, etc) is encoded in these vectors. In the case of self-attention, we assume $d_q = d_k = d_v = d$.

Since the self-attention mechanism can be composed to arbitrary depth, making it a crucial building block of the transformer architecture, we mainly focus on it in what follows. In practice, we need multi-headed attention (cf. [4]), that process independent copies of the data X and combine them with concatenation and matrix multiplication. Let $X = \{x_n\}_{n=1}^N$ be the input set of tokens embedded in $\mathbb{R}^d$. Let us consider n_h-headed attention with the dimension d_h for every head. For every $i \in [n_h]$, let W_i^Q, W_i^K, W_i^V be $d_h \times d$ (query, key, value) matrices associated with the i-th self-attention, and the similarity function

$$a_i(x, y) = \frac{1}{\sqrt{d_h}} \langle W_i^Q x, W_i^K y \rangle. \tag{13}$$

Let $W^O = [W_1^O, \ldots, W_{n_h}^O]$ denote the output projection matrix, where W_i^O is a $d \times d_h$ matrix for every $i \in [n_h]$. For $n \in [N]$, the multi-headed self-attention (MHSelfAtt for short) is then defined by

$$\mathrm{MHSelfAtt}(x_n, X, X) = \sum_{i=1}^{n_h} \sum_{j_i=1}^{n} \mathrm{softmax}(\{\frac{1}{\sqrt{d_h}} \langle W_i^Q x_n, W_i^K x_\ell \rangle\}_{\ell=1}^n)_{j_i} [W_i^O (W_i^V x_{j_i})], \tag{14}$$

that is, an output

$$u_n = \sum_{i=1}^{n_h} W_i^O (W_i^V x_{j_i}), \quad j_i \in [n], \tag{15}$$

occurs with the probability $\prod_{i=1}^{n_h} \mathrm{softmax}(\{\frac{1}{\sqrt{d_h}} \langle W_i^Q x_n, W_i^K x_\ell \rangle\}_{\ell=1}^n)_{j_i}$. As such,

$$\mathrm{MHSelfAtt}(X) = \{\mathrm{MHSelfAtt}(x_n, X, X)\}_{n=1}^N, \tag{16}$$

yields a basic building block of the transformer

$$\mathrm{Transf}(X) = \mathrm{FFN} \circ \mathrm{MHSelfAtt}(X), \tag{17}$$

as in the case of one-headed attention.

2.3. Transformer

In line with successful models, such as large language models, we focus on the decoder-only setting of the transformer, where the model iteratively predicts the next tokens based on a given sequence of tokens. This procedure is called autoregressive since the prediction of new tokens is only based on previous tokens. Such conditional sequence generation using autoregressive transformers is referred to as the transformer architecture.

Specifically, in the transformer architecture defined by a composition of blocks, each block consists of a self-attention layer SelfAtt, a multi-layer perception FFN, and a prediction head layer PH. First, the self-attention layer SelfAtt is the only layer that combines different tokens. Let us denote the input text to the layer by $X = \{x_n\}_{n=1}^N$ embedded in $\mathbb{R}^d$ and focus on the n-th output. For each $n \in [N]$, letting

$$s_j^{(n)} = \frac{1}{\sqrt{d}} \langle W^Q x_n, W^K x_j \rangle, \quad \forall j \in [n], \tag{18}$$

where W^Q and W^K are two $d' \times d$ matrices (i.e., the query and key matrices), we can interpret $S^{(n)} = \{s_j^{(n)}\}_{j=1}^n$ as similarities between the n-th token x_n (i.e., the query) and the other tokens (i.e., keys); for satisfying the autoregressive structure, we only consider $j = 1, \ldots, n$. The softmax layer is given by

$$\text{softmax}(S^{(n)})_j = \frac{e^{s_j^{(n)}}}{\sum_{i=1}^n e^{s_i^{(n)}}}, \quad \forall j \in [n], \tag{19}$$

which can be interpreted as the probability for the n-th query to "attend" to the j-th key. Then, the self-attention layer SelfAtt can be defined as

$$\text{SelfAtt}(X)_n = \sum_{j=1}^n \text{softmax}(S^{(n)})_j W^V x_j, \quad n \in [N], \tag{20}$$

where W^V is the $d \times d$ real matrix such that $W^V x \in \mathbf{T}$ for any $x \in \mathbf{T}$, the output $W^V x_j$ occurring with the probability $\text{softmax}(S^{(n)})_j$ is often referred to as the values of the token x_j. Thus, SelfAtt : $(\mathbb{R}^d)^* \mapsto (\mathbb{R}^d)^*$ is a random map such that $\text{SelfAtt}[(\mathbb{R}^d)^{(N)}] \subset (\mathbb{R}^d)^{(N)}$ for each $N \in \mathbb{N}$.

If the attention is a multi-headed attention with n_h heads of the dimension d_h, where for $i \in [n_h]$, W_i^Q, W_i^K, W_i^V are the $d_h \times d$ (query, key, value) matrices and W_i^O is the $d \times d_h$ (output) matrix of the i-th self-attention, then the multi-headed self-attention layer MHSelfAtt is defined by

$$\text{MHSelfAtt}(X)_n = \sum_{i=1}^{n_h} \sum_{j_i=1}^n \text{softmax}(S_i^{(n)})_{j_i} [W_i^O (W_i^V x_{j_i})], \quad n \in [N], \tag{21}$$

where

$$\text{softmax}(S_i^{(n)})_{j_i} = \frac{\frac{1}{\sqrt{d_h}} \langle W_i^Q x_n, W_i^K x_{j_i} \rangle}{\sum_{\ell=1}^n \frac{1}{\sqrt{d_h}} \langle W_i^Q x_n, W_i^K x_\ell \rangle}, \quad j_i \in [n], \tag{22}$$

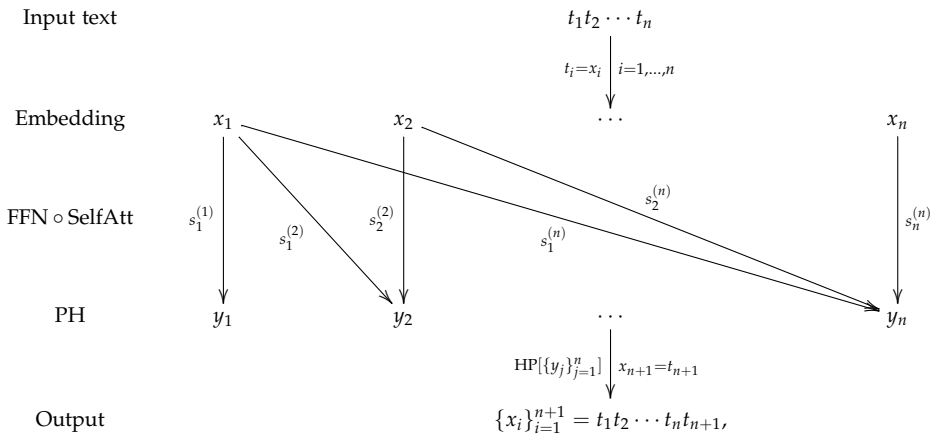
i.e., an output $u_n = \sum_{i=1}^{n_h} W_i^O (W_i^V x_{j_i})$ occurs with the probability $\prod_{i=1}^{n_h} \text{softmax}(S_i^{(n)})_{j_i}$ for each $n \in [N]$. In what follows, we only consider the case of one-headed attention, since the multi-headed case is similar.

Second, the multi-layer perception is a feed-forward neural network FFN such that $y_n = \text{FFN}(W^V x_j)$ with the probability $\text{softmax}(S^{(n)})_j$ ($j \in [n]$) for each $n \in [N]$. Finally, the prediction head layer can be represented as a mapping PH : $(\mathbb{R}^d)^* \mapsto [0,1]^*$, which maps the sequence of $\{y_n\}_{n=1}^N$ to a probability distribution $\{p_n\}_{n=1}^N$, where p_n is the probability of predicting y_n as the next token. Since y_N contains information about the whole input text, we may define

$$\text{PH}[\{y_n\}_{n=1}^N] = \sum_{j=1}^N \text{softmax}(S^{(N)})_j \text{FFN}(W^V x_j), \tag{23}$$

such that the next token $x_{N+1} = y_j = \text{FFN}(W^V x_j)$ with the probability $\text{softmax}(S^{(N)})_j$ for $j \in [N]$.

Hence, a basic building block for the transformer, consisting of a self-attention module (SelfAtt) and a feed-forward network (FFN) followed by a prediction head layer (PH), can be illustrated as follows:

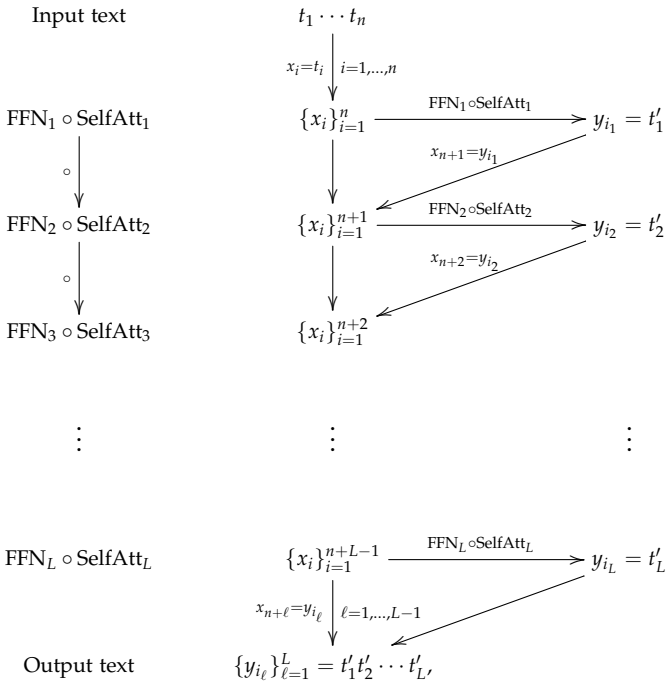


where the input text $t_1t_2\cdots t_n$ is embedded as a sequence $\{x_i\}_{i=1}^n$ in $\mathbb{R}^d$, $y_j = \mathrm{FFN}(W^V x_j)$ occurs with the probability $\mathrm{softmax}(S^{(n)})_j$ for each $j \in [n]$, $x_{n+1} = y_j$ is generated with the probability $\mathrm{softmax}(S^{(n)})_j$ for each $j \in [n]$, and so the output is $x_{n+1} = \mathrm{PH} \circ \mathrm{FFN} \circ \mathrm{SelfAtt}(\{x_i\}_{i=1}^n)$. One can then apply the same operations to the extended sequence $x_1x_2\cdots x_nx_{n+1}$ in the next block, obtaining $x_{n+2} = \mathrm{PH} \circ \mathrm{FFN} \circ \mathrm{SelfAtt}(\{x_i\}_{i=1}^{n+1})$, to iteratively compute further tokens (there is usually a stopping criterion based on a special token or the mapping PH). Below, without loss of generality, we omit the prediction head layer PH.

Typically, a transformer of depth L is defined by a composition of L blocks, denoted by Transf_L, consisting of L self-attention maps $\{\mathrm{SelfAtt}_\ell\}_{\ell=1}^L$ and L feed-forward neural networks $\{\mathrm{FFN}_\ell\}_{\ell=1}^L$, that is,

$$\mathrm{Transf}_L = (\mathrm{FFN}_L \circ \mathrm{SelfAtt}_L) \circ \cdots \circ (\mathrm{FFN}_1 \circ \mathrm{SelfAtt}_1), \tag{24}$$

where the indices of the layers SelfAtt and FFN in (24) indicate the use of different trainable parameters in each of the blocks. This can be illustrated as follows:



that is,

$$\text{Transf}_L(t_1 \cdots t_n) = t_1' t_2' \cdots t_L'. \tag{25}$$

Also, we can consider the transformer of the form

$$\text{Transf}_L = ((\text{id} + \text{FFN}_L) \circ (\text{id} + \text{SelfAtt}_L)) \circ \cdots \circ ((\text{id} + \text{FFN}_1) \circ (\text{id} + \text{SelfAtt}_1)), \tag{26}$$

where id denotes the identity mapping in $\mathbb{R}^d$, commonly known as a skip or residual connection.

2.4. Effect Algebras

For the sake of convenience, we collect some notations and basic properties of σ-effect algebras (cf. [1,11,14] and references therein). Recall that an effect algebra is an algebraic system $(\mathbf{E}, 0, 1, \oplus)$, where $\mathbf{E}$ is a non-empty set, $0, 1 \in \mathbf{E}$, which are called zeroes and unit elements of this algebra, respectively, and $\oplus$ is a partial binary operation on $\mathbf{E}$ that satisfies the following conditions for any $a, b, c \in \mathbf{E}$:

(E1) (Commutative Law): If $a \oplus b$ is defined, then $b \oplus a$ is defined and $b \oplus a = a \oplus b$, which is called the orthogonal sum of a and b;

(E2) (Associative Law): If $a \oplus b$ and $(a \oplus b) \oplus c$ are defined, then $b \oplus c$ and $a \oplus (b \oplus c)$ are defined and

$$(a \oplus b) \oplus c = a \oplus (b \oplus c),$$

which is denoted by $a \oplus b \oplus c$;

(E3) (Orthosupplementation Law): there exists a unique $a' \in \mathbf{E}$ such that $a \oplus a'$ is defined and $a \oplus a' = 1$, such a' is unique and called the orthosupplement of a;

(E4) (Zero–One Law): if $a \oplus 1$ is defined, then $a = 0$.

We simply call $\mathbf{E}$ an effect algebra in the sequel. From the associative law (E2), we can write $a_1 \oplus a_2 \oplus \cdots \oplus a_n$ if this orthogonal sum is defined. For any $a, b \in \mathbf{E}$, we define $a \leq b$ if there exists a $c \in \mathbf{E}$ such that $a \oplus c = b$; this c is unique and denoted by $c = b \ominus a$, so $a' = 1 \ominus a$. We also define $a \perp b$ if $a \oplus b$ is defined; i.e., a is orthogonal to b. It can be shown (cf. [14]) that $(\mathbf{E}, \leq)$ is a bounded partially ordered set (poset for short) and $a \perp b$ if and only if $a \leq b'$. For a sequence $\{a_i\}_{i=1}^{\infty}$ in $\mathbf{E}$, if $a_1 \oplus \cdots \oplus a_n$ is defined for all $n \in \mathbb{N}$ such that $\bigvee_{n=1}^{\infty}(a_1 \oplus \cdots \oplus a_n)$ exists, then the sum $\bigoplus_i a_i$ of $\{a_i\}_{i=1}^{\infty}$ exists and define $\bigoplus_i a_i = \bigvee_{n=1}^{\infty}(a_1 \oplus \cdots \oplus a_n)$. We say that $\mathbf{E}$ is a σ-effect algebra if $\bigoplus_i a_i$ exists for any sequence $\{a_i\}_{i=1}^{\infty}$ in $\mathbf{E}$ satisfying that $a_1 \oplus \cdots \oplus a_n$ is defined for all $n \in \mathbb{N}$. It was shown in (Lemma 3.1, [1]) that $\mathbf{E}$ is a σ-effect algebra if and only if the least upper bound $\bigvee_i a_i$ exists for any monotone sequence $\{a_i\}_{i=1}^{\infty}$, i.e., $a_1 \leq a_2 \leq \cdots$.

Let $\mathbf{E}$ and $\mathbf{F}$ be σ-effect algebras. A map $\phi : \mathbf{E} \mapsto \mathbf{F}$ is said to be additive if for $a, b \in \mathbf{E}$, $a \perp b$ implies that $\phi(a) \perp \phi(b)$ and $\phi(a \oplus b) = \phi(a) \oplus \phi(b)$. An additive map $\phi : \mathbf{E} \mapsto \mathbf{F}$ is σ-additive if for any sequence $\{a_i\}_{i=1}^{\infty}$ such that $\bigoplus_i a_i$ exists, $\bigoplus_i \phi(a_i)$ exists and $\phi(\bigoplus_i a_i) = \bigoplus_i \phi(a_i)$. A σ-additive map $\phi : \mathbf{E} \mapsto \mathbf{F}$ is said to be a σ-morphism if $\phi(1) = 1$; and moreover, ϕ is called a σ-isomorphism if ϕ is a bijective σ-morphism and $\phi^{-1} : \mathbf{F} \mapsto \mathbf{E}$ is a σ-morphism. It can be shown (cf. [1]) that

(1) A map $\phi : \mathbf{E} \mapsto \mathbf{F}$ is additive if and only if ϕ is monotone in the sense that $a \leq b$ implies $\phi(a) \leq \phi(b)$;

(2) An additive map ϕ is σ-additive if and only if $a_1 \leq a_2 \leq \cdots$ implies $\phi(\bigvee_i a_i) = \bigvee_i \phi(a_i)$;

(3) A σ-morphism ϕ satisfies $\phi(a') = \phi(a)'$.

The unit interval $[0, 1]$ is a σ-effect algebra defined as follows: For any $a, b \in [0, 1]$, $a \oplus b$ is defined if $a + b \leq 1$ and in this case $a \oplus b = a + b$. Then, we have that $a' = 1 - a$, and

$0, 1$ are the zero and unit elements, respectively. In what follows, we always regard $[0, 1]$ as a σ-effect algebra in this way. Let $\mathbf{E}$ be a σ-effect algebra, a σ-morphism $\phi : \mathbf{E} \mapsto [0, 1]$ is called a state on $\mathbf{E}$, and we denote by $\mathcal{S}(\mathbf{E})$ the set of all states on $\mathbf{E}$. A subset S of $\mathcal{S}(\mathbf{E})$ is said to be order determining if $\alpha(a) \leq \alpha(b)$ for all $\alpha \in S$ implies $a \leq b$.

Another example of a σ-effect algebra is a measurable space $(\Omega, \mathcal{F})$ defined as follows: For any $A, B \in \mathcal{F}$, $A \oplus B$ is defined if $A \cap B = \emptyset$, and in this case, $A \oplus B = A \cup B$. We then have $0 = \emptyset, 1 = \Omega$, and $A' = \Omega \backslash A$. We always regard a measurable space $(\Omega, \mathcal{F})$ as a σ-effect algebra in this way. Let $\mathbf{E}$ be a σ-effect algebra, a σ-morphism $X : (\Omega, \mathcal{F}) \mapsto \mathbf{E}$ is called an observable on $\mathbf{E}$ with values in $(\Omega, \mathcal{F})$ (a Ω-valued observable for short). The elements of a σ-effect algebra are called effects, and so an observable X maps effects in $\mathcal{F}$ into effects in $\mathbf{E}$; i.e., $X(A)$ is an effect in $\mathbf{E}$ for $A \in \mathcal{F}$. We denote by $\mathcal{O}(\mathbf{E}, \Omega, \mathcal{F})$ the set of all Ω-valued observables. Note that $\mathcal{S}(\Omega, \mathcal{F})$ is equal to the set of all probability measures on $(\Omega, \mathcal{F})$. For $\alpha \in \mathcal{S}(\mathbf{E})$ and $X \in \mathcal{O}(\mathbf{E}, \Omega, \mathcal{F})$, we have $\alpha \circ X \in \mathcal{S}(\Omega, \mathcal{F})$, which is called the probability distribution of X in the state α.

3. Mathematical Formalism

In this section, we introduce a mathematical formalism for generative AI. We utilize the theory of σ-effect algebras to give a mathematical definition for a generative AI system. Let $\mathbf{E}$ be a σ-effect algebra and $(\Omega, \mathcal{F})$ a measurable space. An orthogonal decomposition in $\mathbf{E}$ is a sequence $\{a_i\}$ in $\mathbf{E}$ such that $\bigoplus_i a_i$ exists, and moreover, it is complete if $\bigoplus_i a_i = 1$. We denote by $\mathcal{D}(\mathbf{E})$ the set of all completely orthogonal decomposition in $\mathbf{E}$. A completely orthogonal decomposition in $\mathcal{F}$ is called a countable partition of Ω, i.e., a sequence $\{A_i\}$ of elements in $\mathcal{F}$ such that $A_i \cap A_j$ for $i \neq j$ and $\cup_i A_i = \Omega$. We denote by $\mathcal{D}(\Omega, \mathcal{F})$ the set of all countable partitions of Ω. For $n \in \mathbb{N}$, an ordered n-tuple $\vec{R} = (e_1, \ldots, e_n)$ of effects in $\mathbf{E}$ is called a n-time chain-of-effect, and we interpret $\vec{R}$ as an inference process of an intelligence machine in which the effect e_i occurs at time τ_i for $i \in [n]$, where $\tau_1 < \tau_2 < \cdots < \tau_n$. Alternatively, no specific times may be involved and we regard $\vec{R}$ as a sequential effect in which e_1 occurs first, e_2 occurs second, $\ldots$, and e_n occurs last.

Definition 2. *With the above notations, a generative artificial intelligence system $\mathfrak{S}$ is defined to be a triple $(\mathbf{E}, \Omega, \mathcal{F})$, where $\mathbf{E}$ is a σ-effect algebra, $(\Omega, \mathcal{F})$ is a measurable space, such that*

(G1) The input set $\mathbf{In}(\mathfrak{S})$ of $\mathfrak{S}$ is equal to the set $\mathcal{S}(\mathbf{E})$; i.e., an input is interpreted by a state $\alpha \in \mathcal{S}(\mathbf{E})$;

(G2) The output set $\mathbf{Out}(\mathfrak{S})$ of $\mathfrak{S}$ is equal to the set $\Omega^* = \bigcup_{n=1}^{\infty} \Omega^{(n)}$; i.e., the set of all finite sequences of elements in Ω;

(G3) An inference process in $\mathfrak{S}$ is interpreted by a chain-of-effect $(e_1, \cdots, e_n)$ for $n \in \mathbb{N}$.

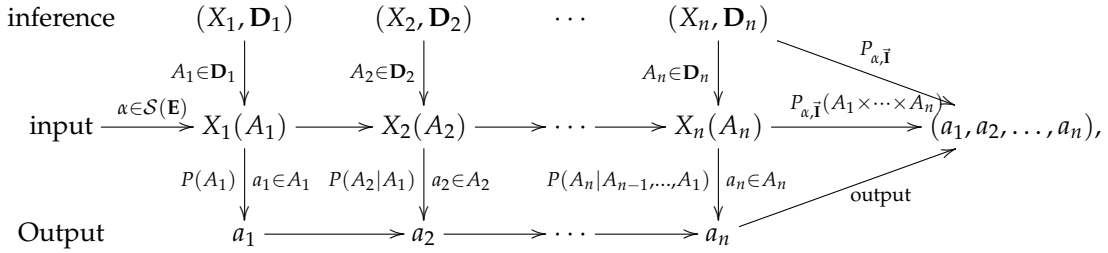
Remark 1. *We refer to [15] for a mathematical definition of general artificial intelligence systems in terms of topos theory, including quantum artificial intelligence systems.*

In practice, we are not concerned with a generative AI system $\mathfrak{S} = (\mathbf{E}, \Omega, \mathcal{F})$ itself but deal with models for $\mathfrak{S}$, such as large language models. To this end, we need to introduce the definition of a model for $\mathfrak{S}$ in terms of joint probability distributions for observables associated with $\mathfrak{S}$.

For $X \in \mathcal{O}(\mathbf{E}, \Omega, \mathcal{F})$ and $A \in \mathcal{F}$, we may view the effect $X(A)$ as the event for which X has a value in A. For a partition $\mathbf{D} = \{A_i\} \in \mathcal{D}(\Omega, \mathcal{F})$, we may view $(X, \mathbf{D})$ as a set of possible alternative events that can occur. One interpretation is that $(X, \mathbf{D})$ represents a building block of an artificial intelligence architecture for processing X and the alternatives result from the dial readings of the block. Given $X_i \in \mathcal{O}(\mathbf{E}, \Omega, \mathcal{F})$, $A_i \in \mathcal{F}$, $i = 1, \ldots, n$, an ordered n-tuple $\vec{R} = (X_1(A_1), \ldots, X_n(A_n))$ of events is called an n-time chain-of-event,

and we interpret $\vec{R}$ as an inference process of an intelligence machine in which X_1 has a value a_1 in A_1 first, X_2 has a_2 in A_2 s, ..., and X_n has a_n in A_n last, so that the output result is $(a_1, a_2, \ldots, a_n)$. We denote the set of all n-time chain-of-events by $\mathbf{R}^{(n)}$ and the set of all chain-of-events by $\mathbf{R}^* = \bigcup_n \mathbf{R}^{(n)}$.

A n-step inference set has the form $\vec{\mathbf{I}} = ((X_1, \mathbf{D}_1), \ldots, (X_n, \mathbf{D}_n))$, where $X_i \in \mathcal{O}(\mathbf{E}, \Omega, \mathcal{F})$, $\mathbf{D}_i \in \mathcal{D}(\Omega, \mathcal{F})$, $i \in [n]$. We interpret $\vec{\mathbf{I}}$ as ordered successive processes of observables X_i with partitions $\mathbf{D}_i$ for $i \in [n]$. We denote the collection of all n-step inference sets by $\mathbf{I}^{(n)}$ and the collection of all inference sets by $\mathbf{I}^* = \bigcup_n \mathbf{I}^{(n)}$. If $\vec{R} = (X_1(A_1), \ldots, X_n(A_n))$ and $\vec{\mathbf{I}} = ((X_1, \mathbf{D}_1), \ldots, (X_n, \mathbf{D}_n))$ such that $A_i \in \mathbf{D}_i$ for every $i \in [n]$, we say the chain-of-event $\vec{R}$ is an element of the inference set $\vec{\mathbf{I}}$ and write $\vec{R} \in \vec{\mathbf{I}}$. This can be illustrated as follows:



which means that the machine firstly obtains a_1 as part of an output with the probability $P(A_1)$, then obtains a_2 with the conditional probability $P(A_2|A_1)$, ..., and lastly obtains a_n with the conditional probability $P(A_n|A_{n-1}, \ldots, A_1)$ and finally combines them to obtain the output result $(a_1, a_2, \ldots, a_n)$ with the probability

$$P_{\alpha,\vec{\mathbf{I}}}(A_1 \times \cdots \times A_n) = P(A_1)P(A_2|A_1) \cdots P(A_n|A_{n-1}, \ldots, A_1), \tag{27}$$

where $P_{\alpha,\vec{\mathbf{I}}}$ will be explained later.

If $\vec{\mathbf{I}}_1 = ((X_1, \mathbf{D}_1), \ldots, (X_n, \mathbf{D}_n))$ and $\vec{\mathbf{I}}_2 = ((Y_1, \mathbf{J}_1), \ldots, (Y_m, \mathbf{J}_m))$ are two inference sets, then we define their sequential product by

$$\vec{\mathbf{I}} = \vec{\mathbf{I}}_1\vec{\mathbf{I}}_2 = ((X_1, \mathbf{D}_1), \ldots, (X_n, \mathbf{D}_n), (Y_1, \mathbf{J}_1), \ldots, (Y_m, \mathbf{J}_m)), \tag{28}$$

and obtain a $(n+m)$-step inference set. Mathematically, we can include the empty inference set $\emptyset$ that satisfies $\emptyset\vec{\mathbf{I}} = \vec{\mathbf{I}}\emptyset = \vec{\mathbf{I}}$, such that $\mathbf{I}^*$ becomes a semigroup under this product.

For a partition $\mathbf{D} \in \mathcal{D}(\Omega, \mathcal{F})$, we denote by $\sigma(\mathbf{D})$ the σ-subalgebra of $\mathcal{F}$ generated by $\mathbf{D}$, and for n partitions $\{\mathbf{D}_i\}_{i=1}^n$, we denote by $\sigma(\{\mathbf{D}_i\}_{i=1}^n)$ the σ-algebra on $\Omega^{(n)}$ generated by $\{\mathbf{D}_i\}_{i=1}^n$, i.e.,

$$\sigma(\{\mathbf{D}_i\}_{i=1}^n) = \sigma(\{A_1 \times \cdots \times A_n \subseteq \Omega^{(n)} : \ A_i \in \mathbf{D}_i, \ i \in [n]\}). \tag{29}$$

We denote by $\mathcal{P}(\Omega^{(n)}, \sigma(\{\mathbf{D}_i\}_{i=1}^n))$ the set of all probability measures on $(\Omega^{(n)}, \sigma(\{\mathbf{D}_i\}_{i=1}^n))$. Also, we write $\sigma(\vec{\mathbf{I}}) = \sigma(\{\mathbf{D}_i\}_{i=1}^n)$ for $\vec{\mathbf{I}} = ((X_1, \mathbf{D}_1), \ldots, (X_n, \mathbf{D}_n))$. Given an input $\alpha \in \mathcal{S}(\mathbf{E})$, for an inference set $\vec{\mathbf{I}} = ((X_1, \mathbf{D}_1), \ldots, (X_n, \mathbf{D}_n))$, we denote by $P_{\alpha,\vec{\mathbf{I}}} \in \mathcal{P}(\Omega^{(n)}, \sigma(\vec{\mathbf{I}}))$ the probability measure such that for $A_1 \times \cdots \times A_n \in \sigma(\{\mathbf{D}_i\}_{i=1}^n)$, $P_{\alpha,\vec{\mathbf{I}}}(A_1 \times \cdots \times A_n)$ is the probability within the inference set $\vec{\mathbf{I}}$ that the event $X_1(A_1)$ occurs first, $X_2(A_2)$ occurs second, ..., $X_n(A_n)$ occurs last. We call $P_{\alpha,\vec{\mathbf{I}}} \in \mathcal{P}(\Omega^{(n)}, \sigma(\vec{\mathbf{I}}))$ the joint probability distribution of an inference set $\vec{\mathbf{I}}$ under the input $\alpha \in \mathcal{S}(\mathbf{E})$.

For interpreting a model for a generative AI system, $P_{\alpha,\vec{\mathbf{I}}}$'s need to satisfy physically motivated axioms as follows.

Definition 3. *With the above notations, a model $\mathcal{M}$ for $\mathfrak{S} = (\mathbf{E}, \Omega, \mathcal{F})$ is defined to be a family of joint probability distributions of inference sets*

$$\mathcal{M} = \bigcup_{n\in\mathbb{N}} \{P_{\alpha,\vec{\mathbf{I}}} \in \mathcal{P}(\Omega^{(n)}, \sigma(\vec{\mathbf{I}})) : \ \alpha \in \mathcal{S}(\mathbf{E}), \ \vec{\mathbf{I}} \in \mathbf{I}^{(n)}\}, \tag{30}$$

that satisfies the following axioms:

(P1) *For $\vec{\mathbf{I}}_1 = (X, \mathbf{D}), \vec{\mathbf{I}}_2 = (Y, \mathbf{J}) \in \mathbf{I}^{(1)}$ and $A \in \sigma(\mathbf{D}), B \in \sigma(\mathbf{J})$, if $P_{\alpha,\vec{\mathbf{I}}_1}(A) = P_{\alpha,\vec{\mathbf{I}}_2}(B)$ for all $\alpha \in \mathcal{S}(\mathbf{E})$, then $X(A) = Y(B)$.*

(P2) *For $\vec{\mathbf{I}} \in \mathbf{I}^*, \vec{\mathbf{I}}_i = (X, \mathbf{D}_i), i = 1, 2$, if $A \in \sigma(\vec{\mathbf{I}})$ and $B \in \sigma(\mathbf{D}_1) \cap \sigma(\mathbf{D}_2)$, then*

$$P_{\alpha,\vec{\mathbf{I}}\vec{\mathbf{I}}_1}(A \times B) = P_{\alpha,\vec{\mathbf{I}}\vec{\mathbf{I}}_2}(A \times B), \tag{31}$$

for every $\alpha \in \mathcal{S}(\mathbf{E})$.

(P3) *For $\vec{\mathbf{I}} \in \mathbf{I}^*, \vec{\mathbf{J}} = (X, \mathbf{D})$ with $\mathbf{D} = \{B_i\}$, if $A \in \sigma(\vec{\mathbf{I}})$ then*

$$P_{\alpha,\vec{\mathbf{I}}\vec{\mathbf{J}}}(A \times \Omega) = \sum_i P_{\alpha,\vec{\mathbf{I}}\vec{\mathbf{J}}}(A \times B_i) = P_{\alpha,\vec{\mathbf{I}}}(A), \quad \forall \alpha \in \mathcal{S}(\mathbf{E}). \tag{32}$$

(P4) *If $\vec{\mathbf{I}}_1 = ((X_1, \mathbf{D}_1), \ldots, (X_n, \mathbf{D}_n))$, $\vec{\mathbf{I}}_2 = ((X_1, \mathbf{J}_1), \ldots, (X_n, \mathbf{J}_n))$, and $A_i \in \mathbf{D}_i \cap \mathbf{J}_i$ for $i \in [n]$, then*

$$P_{\alpha,\vec{\mathbf{I}}_1}(A_1 \times \cdots \times A_n) = P_{\alpha,\vec{\mathbf{I}}_2}(A_1 \times \cdots \times A_n), \tag{33}$$

for every $\alpha \in \mathcal{S}(\mathbf{E})$.

For the physical meanings of the model structure axioms, we remark that

(1) The axiom (P1) means that the input set can distinguish different events;
(2) The axiom (P2) means that the partition of the last processing is irrelevant;
(3) The axiom (P3) means that the last processing does not affect the previous ones;
(4) The axiom (P4) means that the probability of a chain of events does not depend on the partitions and hence is unambiguous. However, for $B \in \sigma(\vec{\mathbf{I}}_1) \cap \sigma(\vec{\mathbf{I}}_2)$ in (P4), $P_{\alpha,\vec{\mathbf{I}}_1}(B) \neq P_{\alpha,\vec{\mathbf{I}}_2}(B)$ in general if X_i's are quantum observables due to quantum interference.

If $\vec{\mathbf{I}}_1 = ((X_1, \mathbf{D}_1), \ldots, (X_n, \mathbf{D}_n))$ and $\vec{\mathbf{I}}_2 = ((Y_1, \mathbf{J}_1), \ldots, (Y_m, \mathbf{J}_m))$ are two inference sets, $A \in \sigma(\{\mathbf{D}_i\}_{i=1}^n)$, $B \in \sigma(\{\mathbf{J}_j\}_{j=1}^m)$, and if α is an input such that $P_{\alpha,\vec{\mathbf{I}}_1}(A) \neq 0$, then we define the conditional probability of B given A within $\vec{\mathbf{I}}_1\vec{\mathbf{I}}_2$ under the input α as follows:

$$P_{\alpha,\vec{\mathbf{I}}_2|\vec{\mathbf{I}}_1}(B|A) = \frac{P_{\alpha,\vec{\mathbf{I}}_1\vec{\mathbf{I}}_2}(A \times B)}{P_{\alpha,\vec{\mathbf{I}}_1}(A)}. \tag{34}$$

Since $P_{\alpha,\vec{\mathbf{I}}_1\vec{\mathbf{I}}_2}$ is a probability measure on $(\Omega^{(n+m)}, \sigma(\{\mathbf{D}'\}_{i=1}^{n+m}))$, where $\mathbf{D}'_i = \mathbf{D}_i$ for $i \in [n]$ and $\mathbf{D}'_{n+j} = \mathbf{J}_j$ for $j \in [m]$, so $P_{\alpha,\vec{\mathbf{I}}_2|\vec{\mathbf{I}}_1}(-|A)$ is a probability measure on $(\Omega^m, \sigma(\{\mathbf{J}\}_{j=1}^m))$, which is called a conditional sequential joint probability distribution.

Proposition 1. *Given $\alpha \in \mathcal{S}(\mathbf{E})$, $\vec{\mathbf{I}} \in \mathbf{I}^*$ and $A \in \sigma(\vec{\mathbf{I}})$, if $P_{\alpha,\vec{\mathbf{I}}}(A) \neq 0$, then the conditional sequential joint probability distribution $P_{\alpha,-|\vec{\mathbf{I}}}(-|A)$ satisfies the axioms* (P2)–(P4) *in Definition 3.*

Proof. By the axiom (P2), we have

$$P_{\alpha,\vec{\mathbf{I}}_1\vec{\mathbf{I}}_2|\vec{\mathbf{I}}}(B \times C|A) = \frac{P_{\alpha,\vec{\mathbf{I}}\vec{\mathbf{I}}_1\vec{\mathbf{I}}_2}(A \times B \times C)}{P_{\alpha,\vec{\mathbf{I}}}(A)} = \frac{P_{\alpha,\vec{\mathbf{I}}\vec{\mathbf{I}}_1\vec{\mathbf{I}}'_2}(A \times B \times C)}{P_{\alpha,\vec{\mathbf{I}}}(A)} = P_{\alpha,\vec{\mathbf{I}}_1\vec{\mathbf{I}}'_2|\vec{\mathbf{I}}}(B \times C|A);$$

hence, $P_{\alpha,-|\vec{\mathbf{I}}}(-|A)$ satisfies the axiom (P2), and so does the axiom (P3). Similarly, the axiom (P4) implies that $P_{\alpha,-|\vec{\mathbf{I}}}(-|A)$ satisfies the axiom (P4); we omit the details. □

We remark that when observables are quantum ones, Bayes' formula need not hold, i.e.,

$$P_{\alpha,\vec{\mathbf{I}}_2|\vec{\mathbf{I}}_1}(B|A)P_{\alpha,\vec{\mathbf{I}}_1}(A) \neq P_{\alpha,\vec{\mathbf{I}}_1|\vec{\mathbf{I}}_2}(A|B)P_{\alpha,\vec{\mathbf{I}}_2}(B), \tag{35}$$

in general. This is because the left-hand side is $P_{\alpha,\vec{\mathbf{I}}_1\vec{\mathbf{I}}_2}(A \times B)$ and the right-hand side is $P_{\alpha,\vec{\mathbf{I}}_2\vec{\mathbf{I}}_1}(B \times A)$, so the order of the occurrences is changed. For instance, consider a qubit with the standard basis $|0\rangle$ and $|1\rangle$. Let $|x\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$. If $X_1(A) = |0\rangle\langle 0|$, $X_2(B) = |x\rangle\langle x|$, and $\alpha = |0\rangle\langle 0|$; then

$$P_{\alpha,\vec{\mathbf{I}}_1\vec{\mathbf{I}}_2}(A \times B) = \mathrm{Tr}[\alpha(X_2(B)^{\frac{1}{2}}X_1(A)^{\frac{1}{2}})^{\dagger}X_2(B)^{\frac{1}{2}}X_1(A)^{\frac{1}{2}}] = \mathrm{Tr}[\alpha X_1(A)^{\frac{1}{2}}X_2(B)X_1(A)^{\frac{1}{2}}] = \frac{1}{2},$$
$$P_{\alpha,\vec{\mathbf{I}}_2\vec{\mathbf{I}}_1}(B \times A) = \mathrm{Tr}[\alpha(X_1(A)^{\frac{1}{2}}X_2(B)^{\frac{1}{2}})^{\dagger}X_1(A)^{\frac{1}{2}}X_2(B)^{\frac{1}{2}}] = \mathrm{Tr}[\alpha X_2(B)^{\frac{1}{2}}X_1(A)X_2(B)^{\frac{1}{2}}] = \frac{1}{4},$$

and so, $P_{\alpha,\vec{\mathbf{I}}_2|\vec{\mathbf{I}}_1}(B|A)P_{\alpha,\vec{\mathbf{I}}_1}(A) \neq P_{\alpha,\vec{\mathbf{I}}_1|\vec{\mathbf{I}}_2}(A|B)P_{\alpha,\vec{\mathbf{I}}_2}(B)$. We refer to Section 4 for more details.

4. Physical Models for Generative AI

Physical models for generative AI are usually described by using systems of mean-field interacting particles, such as large language models based on attention mechanisms (cf. [7,8] and references therein); i.e., generative AI systems are regarded as classical statistical ensembles. However, since modern chips process data through controlling the flow of electric current, i.e., the dynamics of largely many electrons, they should be regarded as quantum statistical ensembles from a physical perspective (cf. [10]). Consequently, we need to model modern intelligence machines involving open quantum systems. To this end, combining the history theory of quantum systems (cf. [2]) and the theory of effect algebras (cf. [1,14]), we construct physical models realizing generative AI systems as open quantum systems.

Let $\mathbb{H}$ be a separable complex Hilbert space with the inner product $\langle -|-\rangle$ being conjugate-linear at the first variable and linear at the second variable. We denote by $\mathcal{L}(\mathbb{H})$ the set of all bounded linear operators on $\mathbb{H}$, by $\mathcal{O}(\mathbb{H})$ the set of all bounded self-adjoint operators, and by $\mathcal{P}(\mathbb{H})$ the set of all orthogonal projection operators. We denote by I the identity operator on $\mathbb{H}$. Unless stated otherwise, an operator means a bounded linear operator in the sequel. An operator T is positive if $\langle x|Tx\rangle \geq 0$ for all $x \in \mathbb{H}$, and in this case we write $T \geq 0$. We define $\mathrm{Tr}[T] = \sum_i \langle x_i|Tx_i\rangle$ for a positive operator T, where $\{x_i\}$ is an orthogonal basis of $\mathbb{H}$. It is known that $\mathrm{Tr}[T]$ is independent of the choice of the basis, and it is called the trace of T if $\mathrm{Tr}[T] < \infty$. A positive operator ρ is a density operator if $\mathrm{Tr}[\rho] = 1$, and the set of all density operators on $\mathbb{H}$ is denoted by $\mathcal{S}(\mathbb{H})$. Each positive operator is self-adjoint, and if two self-adjoint operators S, T such that $T - S \geq 0$, we write $T \geq S$ or $S \leq T$. We refer to [16–18] for more details on the theory of operators on Hilbert spaces.

A self-adjoint operator E that satisfies $0 \leq E \leq I$ is called an effect, and the set of all effects on $\mathbb{H}$ is denoted by $\mathcal{E}(\mathbb{H})$. For $E, F \in \mathcal{E}(\mathbb{H})$, we define $E \oplus F = E + F$ if $E + F \leq I$, and in this case we write $E \perp F$. It can be shown (cf. (Lemma 5.1, [1])) that $(\mathcal{E}(\mathbb{H}), 0, I, \oplus)$ is a σ-effect algebra, and each state α on $\mathcal{E}(\mathbb{H})$ has the form $\alpha(E) = \mathrm{Tr}[\rho E]$ for every $E \in \mathcal{E}(\mathbb{H})$, where ρ is a unique density operator on $\mathbb{H}$, and vice versa. Thus, we identify $\mathcal{S}(\mathcal{E}(\mathbb{H})) = \mathcal{S}(\mathbb{H})$. Let $(\Omega, \mathcal{F})$ be a measurable space. An observable $X \in \mathcal{O}(\mathcal{E}(\mathbb{H}), \Omega, \mathcal{F})$ is a positive operator valued (POV for short) measure on $(\Omega, \mathcal{F})$; i.e.,

(1) $X(F)$ is an effect on $\mathcal{E}(\mathbb{H})$ for any $F \in \mathcal{F}$;
(2) $X(\emptyset) = 0$ and $X(\Omega) = I$;
(3) For an orthogonal decomposition $\{F_j\}$ in $\mathcal{F}$,

$$X(\bigcup_j F_j) = \sum_j X(F_j), \tag{36}$$

where the series on the right-hand side is convergent in the strong operator topology on $\mathcal{L}(\mathbb{H})$, i.e.,

$$X(\bigcup_j F_j)h = \sum_j X(F_j)h, \tag{37}$$

for every $h \in \mathbb{H}$.

To understand the inference process, let us show the conventional interpretation of joint probability distributions in an open quantum system that is subject to measurements by an external observer. To this end, let $\mathcal{E}(t,s) = \{K_m(t,s)\}$ denote the time-evolution operator from time s to t, where $K_m(t,s)$'s are usually called Kraus operators, such that

$$\sum_m K_m(t,s)^\dagger K_m(t,s) = I. \tag{38}$$

That is, $\mathcal{E}(t,s)$ are quantum operations (cf. [19]) such that for every state $\rho \in \mathcal{S}(\mathbb{H})$,

$$\mathcal{E}(t,s)\rho = \sum_m K_m(t,s)\rho K_m(t,s)^\dagger, \tag{39}$$

in the Schrödinger picture, while for each observable $X \in \mathcal{O}(\mathbb{H})$,

$$\mathcal{E}(t,s)X = \sum_m K_m(t,s)^\dagger X K_m(t,s), \tag{40}$$

in the Heisenberg picture. We refer to [9] for the details on the theory of open quantum systems.

Then the density operator state $\rho(t_0)$ at time t_0 evolves in time $t_1 - t_0$ to the state $\rho(t_1)$, where

$$\rho(t_1) = \mathcal{E}(t_1,t_0)\rho(t_0) = \sum_m K_m(t_1,t_0)\rho(t_0)K_m(t_1,t_0)^\dagger. \tag{41}$$

Suppose that a measurement $(X_1, \mathbf{D}_1)$ is made at time t_1, where $X_1 \in \mathcal{O}(\mathcal{E}(\mathbb{H}), \Omega, \mathcal{F})$ and $\mathbf{D}_1 \in \mathcal{D}(\Omega, \mathcal{F})$. Then the probability that an event $X_1(A_1)$ with $A_1 \in \mathbf{D}_1$ occurs is

$$\mathrm{P}(X_1(A_1), \rho(t_1)) = \mathrm{Tr}[X_1(A_1)\rho(t_1)]. \tag{42}$$

If the result of this measurement is kept, then, according to the von Neumann–Lüders reduction postulate, the appropriate density operator to use for any further calculation is

$$\rho_{\mathrm{red}}(t_1) = \frac{X_1(A_1)^{\frac{1}{2}}\rho(t_1)X_1(A_1)^{\frac{1}{2}}}{\mathrm{Tr}[X_1(A_1)\rho(t_1)]}. \tag{43}$$

Next, suppose a measurement $(X_2, \mathbf{D}_2)$ is performed at time $t_2 > t_1$. Then, according to the above, the conditional probability of an event $X_2(A_2)$ for $A_2 \in \mathbf{D}_2$ occurs at time t_2 given that the event $X_1(A_1)$ occurs at time t_1 (and that the original state was $\rho(t_0)$) is

$$\mathrm{P}(X_2(A_2)|X_1(A_1), \rho(t_0)) = \mathrm{Tr}[X_2(A_2)\rho(t_2)], \tag{44}$$

where $\rho(t_2) = \mathcal{E}(t_2,t_0)\rho_{\mathrm{red}}(t_1)$, and the appropriate density operator to use for any further calculation is

$$\rho_{\mathrm{red}}(t_2) = \frac{X_2(A_2)^{\frac{1}{2}}\rho(t_2)X_2(A_2)^{\frac{1}{2}}}{\mathrm{Tr}[X_2(A_2)\rho(t_2)]}. \tag{45}$$

The joint probability of $X_1(A_1)$ occurring at t_1 and $X_2(A_2)$ occurring at t_2 is then

$$\begin{aligned} \mathrm{P}((X_1(A_1), X_2(A_2)), \rho(t_0)) =& \mathrm{P}(X_1(A_1), \rho(t_1))\mathrm{P}(X_2(A_2)|X_1(A_1), \rho(t_0)) \\ =& \mathrm{Tr}[X_1(A_1)\rho(t_1)]\mathrm{Tr}[X_2(A_2)\rho(t_2)]. \end{aligned} \tag{46}$$

Generalizing to a sequence of measurements $(X_1, \mathbf{D}_1), (X_2, \mathbf{D}_2), \ldots, (X_n, \mathbf{D}_2)$ at times $t_1 < t_2 < \cdots < t_n$, where $X_i \in \mathcal{O}(\mathcal{E}(\mathbb{H}), \Omega, \mathcal{F})$ and $\mathbf{D}_i \in \mathcal{D}(\Omega, \mathcal{F})$ for $i \in [n]$, the sequential joint probability of associated events $X_i(A_i)$ with $A_i \in \mathbf{D}_i$ occurring at t_i for $i \in [n]$ is

$$\begin{aligned} &\mathrm{P}((X_1(A_1), X_2(A_2), \ldots, X_n(A_n)), \rho(t_0)) \\ &\quad = \mathrm{Tr}[X_1(A_1)\rho(t_1)]\mathrm{Tr}[X_2(A_2)\rho(t_2)] \cdots \mathrm{Tr}[X_n(A_n)\rho(t_n)], \end{aligned} \tag{47}$$

where $\rho(t_i) = \mathcal{E}(t_i, t_0)\rho_{\mathrm{red}}(t_{i-1})$ for $i \in [n]$, $\rho_{\mathrm{red}}(t_0) = \rho(t_0)$, and

$$\rho_{\mathrm{red}}(t_i) = \frac{X_i(A_i)^{\frac{1}{2}}\rho(t_i)X_i(A_i)^{\frac{1}{2}}}{\mathrm{Tr}[X_i(A_i)\rho(t_i)]}. \tag{48}$$

for $i \in [n-1]$.

Therefore, given an inference set $\vec{\mathbf{I}} = ((X_1, \mathbf{D}_1), \ldots, (X_n, \mathbf{D}_n))$, for an input $\rho(t_0) \in \mathcal{S}(\mathbf{E})$, the sequential joint probability within the inference set $\vec{\mathbf{I}}$ that the event $X_1(A_1)$ occurs at t_1, $X_2(A_2)$ occurs at $t_2, \ldots, X_n(A_n)$ occurs at t_n, where $A_i \in \mathbf{D}_i$ for $i \in [n]$ and $t_0 < t_1 < t_2 < \cdots < t_n$, is given by

$$P_{\rho(t_0),\vec{\mathbf{I}}}(A_1 \times \cdots \times A_n) = \mathrm{Tr}[X_1(A_1)\rho(t_1)]\mathrm{Tr}[X_2(A_2)\rho(t_2)] \cdots \mathrm{Tr}[X_n(A_n)\rho(t_n)], \tag{49}$$

where $\rho(t_i) = \mathcal{E}(t_i, t_0)\rho_{\mathrm{red}}(t_{i-1})$ and

$$\mathcal{E}(t_i, t_0)\rho_{\mathrm{red}}(t_{i-1}) = \sum_m K_m(t_i, t_0)\rho_{\mathrm{red}}(t_{i-1})K_m(t_i, t_0)^\dagger, \tag{50}$$

for $i \in [n]$, in the Schrödinger picture operator defined with respect to the fiducial time t_0.

Proposition 2. *Let $\mathbb{H}$ be a separable complex Hilbert space, and let $(\Omega, \mathcal{F})$ be a measurable space. A physical model associated with $(\mathcal{E}(\mathbb{H}), \Omega, \mathcal{F})$ defined by*

$$\mathcal{M} = \bigcup_{n \in \mathbb{N}} \{P_{\rho,\vec{\mathbf{I}}} \in \mathcal{P}(\Omega^{(n)}, \sigma(\vec{\mathbf{I}})) : \rho \in \mathcal{S}(\mathbb{H}),\ \vec{\mathbf{I}} \in \mathbf{I}^{(n)}\}, \tag{51}$$

where $P_{\rho,\vec{\mathbf{I}}}$s′ are given by (49), satisfies the axioms in Definition 3.

Proof. For $\vec{\mathbf{I}}_1 = (X, \mathbf{D}), \vec{\mathbf{I}}_2 = (Y, \mathbf{J}) \in \mathbf{I}^{(1)}$ and $A \in \sigma(\mathbf{D}), B \in \sigma(\mathbf{J})$, by (49) we have

$$P_{\rho,\vec{\mathbf{I}}_1}(A) = \mathrm{Tr}[\rho X(A)], \quad P_{\rho,\vec{\mathbf{I}}_2}(B) = \mathrm{Tr}[\rho Y(B)].$$

If $P_{\rho,\vec{\mathbf{I}}_1}(A) = P_{\rho,\vec{\mathbf{I}}_2}(B)$ for all $\rho \in \mathcal{S}(\mathbb{H})$, then $X(A) = Y(B)$, i.e., the axiom (P1) holds.

For $\vec{\mathbf{I}} = ((X_1, \mathbf{D}_1), \ldots, (X_n, \mathbf{D}_n)) \in \mathbf{I}^*$, $\vec{\mathbf{I}}_1 = (Y, \mathbf{J}_1)$, if $A_i \in \mathbf{D}_i$ for $i \in [n]$ and $B \in \mathbf{J}_1$, by (49), we have

$$P_{\rho,\vec{\mathbf{I}}\vec{\mathbf{I}}_1}(A_1 \times \cdots \times A_n \times B) = \mathrm{Tr}[X_1(A_1)\rho(t_1)] \cdots \mathrm{Tr}[X_n(A_n)\rho(t_n)]\mathrm{Tr}[Y(B)\rho(t_{n+1})],$$

where $t_0 < t_1 < \cdots < t_n < t_{n+1}$, $\rho(t_0) = \rho$, $\rho(t_i) = \mathcal{E}(t_i, t_0)\rho_{\mathrm{red}}(t_{i-1})$, $\rho_{\mathrm{red}}(t_0) = \rho(t_0)$,

$$\rho_{\mathrm{red}}(t_i) = \frac{X_i(A_i)^{\frac{1}{2}}\rho(t_i)X_i(A_i)^{\frac{1}{2}}}{\mathrm{Tr}[X_i(A_i)\rho(t_i)]}$$

for $i \in [n]$, and

$$\rho_i(t_i) = \sum_m K_m(t_i, t_0)\rho_{\mathrm{red}}(t_{i-1})K_m(t_i, t_0)^\dagger,$$

for $i \in [n+1]$. Also, for $\vec{\mathbf{I}}_2 = (Y, \mathbf{J}_2)$ and $B \in \mathbf{J}_2$, by (49) we have

$$P_{\alpha,\vec{\mathbf{II}}_2}(A_1 \times \cdots \times A_n \times B) = \mathrm{Tr}[X_1(A_1)\rho(t_1)] \cdots \mathrm{Tr}[X_n(A_n)\rho(t_n)]\mathrm{Tr}[Y(B)\rho(t_{n+1})].$$

Hence, we have

$$P_{\rho,\vec{\mathbf{II}}_1}(A_1 \times \cdots \times A_n \times B) = P_{\alpha,\vec{\mathbf{II}}_2}(A_1 \times \cdots \times A_n \times B)$$

for $B \in \sigma(\mathbf{J}_1) \cap \sigma(\mathbf{J}_2)$. Since $\sigma(\mathbf{I})$ is generated by $\mathbf{D}_i$'s, this concludes the axiom (P2). Similarly, we can check the axioms (P3) and (P4) and omit the details. □

Remark 2. *Note that the probability family $P_{\rho(t_0),\vec{\mathbf{I}}}$'s are determined by the time-evolution operators $\mathcal{E}(t,s)$'s. Therefore, a family of the discrete-time evolution operators $\{\mathcal{E}(t_i,s)\}_{i=1}^n$ defines a physical model realizing a generative AI system, based on the mathematical formalism in Definition 3 for models of generative AI systems.*

5. Large Language Models

In this section, we describe physical models for large language models based on a transformer architecture in the Fock space over the Hilbert space of tokens. Consider a large language model $\mathfrak{S}$ with the set $\mathbf{T}$ of N tokens. A finite sequence $\{x_i\}_{i=1}^n$ of tokens is called a text for $\mathfrak{S}$, simply denoted by $T = x_1x_2\cdots x_n$ or $(x_1, x_2, \cdots, x_n)$, where n is called the length of the text T.

Let $\mathbf{h}$ be the Hilbert space with $\{|x\rangle : x \in \mathbf{T}\}$ being an orthogonal basis, and we identify $x = |x\rangle$ for $x \in \mathbf{T}$. Let $\mathbb{H} = \mathcal{F}(\mathbf{h})$ be the Fock space over $\mathbf{h}$, that is,

$$\mathcal{F}(\mathbf{h}) = \mathbb{C} \oplus \bigoplus_{n=1}^{\infty} \mathbf{h}^{\otimes n}, \tag{52}$$

where $\mathbf{h}^{\otimes n}$ is the n-fold tensor product of $\mathbf{h}$. We refer to [16] for the details of Fock spaces. In what follows, for the sake of convenience, we involve the finite Fock space

$$\mathbb{H} = \mathcal{F}^{(M)}(\mathbf{h}) = \mathbb{C} \oplus \bigoplus_{n=1}^{M} \mathbf{h}^{\otimes n}, \tag{53}$$

for a large integer $M \gg N$. Note that an operator $A^{(n)} = A_1 \otimes \cdots \otimes A_n \in \mathcal{L}(\mathbf{h}^{\otimes n})$ for $A_j \in \mathcal{L}(\mathbf{h})$ satisfies that for all $h^{(n)} = h_1 \otimes \ldots \otimes h_n \in \mathbf{h}^{\otimes n}$,

$$Ah^{(n)} = (A_1h_1) \otimes \cdots \otimes (A_nh_n) \in \mathbf{h}^{\otimes n}, \tag{54}$$

and in particular, if $\rho_i \in \mathcal{S}(\mathbf{h})$ for $i \in [n]$, then $\rho^{(n)} = \rho_1 \otimes \cdots \otimes \rho_n \in \mathcal{S}(\mathbf{h}^{\otimes n})$. Given $\alpha \in \mathbb{C}$ and a sequence $A^{(n)} \in \mathcal{L}(\mathbf{h}^{\otimes n})$ for $n \in [M]$, the operator $\mathrm{diag}(\alpha, A^{(1)}, \cdots, A^{(M)}) \in \mathcal{L}(\mathbb{H})$ is defined by

$$\mathrm{diag}(\alpha, A^{(1)}, \cdots, A^{(M)})\mathrm{h}^{(M)} = (\alpha c, A^{(1)}h^{(1)}, \cdots, A^{(M)}h^{(M)}), \tag{55}$$

for every $\mathrm{h}^{(M)} = (c, h^{(1)}, \cdots, h^{(M)}) \in \mathbb{H}$. In particular, if $\rho^{(n)} \in \mathcal{S}(\mathbf{h}^{\otimes n})$, then

$$\rho^{(M)} = \mathrm{diag}(0, 0^{(1)}, \cdots, 0^{(n-1)}, \rho^{(n)}, 0^{(n+1)}, \cdots, 0^{(M)}) \in \mathcal{S}(\mathbb{H}), \tag{56}$$

where $0^{(i)}$ denotes the zero operator in $\mathbf{h}^{\otimes i}$ for $i \geq 1$.

Since large language models are based on a transformer architecture, we suffice to construct a physical model in the Fock space $\mathbb{H} = \mathcal{F}^{(M)}(\mathbf{h})$ ($M \gg L$) for a transformer Transf_L

(24) with a composition of L blocks, consisting of L self-attention maps $\{\mathrm{SelfAtt}_\ell\}_{\ell=1}^L$ and L feed-forward neural networks $\{\mathrm{FFN}_\ell\}_{\ell=1}^L$. Precisely, let us denote the input text to the layer by $T = \{x_i\}_{i=1}^n$. As noted above,

$$\mathrm{FFN}_\ell \circ \mathrm{SelfAtt}_\ell(T) = \sum_{i=1}^{n+\ell-1} \mathrm{softmax}(S_\ell^{(n+\ell-1)})_i \mathrm{FFN}_\ell(W^{V_\ell} x_i), \tag{57}$$

where $S_\ell^{(n+\ell-1)} = \{s_i^{(\ell)}\}_{i=1}^{n+\ell-1}$ and

$$s_i^{(\ell)} = \frac{1}{\sqrt{d}} \langle W^{Q_\ell} x_{n+\ell-1}, W^{K_\ell} x_i \rangle, \quad \forall i \in [n+\ell-1]. \tag{58}$$

Then, a physical model for Transf_L consists of an input $\rho(t_0)$ and a sequence of quantum operations $\{\mathcal{E}(t_\ell, t_0)\}_{\ell=1}^L$ in the Fock space $\mathbb{H}$ defined above, where $t_0 < t_1 < \cdots < t_L$. We show how to construct this model step by step as follows.

To this end, we denote by $\Omega = \{\diamond\} \cup \mathbf{T}$ and write $\mathbf{D} = (\{\omega\} : \omega \in \Omega)$. At first, the input state ρ_T is given as

$$\rho_T = \rho(t_0) = \mathrm{diag}(0, 0^{(1)}, \cdots, 0^{(n-1)}, |x_1\rangle\langle x_1| \otimes \ldots \otimes |x_n\rangle\langle x_n|, 0^{(n+1)}, \cdots) \in \mathcal{S}(\mathbb{H}). \tag{59}$$

Then there is a physical operation $\mathcal{E}(t_1, t_0)$ in $\mathbb{H}$ (see Proposition 3 below), depending only on the attention mechanism (W_1^Q, W_1^K, W_1^V) and FFN_1, such that

$$\begin{aligned} &\mathcal{E}(t_1,t_0)\rho(t_0) \\ &= \sum_{i_1=1}^n \mathrm{softmax}(S_1^{(n)})_{i_1} \mathrm{diag}(0, 0^{(1)} \cdots, 0^{(n)}, |x_1\rangle\langle x_1| \otimes \ldots \otimes |x_n\rangle\langle x_n| \otimes |y_{i_1}^{(1)}\rangle\langle y_{i_1}^{(1)}|, 0^{(n+2)}, \cdots), \end{aligned} \tag{60}$$

where $y_{i_1}^{(1)} = \mathrm{FFN}_1(W^{V_1} x_{i_1})$ and $\{y_{i_1}^{(1)}\}_{i_1=1}^n \subset \{|x\rangle : x \in \mathbf{T}\}$. Define $X_1 : 2^\Omega \mapsto \mathcal{E}(\mathbb{H})$ by

$$X_1(\{\diamond\}) = \mathrm{diag}(1, I_{\mathbf{h}}, \ldots, I_{\mathbf{h}}^{\otimes n}, 0^{(n+1)}, I_{\mathbf{h}^{\otimes(n+2)}}, \cdots), \tag{61}$$

and for every $x \in \mathbf{T}$,

$$X_1(\{x\}) = \mathrm{diag}(0, 0^{(1)}, \ldots, 0^{(n)}, \underbrace{I_{\mathbf{h}} \otimes \cdots \otimes I_{\mathbf{h}}}_{n} \otimes |x\rangle\langle x|, 0^{(n+2)}, \cdots). \tag{62}$$

Making a measurement $(X_1, \mathbf{D})$ at time t_1, we obtain an output $y_{i_1}^{(1)}$ with probability $\mathrm{softmax}(S_1^{(n)})_{i_1}$, and the appropriate density operator to use for any further calculation is

$$\begin{aligned} \rho_{\mathrm{red}}(t_1)_{i_1} =& \frac{E_{i_1}^{(1)} \rho(t_1) E_{i_1}^{(1)}}{\mathrm{Tr}[E_{i_1}^{(1)} \rho(t_1)]} \\ =& \mathrm{diag}(0, 0^{(1)} \cdots, 0^{(n)}, |x_1\rangle\langle x_1| \otimes \ldots \otimes |x_n\rangle\langle x_n| \otimes |y_{i_1}^{(1)}\rangle\langle y_{i_1}^{(1)}|, 0^{(n+2)}, \cdots), \end{aligned} \tag{63}$$

for every $i_1 \in [n]$, where $\rho(t_1) = \mathcal{E}(t_1, t_0)\rho(t_0)$, and

$$E_{i_1}^{(1)} = \mathrm{diag}(0, 0^{(1)}, \ldots, 0^{(n)}, \underbrace{I_{\mathbf{h}} \otimes \ldots \otimes I_{\mathbf{h}}}_{n} \otimes |y_{i_1}^{(1)}\rangle\langle y_{i_1}^{(1)}|, 0^{(n+2)}, \cdots). \tag{64}$$

Next, there is a physical operation $\mathcal{E}(t_2, t_0)$ in $\mathbb{H}$ (see Proposition 3 again), depending only on the attention mechanism (W_2^Q, W_2^K, W_2^V) and FFN_2 at time t_2, such that

$$\begin{aligned}\mathcal{E}(t_2,t_0)\rho_{\mathrm{red}}(t_1)_{i_1} &= \sum_{i_2=1}^{n+1} \mathrm{softmax}(S_2^{(n+1)})_{i_2} \\ &\times \mathrm{diag}(0,0^{(1)},\cdots,0^{(n+1)},|x_1\rangle\langle x_1|\otimes\ldots\otimes|x_n\rangle\langle x_n|\otimes|y_{i_1}^{(1)}\rangle\langle y_{i_1}^{(1)}|\otimes|y_{i_2}^{(2)}\rangle\langle y_{i_2}^{(2)}|,0^{(n+3)},\cdots)\end{aligned} \tag{65}$$

for $i_1 \in [n]$, where $y_{i_2}^{(2)} = \mathrm{FFN}_2(W^{V_2}x_{i_2})$ (with $x_{n+1} = y_{i_1}^{(1)}$) and $\{y_i^{(2)}\}_{i=1}^{n+1} \subset \{|x\rangle : x \in \mathbf{T}\}$. Define $X_2 : 2^{\Omega} \mapsto \mathcal{E}(\mathbb{H})$ by

$$X_2(\{\diamond\}) = \mathrm{diag}(1, I_{\mathbf{h}}, \ldots, I_{\mathbf{h}}^{\otimes(n+1)}, 0^{(n+2)}, I_{\mathbf{h}^{\otimes(n+3)}}, \cdots), \tag{66}$$

and for every $x \in \mathbf{T}$,

$$X_2(\{x\}) = \mathrm{diag}(0, 0^{(1)}, \ldots, 0^{(n+1)}, \underbrace{I_{\mathbf{h}} \otimes \cdots \otimes I_{\mathbf{h}}}_{n+1} \otimes |x\rangle\langle x|, 0^{(n+3)}, \cdots). \tag{67}$$

Making a measurement $(X_2, \mathbf{D})$ at time t_2, we obtain an output $y_{i_2}^{(2)}$ with probability $\mathrm{softmax}(S_2^{(n+1)})_{i_2}$, and the appropriate density operator to use for any further calculation is

$$\begin{aligned}\rho_{\mathrm{red}}(t_2)_{i_1,i_2} =& \frac{E_{i_2}^{(2)}\rho(t_2)_{i_1}E_{i_2}^{(2)}}{\mathrm{Tr}[E_{i_2}^{(2)}\rho(t_2)_{i_1}]} \\ =& \mathrm{diag}(0,0^{(1)},\cdots,0^{(n+1)},|x_1\rangle\langle x_1|\otimes\ldots\otimes|x_n\rangle\langle x_n|\otimes|y_{i_1}^{(1)}\rangle\langle y_{i_1}^{(1)}|\otimes|y_{i_2}^{(2)}\rangle\langle y_{i_2}^{(2)}|,0^{(n+3)},\cdots),\end{aligned} \tag{68}$$

for each $i_2 \in [n+1]$, where $\rho(t_2)_{i_1} = \mathcal{E}(t_2,t_0)\rho_{\mathrm{red}}(t_1)_{i_1}$ and

$$E_{i_2}^{(2)} = \mathrm{diag}(0, 0^{(1)}, \ldots, 0^{(n+1)}, \underbrace{I_{\mathbf{h}} \otimes \ldots \otimes I_{\mathbf{h}}}_{n+1} \otimes |y_{i_2}^{(2)}\rangle\langle y_{i_2}^{(2)}|, 0^{(n+3)}, \cdots). \tag{69}$$

Step by step, we can obtain a physical model $\{\mathcal{E}(t_\ell,t_0)\}_{\ell=1}^{L}$ with the input state $\rho(t_0)$ such that a text $(y_{i_1}^{(1)}, y_{i_2}^{(2)}, \ldots, y_{i_L}^{(L)})$ is generated with the probability

$$\mathrm{P}_T(y_{i_1}^{(1)}, y_{i_2}^{(2)}, \ldots, y_{i_L}^{(L)}) = \mathrm{softmax}(S_1^{(n)})_{i_1} \cdots \mathrm{softmax}(S_L^{(n+L-1)})_{i_L}, \tag{70}$$

within the inference $((X_1, \mathbf{D}), \ldots, (X_L, \mathbf{D}))$.

Thus, we can obtain a physical model for Transf_L if we prove that $\mathcal{E}(t_\ell, t_0)$'s exist.

Proposition 3. *With the above notations, there exists a physical model* $\{\mathcal{E}(t_\ell,t_0)\}_{\ell=1}^{L}$ *in* $\mathbb{H} = \mathcal{F}^{(M)}(\mathbf{h})$ *(*$M \gg L$*) for a transformer* Transf_L (24) *such that given an input text* $T = \{x_i\}_{i=1}^{n}$*, a text* $(y_{i_1}^{(1)}, y_{i_2}^{(2)}, \ldots, y_{i_L}^{(L)})$ *is generated with the probability*

$$\mathrm{P}_T(y_{i_1}^{(1)}, y_{i_2}^{(2)}, \ldots, y_{i_L}^{(L)}) = \mathrm{softmax}(S_1^{(n)})_{i_1} \cdots \mathrm{softmax}(S_L^{(n+L-1)})_{i_L}, \tag{71}$$

within the inference $((X_1, \mathbf{D}), \ldots, (X_L, \mathbf{D}))$.

Proof. We regard 1, $|x\rangle\langle x|$, and $|x_1\rangle\langle x_1| \otimes \ldots \otimes |x_n\rangle\langle x_n|$ as elements in $\mathcal{L}(\mathbb{H})$ in a natural way, i.e.,

$$\begin{aligned} 1 &\simeq \mathrm{diag}(1, 0^{(1)}, 0^{(2)}, \cdots), \\ |x\rangle\langle x| &\simeq \mathrm{diag}(0, |x\rangle\langle x|, 0^{(2)}, , \cdots), \\ |x_1\rangle\langle x_1| \otimes \ldots \otimes |x_n\rangle\langle x_n| &\simeq \mathrm{diag}(0, 0^{(1)}, \cdots, 0^{(n-1)}, |x_1\rangle\langle x_1| \otimes \ldots \otimes |x_n\rangle\langle x_n|, 0^{(n+1)}, \cdots), \end{aligned}$$

for $n \geq 1$. We need to construct $\mathcal{E}(t_1, t_0)$ to satisfy (60). We first define

$$\Phi(1) = |x_0\rangle\langle x_0|,$$

where $x_0 \in \mathbf{T}$ is a certain token. Secondly, define

$$\Phi(|x\rangle\langle x|) = \mathrm{diag}(0, 0^{(1)}, |x\rangle\langle x| \otimes |\mathrm{FFN}(W^V x)\rangle\langle \mathrm{FFN}(W^V x)|, 0^{(3)}, \cdots), \quad \forall x \in \mathbf{T},$$

and in general, for $n \in [L]$ define

$$\begin{aligned} &\Phi(|x_1\rangle\langle x_1| \otimes \ldots \otimes |x_n\rangle\langle x_n|) \\ &\quad = \sum_{i=1}^{n} \mathrm{softmax}(S_1^{(n)})_i \mathrm{diag}(0, 0^{(1)} \cdots, 0^{(n)}, |x_1\rangle\langle x_1| \otimes \ldots \otimes |x_n\rangle\langle x_n| \otimes |y_i^{(1)}\rangle\langle y_i^{(1)}|, 0^{(n+2)}, \cdots), \end{aligned}$$

for any $x_i \in \mathbf{T}$ and $i \in [n]$. Let

$$\mathbf{S} = \mathrm{span}\{1, |x_1\rangle\langle x_1| \otimes \ldots \otimes |x_\ell\rangle\langle x_\ell| : \ x_i \in \mathbf{T}, i \in [\ell]; \ \ell = 1, \ldots, L\}.$$

Then Φ extends uniquely to a positive map $\mathcal{E}_\Phi$ from $\mathbf{S}$ into $\mathcal{L}(\mathbb{H})$, that is,

$$\begin{aligned} &\mathcal{E}_\Phi\Big(a_0 + \sum_{n\geq 1} \sum_{x_1, \cdots, x_n \in \mathbf{T}} a_{x_1, \cdots, x_n} |x_1\rangle\langle x_1| \otimes \ldots \otimes |x_n\rangle\langle x_n|\Big) \\ &\quad = a_0 |x_0\rangle\langle x_0| + \sum_{n\geq 1} \sum_{x_1, \cdots, x_n \in \mathbf{T}} a_{x_1, \cdots, x_n} \Phi(|x_1\rangle\langle x_1| \otimes \ldots \otimes |x_n\rangle\langle x_n|), \end{aligned}$$

where $a_0, a_{x_1, \cdots, x_n}$ are any complex numbers for $n \geq 1$. Since $\mathbf{S}$ is a commutative C^*-algebra, by Stinespring's theorem (cf. (Theorem 3.11, [20])), it follows that $\mathcal{E}_\Phi : \mathbf{S} \mapsto \mathcal{L}(\mathbb{H})$ is completely positive. Hence, by Arveson's extension theorem (cf. (Theorem 7.5, [20])), $\mathcal{E}_\Phi$ extends to a completely positive operator $\mathcal{E}(t_1, t_0)$ in $\mathcal{L}(\mathbb{H})$ (note that $\mathcal{E}(t_1, t_0)$ is not necessarily unique), i.e., a quantum operation in $\mathbb{H}$. By the construction, $\mathcal{E}(t_1, t_0)$ satisfies (60). Also, by Kraus's theorem (cf. [19]), we conclude that $\mathcal{E}(t_1, t_0)$ has the Kraus decomposition (39).

In the same way, we can prove that $\mathcal{E}(t_2, t_0)$ exists and satisfies (65). Step by step, we can thus obtain a physical model $\{\mathcal{E}(t_\ell, t_0)\}_{\ell=1}^L$ as required. □

Remark 3. *A physical model for the transformer with a multi-headed attention* (21) *can be constructed in a similar way. Also, we can construct physical models for the transformer of the form* (26)*, even for the transformer of a more complex structure (cf. [21] and reference therein). We omit the details.*

Physical models satisfying the above joint probability distributions associated with a transformer Transf_L are not necessarily unique. However, a physical model $\{\mathcal{E}(t_\ell, t_0)\}_{\ell=1}^L$ uniquely determines the joint probability distributions; that is, it defines a unique physical process for operating the large language model based on Transf_L. Therefore, in a physical model $\{\mathcal{E}(t_\ell, t_0)\}_{\ell=1}^L$ for Transf_L, training for Transf_L corresponds to training for the Kraus

operators $\mathcal{E}(t_\ell, t_0) = \{K_j^{(\ell)}(t_\ell, t_0)\}$, which are adjustable and learned during the training process, determining the physical model, as corresponding to the parameters W_ℓ^Q, W_ℓ^K and W_ℓ^V in Transf_L. From a physical perspective, to train for a large language model is just to determine the Kraus operators $\mathcal{E}(t_\ell, t_0) = \{K_j^{(\ell)}(t_\ell, t_0)\}$ associated with the corresponding physical system (cf. [22]).

Example 1. *Let* $\mathbf{T} = \{e_0, e_1\}$ *be the set of two tokens embedded in* $\mathbb{R}^2$ *such that* $e_0 = (1,0)$ *and* $e_1 = (0,1)$. *Then,* $\mathbf{h} = \mathbb{C}^2$ *with the standard basis* $|0\rangle = |e_0\rangle$ *and* $|1\rangle = |e_1\rangle$. *Let*

$$\mathbb{H} = \mathcal{F}^{(3)}(\mathbb{C}^2) = \mathbb{C} \oplus \mathbb{C}^2 \oplus [\mathbb{C}^2 \otimes \mathbb{C}^2] \oplus [\mathbb{C}^2 \otimes \mathbb{C}^2 \otimes \mathbb{C}^2].$$

Suppose that $W^Q = W^K = \text{FFN} = I$ *in* $\mathbb{R}^2$, *and let* $W^V = \sigma_x$, *i.e.,* $W^V e_0 = e_1$ *and* $W^V e_1 = e_0$. *Below, we construct a quantum operation* $\mathcal{E}$ *associated with* $\text{SelfAtt} = (I, I, \sigma)$ *and* $\text{FFN} = I$ *in* $\mathbb{R}^2$. *To this end, define* $\Phi(1) = |0\rangle\langle 0|$,

$$\begin{aligned}
\Phi(|0\rangle\langle 0|) =& |0\rangle\langle 0| \times |W^V e_0\rangle\langle W^V e_0| = |0\rangle\langle 0| \times |1\rangle\langle 1|,\\
\Phi(|1\rangle\langle 1|) =& |1\rangle\langle 1| \times |W^V e_1\rangle\langle W^V e_1| = |1\rangle\langle 1| \times |0\rangle\langle 0|;
\end{aligned}$$

and

$$\begin{aligned}
\Phi(|0\rangle\langle 0| \otimes |0\rangle\langle 0|) =& |0\rangle\langle 0| \times |0\rangle\langle 0| \times |1\rangle\langle 1|,\\
\Phi(|0\rangle\langle 0| \times |1\rangle\langle 1|) =& \frac{e}{1+e}|0\rangle\langle 0| \times |1\rangle\langle 1| \times |0\rangle\langle 0|\\
&+ \frac{1}{1+e}|0\rangle\langle 0| \times |1\rangle\langle 1| \times |1\rangle\langle 1|,\\
\Phi(|1\rangle\langle 1| \otimes |0\rangle\langle 0|) =& \frac{1}{1+e}|1\rangle\langle 1| \times |0\rangle\langle 0| \times |0\rangle\langle 0|\\
&+ \frac{e}{1+e}|1\rangle\langle 1| \times |0\rangle\langle 0| \times |1\rangle\langle 1|,\\
\Phi(|1\rangle\langle 1| \otimes |1\rangle\langle 1|) =& |1\rangle\langle 1| \times |1\rangle\langle 1| \times |0\rangle\langle 0|.
\end{aligned}$$

We regard $1, |e_i\rangle\langle e_i|, |e_j\rangle\langle e_j| \otimes |e_k\rangle\langle e_k|$ $(i,j,k = 0,1)$ *as elements in* $\mathcal{L}(\mathcal{F}^{(3)}(\mathbb{C}^2))$ *in a natural way. Let*

$$\mathbf{S} = \text{span}\{1, |e_i\rangle\langle e_i|, |e_j\rangle\langle e_j| \otimes |e_k\rangle\langle e_k| : \ i,j,k = 0,1\}.$$

Then $\mathbf{S}$ *is a subspace of* $\mathcal{L}(\mathcal{F}^{(3)}(\mathbb{C}^2))$ *and* Φ *extends uniquely to a positive map* $\mathcal{E}$ *from* $\mathbf{S}$ *into* $\mathcal{L}(\mathcal{F}^{(3)}(\mathbb{C}^2))$, *i.e.,*

$$\begin{aligned}
&\mathcal{E}\Big(a + \sum_{i=0,1} b_i |e_i\rangle\langle e_i| + \sum_{j,k=0,1} c_{j,k} |e_j\rangle\langle e_j| \otimes |e_k\rangle\langle e_k|\Big)\\
&= a|0\rangle\langle 0| + \sum_{i=0,1} b_i \Phi(|e_i\rangle\langle e_i|) + \sum_{j,k=0,1} c_{j,k} \Phi(|e_j\rangle\langle e_j| \otimes |e_k\rangle\langle e_k|),
\end{aligned}$$

for any $a, b_i, c_{j,k} \in \mathbb{C}$. *As shown in Proposition 3,* $\mathcal{E}$ *can extend to a completely positive operator in* $\mathcal{L}(\mathcal{F}^{(3)}(\mathbb{C}^2))$, *which is a quantum operation in* $\mathbb{H} = \mathcal{F}^{(3)}(\mathbb{C}^2)$ *associated with* $\text{SelfAtt} = (I, I, \sigma)$ *and* $\text{FFN} = I$ *in* $\mathbb{R}^2$. *Note that* $\mathcal{E}$ *is not necessarily unique.*

Example 2. *As in Example 1,* $\mathbf{T} = \{x_0, x_1\}$ *is the set of two tokens embedded in* $\mathbb{R}^2$ *such that* $x_0 = (1,0)$ *and* $x_1 = (0,1)$. *Then* $\mathbf{h} = \mathbb{C}^2$ *with the standard basis* $|0\rangle = |x_0\rangle$ *and* $|1\rangle = |x_1\rangle$. *Let* $\mathbb{H} = \mathcal{F}^{(6)}(\mathbb{C}^2)$. *Assume an input text* $T = (x_0, x_1, x_0)$. *The input state* ρ_T *is then given by*

$$\rho_T = \rho(t_0) = \text{diag}(0, 0^{(1)}, 0^{(2)}, |x_0\rangle\langle x_0| \otimes |x_1\rangle\langle x_1| \otimes |x_0\rangle\langle x_0|, 0^{(4)}, 0^{(5)}, 0^{(6)}).$$

If $W_1^Q = W_1^K = \mathrm{FFN}_1 = I$ *and* $W_1^V = \sigma_x$ *in* $\mathbb{R}^2$, *an associated physical operation* $\mathcal{E}(t_1, t_0)$ *at time* t_1 *satisfies*

$$\begin{aligned}\mathcal{E}(t_1,t_0)\rho(t_0) =& \frac{1}{2e+1}\mathrm{diag}(0,0^{(1)},0^{(2)},0^{(3)},|x_0\rangle\langle x_0|\otimes|x_1\rangle\langle x_1|\otimes|x_0\rangle\langle x_0|\otimes|x_0\rangle\langle x_0|,0^{(5)},0^{(6)})\\ &+\frac{2e}{2e+1}\mathrm{diag}(0,0^{(1)},0^{(2)},0^{(3)},|x_0\rangle\langle x_0|\otimes|x_1\rangle\langle x_1|\otimes|x_0\rangle\langle x_0|\otimes|x_1\rangle\langle x_1|,0^{(5)},0^{(6)}).\end{aligned}$$

By measurement, we obtain x_0 *with probability* $\frac{1}{2e+1}$ *and obtain* x_1 *with probability* $\frac{2e}{2e+1}$, *while*

$$\begin{aligned}\rho_{\mathrm{red}}(t_1)_0 =& \mathrm{diag}(0,0^{(1)},0^{(2)},0^{(3)},|x_0\rangle\langle x_0|\otimes|x_1\rangle\langle x_1|\otimes|x_0\rangle\langle x_0|\otimes|x_0\rangle\langle x_0|,0^{(5)},0^{(6)}),\\ \rho_{\mathrm{red}}(t_1)_1 =& \mathrm{diag}(0,0^{(1)},0^{(2)},0^{(3)},|x_0\rangle\langle x_0|\otimes|x_1\rangle\langle x_1|\otimes|x_0\rangle\langle x_0|\otimes|x_1\rangle\langle x_1|,0^{(5)},0^{(6)}).\end{aligned}$$

If $W_2^Q = W_2^V = \mathrm{FFN}_2 = I$ *and* $W_2^K = \sigma_x$ *in* $\mathbb{R}^2$, *an associated quantum operation* $\mathcal{E}(t_2, t_0)$ *at time* t_2 *satisfies*

$$\begin{aligned}&\mathcal{E}(t_2,t_0)\rho_{\mathrm{red}}(t_1)_0\\ &=\frac{3}{e+3}(0,0^{(1)},0^{(2)},0^{(3)},0^{(4)},|x_0\rangle\langle x_0|\otimes|x_1\rangle\langle x_1|\otimes|x_0\rangle\langle x_0|\otimes|x_0\rangle\langle x_0|\otimes|x_0\rangle\langle x_0|,0^{(6)})\\ &\quad+\frac{e}{e+3}(0,0^{(1)},0^{(2)},0^{(3)},0^{(4)},|x_0\rangle\langle x_0|\otimes|x_1\rangle\langle x_1|\otimes|x_0\rangle\langle x_0|\otimes|x_0\rangle\langle x_0|\otimes|x_1\rangle\langle x_1|,0^{(6)}),\end{aligned}$$

and

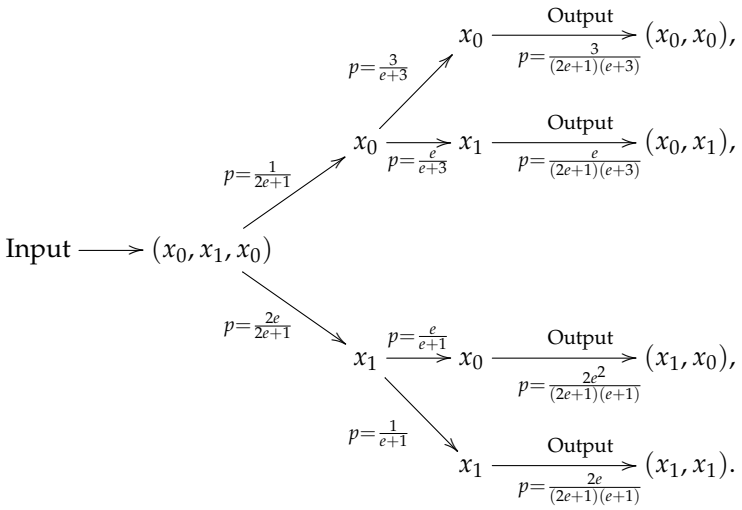
$$\begin{aligned}&\mathcal{E}(t_2,t_0)\rho_{\mathrm{red}}(t_1)_1\\ &=\frac{e}{e+1}(0,0^{(1)},0^{(2)},0^{(3)},0^{(4)},|x_0\rangle\langle x_0|\otimes|x_1\rangle\langle x_1|\otimes|x_0\rangle\langle x_0|\otimes|x_1\rangle\langle x_1|\otimes|x_0\rangle\langle x_0|,0^{(6)})\\ &\quad+\frac{1}{e+1}(0,0^{(1)},0^{(2)},0^{(3)},0^{(4)},|x_0\rangle\langle x_0|\otimes|x_1\rangle\langle x_1|\otimes|x_0\rangle\langle x_0|\otimes|x_1\rangle\langle x_1|\otimes|x_1\rangle\langle x_1|,0^{(6)}).\end{aligned}$$

By measurement at time t_2, *when* x_0 *occurs at* t_1, *we obtain* x_0 *with probability* $\frac{3}{e+3}$ *and obtain* x_1 *with probability* $\frac{e}{e+3}$; *when* x_1 *occurs at* t_1, *we obtain* x_0 *with probability* $\frac{e}{e+1}$ *and obtain* x_1 *with probability* $\frac{1}{e+1}$.

Therefore, we obtain the joint probability distributions:

$$\begin{aligned}P_T(x_0,x_0) =& \frac{1}{2e+1}\frac{3}{e+3} = \frac{3}{(2e+1)(e+3)},\\ P_T(x_0,x_1) =& \frac{1}{2e+1}\frac{e}{e+3} = \frac{e}{(2e+1)(e+3)},\\ P_T(x_1,x_0) =& \frac{2e}{2e+1}\frac{e}{e+1} = \frac{2e^2}{(2e+1)(e+1)},\\ P_T(x_1,x_1) =& \frac{2e}{2e+1}\frac{1}{e+1} = \frac{2e}{(2e+1)(e+1)}.\end{aligned}$$

This can be illustrated as follows:



6. Conclusions

Our primary innovative points are summarized as follows:

- *Mathematical formalism for generative AI models.* We present a mathematical formalism for generative AI models by using the theory of the history approach to physical systems developed by Isham and Gudder [1,2].
- *Physical models realizing generative AI systems.* We give a construction of physical models realizing generative AI systems as open quantum systems by using the theories of σ-effect algebras and open quantum systems.
- *Large language models realized as open quantum system.* We construct physical models realizing large language models based on the transformer architecture as open quantum systems in the Fock space over the Hilbert space of tokens. The Fock space structure plays a crucial role in this construction.

In conclusion, we present a mathematical formalism for generative AI and describe physical models realizing generative AI systems as open quantum systems. Our formalism shows that a transformer architecture used for generative AI systems is characterized by a family of sequential joint probability distributions. The physical models realizing generative AI systems are described by sequential event histories in open quantum systems. The Kraus operators in the physical models correspond to the query, key, and value matrices in the attention mechanism of a transformer, which are adjustable and learned during the training process. As an illustration, we construct physical models in the Fock space over the Hilbert space of tokens, realizing large language models based on a transformer architecture as open quantum systems. This means that our physical models underlie the transformer architecture for large language models. We refer to [23] for an argument on the physical principle of generic AI and to [15] for a mathematical foundation of general AI, including quantum AI.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available in this article.

Acknowledgments: The author thanks the anonymous referees for making helpful comments and suggestions, which have been incorporated into this version of the paper.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Gudder, S. A histories approach to quantum mechanics. *J. Math. Phys.* **1998**, *39*, 5772–5788. [CrossRef]
2. Isham, C.J. Quantum logic and the histories approach to quantum theory. *J. Math. Phys.* **1994**, *35*, 2157–2185. [CrossRef]
3. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
4. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
5. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A survey of large language models. *arXiv* **2023**, arXiv:2303.18223.
6. Minaee, S.; Mikolov, T.; Nikzad, N.; Chenaghlu, M.; Socher, R.; Amatriain, X.; Gao, J. Large language models: A Survey. *arXiv* **2024**, arXiv:2402.06196.
7. Geshkovski, B.; Letrouit, C.; Polyyanskiy, Y.; Rigollet, P. A mathematical perspective on transformers. *Bull. Am. Math. Soc.* 2025, *in press*.
8. Vuckovic, J.; Baratin, A.; Combes, R.T. A mathematical theory of attention. *arXiv* **2020**, arXiv:2007.02876.
9. Breuer, H.P.; Petruccione, F. *The Theory of Open Quantum Systems*; Oxford University Press: Oxford, UK, 2002.
10. Villas-Boas, C.J.; Máximo, C.E.; Paulino, P.J.; Bachelard, R.P.; Rempe, G. Bright and dark states of light: The quantum origin of classical interference. *Phys. Rev. Lett.* **2025**, *134*, 133603. [CrossRef] [PubMed]
11. Dvurečenskij, A.; Pulmannová, S. *New Trends in Quantum Structures*; Springer: Berlin/Heidelberg, Germany, 2000.
12. Petersen, P.; Zech, J. Mathematical Theory of Deep Learning. *arXiv* **2025**, arXiv:2407.18384v3.
13. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
14. Foulis, D.J.; Bennett, M.K. Effect algebras and unsharp quantum logics. *Found. Phys.* **1994**, *24*, 1331–1352. [CrossRef]
15. Chen, Z.; Ding, L.; Liu, H.; Yu, J. A topos-theoretic formalism of quantum artificial intellegence. *Sci. Sin. Math.* **2025**, *55*, 1–32. (In Chinese) [CrossRef]
16. Reed, M.; Simon, B. *Method of Mordern Mathematical Physics, Vol. I*; Academic Press: San Diego, CA, USA, 1980.
17. Reed, M.; Simon, B. *Method of Mordern Mathematical Physics, Vol. II*; Academic Press: Cambridge, UK, 1980.
18. Rudin, W. *Functional Analysis*, 2nd ed.; The McGraw-Hill Companies, Inc.: New York, NY, USA, 1991.
19. Nielsen, M.A.; Chuang, I.L. *Quantum Computation and Quantum Information*; Cambridge University Press: Cambridge, UK, 2001.
20. Paulsen, V. *Completely Bounded Maps and Operator Algebras*; Cambridge University Press: Cambridge, UK, 2002.
21. Zhang, Y.; Liu, Y.; Yuan, H.; Qin, Z.; Yuan, Y.; Gu, Q.; Yao, A.C. Tensor product attention is all you need. *arXiv* **2025**, arXiv:2501.06425.
22. Sharma, K.; Cerezo, M.; Cincio, L.; Coles, P.J. Trainability of dissipative perceptron-based quantum neural networks. *Phys. Rev. Lett.* **2022**, *128*, 180505. [CrossRef] [PubMed]
23. Chen, Z. Turing's thinking machine and 't Hooft's principle of superposition of states. *ChinaXiv* **2024**, 10.12074/202405.00127V1. [CrossRef]