# Mathematical formalism and physical models for generative artificial intelligence

**Authors:** Zeqian Chen

**Date:** 2025-05-07T10:46:16+00:00

## 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 as in [**?**, **?**]). From a physical perspective on modern chips, we then construct physical models realizing GAI systems as open quantum systems. Finally, as 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.

## Full Text

## Preamble

### MATHEMATICAL FORMALISM AND PHYSICAL MODELS FOR GENERATIVE ARTIFICIAL INTELLIGENCE

### ZEQIAN CHEN

**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 as in [**?**, **?**]). From a physical perspective on modern chips, we then construct physical models realizing GAI systems as open quantum systems. Finally, as 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.

## 1. Introduction

Generative artificial intelligence (AI) models are important for modelling intelligence machines. Generative AI is based on deep neural networks (DNNs for short), and a common characteristic of DNNs is their compositional nature (cf. [**?**]): 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. [**?**]). 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. [**?**]).

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. [**?**]).

The modularization framework of generative AI models has two origins. One is a mathematical origin that a joint probability distribution can be computed by sequentially conditional probabilities. For instance, the probability of 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

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

where the conditional probability $P_X(t_\ell|t_1,\ldots,t_{\ell-1})$ is given by the $\ell$-th attention block in the transformer. Another is a physical origin that a physical process is considered to be a sequence of events (history). That generating a text $t_{1t}2\cdots t_N$ given an input $X$ in a physical machine is a process that given an input $X$ at time $\tau_0$, an event $|t_1\rangle\langle t_1|$ occurs at time $\tau_1$, an event $|t_2\rangle\langle t_2|$ occurs at time $\tau_2$, ..., and last, an event $|t_N\rangle\langle t_N|$ occurs at time $\tau_N$. A theory of the "histories" approach to physical systems is established by Isham [**?**], and the mathematical theory of it associated with joint probability distributions is then developed by Gudder [**?**]. According to their theory, in this paper, we present a mathematical formalism for generative AI and construct 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. [**?**, **?**] and references therein), i.e., generative AI models are regarded as classical statistical systems. However, since modern chips process data through controlling the flow of electric current, i.e., the flow of many electrons, so they should be regarded as quantum statistical ensembles or open quantum systems from a physical perspective (cf. [**?**]). Consequently, according to the mathematical formalism

for generative AI, we then construct physical models realizing generative AI systems as open quantum systems. In particular, 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.

**Key words:** Generative artificial intelligence; attention mechanism; large language model; sequential joint probability; event histories; open quantum system; Kraus operator; Fock space.

## 2. Preliminaries

In this section, we present a mathematical description of attention mechanism and transformer architecture for generative AI, and include some notations and basic properties of $\sigma$-effect algebras (cf. [**?**]). 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 $\{1, \dots, n\}$. For $d \in \mathbb{N}$, we denote by $\mathbb{R}^d$ the $d$-dimensional Euclidean 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 by $S^* = \bigcup_{n\in\mathbb{N}} S^{(n)}$, where $S^{(n)}$ is the set of all sequences $(s_1, \dots, s_n)$ of $n$ elements in $S$, i.e., $S^*$ is the set of all finite sequences of elements in $S$.

### 2.1. Deep Neural Networks

Generative artificial intelligence is based on DNNs. A neural network is constructed by connecting multiple neurons, and a common characteristic of it is their compositional nature: data is processed sequentially, layer by layer, resulting in a discrete-time dynamical system.

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. Precisely, let $\{X_\ell\}_{\ell=0}^L$ be a sequence of sets where $X_\ell$ indexes the neurons in layer $\ell$, and let $\{V_\ell\}_{\ell=0}^L$ be a sequence of vector spaces. A mapping $\Phi : \mathrm{Hom}(X_0, V_0) \to \mathrm{Hom}(X_L, V_L)$ is called a feed-forward neural network of depth $L$, if there exist a sequence $\{W_\ell\}_{\ell=1}^L$ of maps $W_\ell : \mathrm{Hom}(X_{\ell-1}, V_{\ell-1}) \to \mathrm{Hom}(X_\ell, V_\ell)$ and a sequence $\{\sigma_\ell\}_{\ell=1}^{L-1}$ of maps $\sigma_\ell : V_\ell \to V_\ell$ that is called the activation function at the layer $\ell$, such that

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

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

In their most basic form, $X_\ell$ is a finite set of $n_\ell$ elements and $V_\ell = \mathbb{R}$, a feed-forward neural network $\Phi : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$ is a function of the form: the input $x^{(0)} = x \in \mathbb{R}^{n_0}$, $x^{(\ell)} = \sigma_\ell(W_\ell(x^{(\ell-1)}))$ for $\ell = 1, \dots, L-1$, and where $x^{(L)} \in \mathbb{R}^{n_L}$ is the output. This can be illustrated as follows

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

where the map $W_\ell : \mathbb{R}^{n_\ell - 1} \to \mathbb{R}^{n_\ell}$ is usually of the form

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

where $A_\ell$ is a $n_\ell \times n_{\ell-1}$-matrix called a weight matrix and $b_\ell \in \mathbb{R}^{n_\ell}$ called a bias vector for each $\ell$, and the function $\sigma_\ell : \mathbb{R}^{n_\ell} \to \mathbb{R}^{n_\ell}$ represents the activation function at the $\ell$-th layer. The set of all entries of the weight matrices and bias vectors of a neural network $\Phi$ are called the parameters of $\Phi$. 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 $\Phi$ are called the hyperparameters of $\Phi$. 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 (1) defines a specific function $\Phi$, and this function is often referred to as a model.

In a feed-forward neural network, the inputs to neurons in the $\ell$-st 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 $\ell$ may have $x^{(0)}, \dots, x^{(\ell-1)}$ as their input (and not just $x^{(\ell-1)}$). In their most basic form, $x^{(0)} = x \in \mathbb{R}^d$, $x^{(\ell)} = x^{(\ell-1)} + Q_\ell \sigma(A_\ell x^{(\ell-1)} + b_\ell)$, where $\sigma : \mathbb{R}^d \to \mathbb{R}^d$ is a vector function, $Q_\ell, A_\ell$'s are $d \times d$-matrices, and $b_\ell$'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 $x^{(\ell-1)}, x^{(\ell+1)}, \dots, x^{(L)}$ may serve as input for the neurons in layer $\ell$ and not just $x^{(\ell-1)}$. We refer to [**?**] for more details, such as training for a neural network.

### 2.2. Attention

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

**Definition 2.1.** (cf. [**?**]) Let $a : Q \times K \to \mathbb{R}$ be a function. Let $K = \{k_1, \dots, k_N\} \subset K$ be a set of keys and $V \subset V$ a set of values. Given a $q \in Q$, the attention $\mathrm{Att} : (q, K, V) \to \mathbb{R}$ is defined by

$$\mathrm{Att}(q, K, V) = \sum \mathrm{softmatch}_a(q, K)_n \cdot v_n,$$

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

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

This means that a value $v_n$ in (3) occurs with probability $\mathrm{softmatch}_a(q,K)_n$ for $n\in[N]$.

For $Q=\{q_1,\dots,q_M\}\subset Q$, we define $\mathrm{Att}(Q,K,V)=\{\mathrm{Att}(q_m,K,V)\}_{m=1}^M$. In particular, when $Q=K=V$, $\mathrm{Att}(Q,Q,Q)$ is said to be self-attention at $Q$, and the following mapping is called the self-attention map, denoted by

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

We remark that: 1. For a finite sequence $\{x_n\}_{n=1}^N$ of real numbers, define

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

Therefore, $\mathrm{softmatch}_a(q,K)_n=\mathrm{softmax}(\{a(q,k_j)\}_{j=1}^N)_n$ as usual in the literatures. 2. We have $|K|=|V|=N$, but $|Q|=M\neq N$ in general. 3. The function $a:Q\times K\to\mathbb{R}$ is called a similarity function, usually given by $a(q,k)=\langle W^Qq,W^Kk\rangle$, 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 $q\in Q$, $k\in K$, the real number $a(q,k)$ is interpreted as 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 $t\in T$ 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. [**?**]), 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 $W_i^Q$ be $d_h\times d$ (query, key, value) matrices associated with the $i$-th self-attention, and the similarity function

$$a_i(x,y)=\langle W_i^Qx,W_i^Ky\rangle.$$

Let $W^O=[W_1^O,\dots,W_{n_h}^O]$ 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

$$\mathrm{MHSelfAtt}(x_n,X,X)=\sum_{i=1}^{n_h}\sum_{j_i\in[n]}\mathrm{softmax}(\{\langle W_i^Qx_n,W_i^Kx_\ell\rangle\}_{\ell=1}^n)_{j_i}[W_i^O(W_i^Vx_{j_i})],$$

that is, an output $W_i^O(W_i^V x_{j_i})$, $j_i \in [n]$, occurs with the probability $\prod_{i=1}^{n_h} \mathrm{softmax}(\{\frac{1}{\sqrt{d_k}}\langle W_i^Q x_n, W_i^K x_\ell\rangle\}_{\ell=1}^n)_{j_i}$. As such,

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

yields a basic building block of the transformer

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

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 coined 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.

Precisely, 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

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

where $W^Q$ and $W^K$ are two $d' \times d$ matrices (i.e., the query and key matrixes), 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 $j = 1, \dots, n$. The softmax layer is given by

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

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

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

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

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

where $\mathrm{softmax}(S_i^{(n)})_{j_i} = \frac{\exp(\langle W_i^Q x_n, W_i^K x_{j_i} \rangle)}{\sum_{\ell=1}^{n} \exp(\langle W_i^Q x_n, W_i^K x_\ell \rangle)}$, $j_i \in [n]$, i.e., an output $u_n = \sum_{i=1}^{n_h} W_i^O (W_i^V x_{j_i})$ occurs with the probability $\prod_{i=1}^{n_h} \mathrm{softmax}(S_i^{(n)})_{j_i}$ 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 $y_n = \mathrm{FFN}(W^V x_j)$ $(j = 1, \ldots, n)$ with the probability $\mathrm{softmax}(S^{(n)})_j$ for each $n \in [N]$. Finally, the prediction head layer can be represented as a mapping $\mathrm{PH} : (\mathbb{R}^d)^* \to [0,1]^*$, 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

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

such that the next token $x_{N+1} = y_j = \mathrm{FFN}(W^V x_j)$ with the probability $\mathrm{softmax}(S^{(N)})_j$ 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 (HP), can be illustrated as follows: 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 a 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 HP). In the sequel, 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 $\mathrm{Transf}_L$, 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,

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

where the indices of the layers SelfAtt and FFN in (9) indicate the use of different trainable parameters in each of the block. This can be illustrated as follows:

the input text $t_1 \cdots t_n$ is embedded as $\{x_i\}_{i=1}^n$, then $\mathrm{FFN}_1 \circ \mathrm{SelfAtt}_1$ produces $y_{i_1} = t'_1$, generating $x_{n+1} = y_{i_1}$, and so on through $L$ blocks, yielding output text $\mathrm{Transf}_L(t_1 \cdots t_n) = t'_1 t'_2 \cdots t'_L$. Also, we can consider the transformer of the form

$$\mathrm{Transf}_L = ((\mathrm{id} + \mathrm{FFN}_L) \circ (\mathrm{id} + \mathrm{SelfAtt}_L)) \circ \cdots \circ ((\mathrm{id} + \mathrm{FFN}_1) \circ (\mathrm{id} + \mathrm{SelfAtt}_1))$$

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 $\sigma$-effect algebras (cf. [**?**, **?**, **?**] and references therein). Recall that an effect algebra is an algebraic system $(E, 0, 1, \oplus)$, where $E$ is a non-empty set, $0, 1 \in E$ which are called zero and unit elements of this algebra respectively, and $\oplus$ is a partial binary operation on $E$, that satisfies the following conditions for any $a, b, c \in E$:

(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 $a' \in E$ 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 $E$ 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 $a, b \in E$, we define $a \leq b$ if there exists a $c \in E$ such that $a \oplus c = b$; such $c$ 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. [**?**]) that $(E, \leq)$ 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 $E$, 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 we call the sum $\bigoplus_{i=1}^\infty a_i = \bigvee_{n=1}^\infty (a_1 \oplus \cdots \oplus a_n)$. We say that $E$ is a $\sigma$-effect algebra if $\bigoplus_{i=1}^\infty a_i$ exists for any sequence $\{a_i\}_{i=1}^\infty$ in $E$ satisfying that $a_1 \oplus \cdots \oplus a_n$ is defined for all $n \in \mathbb{N}$. It was shown in [**?**, **?**] that $E$ is a $\sigma$-effect algebra if and only if the least upper bound $\bigvee_{i=1}^\infty a_i$ exists for any monotone sequence $\{a_i\}_{i=1}^\infty$, i.e., $a_1 \leq a_2 \leq \cdots$.

Let $E$ and $F$ be $\sigma$-effect algebras. A map $\phi : E \to F$ is said to be additive if for $a, b \in E$, $a \perp b$ implies that $\phi(a) \perp \phi(b)$ and $\phi(a \oplus b) = \phi(a) \oplus \phi(b)$. An additive map $\phi : E \to F$ is $\sigma$-additive if for any sequence $\{a_i\}_{i=1}^\infty$ of pairwise orthogonal elements in $E$, $\phi(\bigoplus_{i=1}^\infty a_i) = \bigoplus_{i=1}^\infty \phi(a_i)$. A $\sigma$-additive map $\phi : E \to F$ is said to be a $\sigma$-morphism if $\phi(1) = 1$; and moreover, $\phi$ is called a $\sigma$-isomorphism if $\phi$ is a bijective $\sigma$-morphism and $\phi^{-1} : F \to E$ is a $\sigma$-morphism. It can be shown

(cf. [?]) that: 1. a map $\phi : E \to F$ is additive if and only if $\phi$ is monotone in the sense that $a \leq b$ implies $\phi(a) \leq \phi(b)$; 2. an additive map $\phi$ is $\sigma$-additive if and only if $a_1 \leq a_2 \leq \cdots$ implies $\phi(\bigvee_{i=1}^{\infty} a_i)$ exists and $\phi(\bigvee_{i=1}^{\infty} a_i) = \bigvee_{i=1}^{\infty} \phi(a_i)$; and 3. a $\sigma$-morphism $\phi$ satisfies $\phi(a') = \phi(a)'$.

The unit interval $[0, 1]$ is a $\sigma$-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 $\sigma$-effect algebra in such way.

Let $E$ be a $\sigma$-effect algebra, a $\sigma$-morphism $\phi : E \to [0, 1]$ is called a state on $E$, and we denote by $S(E)$ the set of all states on $E$. A subset $S$ of $S(E)$ 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 $\sigma$-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 $A' = \Omega \quad A$. We always regard a measurable space $(\Omega, \mathcal{F})$ as a $\sigma$-effect algebra in such way. Let $E$ be a $\sigma$-effect algebra, a $\sigma$-morphism $X : (\Omega, \mathcal{F}) \to E$ is called an observable on $E$ with valued in $(\Omega, \mathcal{F})$ (a $\Omega$-valued observable for short). The elements of a $\sigma$-effect algebra are called effects, and so an observable $X$ maps effects in $\mathcal{F}$ into effects in $E$, i.e., $X(A)$ is an effect in $E$ for $A \in \mathcal{F}$. We denote by $O(E, \Omega, \mathcal{F})$ the set of all $\Omega$-valued observables. Note that $S(\Omega, \mathcal{F})$ is equal to the set of all probability measures on $(\Omega, \mathcal{F})$. For $\alpha \in S(E)$ and $X \in O(E, \Omega, \mathcal{F})$, we have $\alpha \circ X \in S(\Omega, \mathcal{F})$, which is called the probability distribution of $X$ in the state $\alpha$.

## 3. Mathematical Formalism

In this section, we introduce a mathematical formalism for generative AI. We utilize the theory of $\sigma$-effect algebras to give a mathematical definition for a generative AI system. Let $E$ be a $\sigma$-effect algebra and $(\Omega, \mathcal{F})$ a measurable space. An orthogonal decomposition in $E$ is a sequence $\{a_i\}$ in $E$ such that $\bigoplus_i a_i = 1$. We denote by $D(E)$ the set of all completely orthogonal decomposition in $E$. A completely orthogonal decomposition in $\mathcal{F}$ is called a countable partition of $\Omega$, i.e., a sequence $\{A_i\}$ of elements in $\mathcal{F}$ such that $A_i \cap A_j = \emptyset$ for $i \neq j$ and $\cup_i A_i = \Omega$. We denote by $D(\Omega, \mathcal{F})$ the set of all countable partitions of $\Omega$. For $n \in \mathbb{N}$, an ordered $n$-tuple $\vec{R} = (e_1, \ldots, e_n)$ of effects in $E$ 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 $\tau_i$ 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, ..., and $e_n$ occurs last.

**Definition 3.1.** With the above notations, a generative artificial intelligence system $S$ is defined to be a triple $(E, \Omega, \mathcal{F})$, where $E$ is a $\sigma$-effect algebra, $(\Omega, \mathcal{F})$ is a measurable space, such that:

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

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

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

**Remark 3.1.** We refer to [**?**] for a mathematical definition of general artificial intelligence systems in terms of topos theory, including quantum artificial intelligence systems.

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

For $X \in O(E, \Omega, \mathcal{F})$ 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 $D = \{A_i\} \in D(\Omega, \mathcal{F})$, we may view $(X, D)$ as a set of possible alternative events that can occur. One interpretation is that $(X, D)$ represents a building block of an artificial intelligence architecture for processing $X$ and the alternatives result from the dial readings of the block. Given $X_i \in O(E, \Omega, \mathcal{F})$, $A_i \in \mathcal{F}$, $i = 1, \dots, n$, an ordered $n$-tuple $\vec{R} = (X_1(A_1), \dots, X_n(A_n))$ of events is called a $n$-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 $A_2$ second, ..., and $X_n$ has $a_n$ in $A_n$ last, so that the output result is $(a_1, a_2, \dots, a_n)$. We denote the set of all $n$-time chain-of-events by $\mathcal{R}^{(n)}$ and the set of all chain-of-events by $\mathcal{R}^* = \bigcup_{n=1}^{\infty} \mathcal{R}^{(n)}$.

A $n$-step inference set has the form $\vec{I} = ((X_1, D_1), \dots, (X_n, D_n))$, where $X_i \in O(E, \Omega, \mathcal{F})$, $D_i \in D(\Omega, \mathcal{F})$, $i \in [n]$. We interpret $\vec{I}$ as ordered successive processes of observables $X_i$ with partitions $D_i$, $i = 1, \dots, n$. We denote the collection of all $n$-step inference sets by $\mathcal{J}^{(n)}$ and the collection of all inference sets by $\mathcal{J}^* = \bigcup_{n=1}^{\infty} \mathcal{J}^{(n)}$. If $\vec{R} = (X_1(A_1), \dots, X_n(A_n))$ and $\vec{I} = ((X_1, D_1), \dots, (X_n, D_n))$ such that $A_i \in D_i$ for every $i \in [n]$, we say the chain-of-event $\vec{R}$ is an element of the inference set $\vec{I}$ and write $\vec{R} \in \vec{I}$. This can be illustrated as follows: the inference process takes an input $\alpha \in S(E)$, processes through $X_1(A_1), X_2(A_2), \dots, X_n(A_n)$, yielding output $(a_1, a_2, \dots, a_n)$ with probability $P_{\alpha,\vec{I}}(A_1 \times \cdots \times A_n) = P(A_1)P(A_2|A_1) \cdots P(A_n|A_{n-1}, \dots, A_1)$, where $P_{\alpha,\vec{I}}$ will be explained later.

If $\vec{I}_1 = ((X_1, D_1), \dots, (X_n, D_n))$ and $\vec{I}_2 = ((Y_1, J_1), \dots, (Y_m, J_m))$ are two inference sets, then we define their sequential product by $\vec{I} = \vec{I}_1\vec{I}_2 = ((X_1, D_1), \dots, (X_n, D_n), (Y_1, J_1), \dots, (Y_m, J_m))$ and obtain a $(n + m)$-step inference set. Mathematically, we can include the empty inference set $\emptyset$ that satisfies $\emptyset\vec{I} = \vec{I}\emptyset = \vec{I}$, such that $\mathcal{J}^*$ becomes a semigroup under this product, but we need not this property in the sequel.

For a partition $D \in D(\Omega, \mathcal{F})$, we denote by $\sigma(D)$ the $\sigma$-subalgebra of $\mathcal{F}$ generated by $D$, and for $n$ partitions $\{D_i\}_{i=1}^n$, we denote by $\sigma(\{D_i\}_{i=1}^n)$ the $\sigma$-algebra on $\Omega^{(n)}$ generated by $\{D_i\}_{i=1}^n$, i.e., $\sigma(\{D_i\}_{i=1}^n) = \sigma(\{A_1 \times \cdots \times A_n \subseteq \Omega^{(n)} : A_i \in D_i, i \in [n]\})$. We denote by $P(\Omega^{(n)}, \sigma(\{D_i\}_{i=1}^n))$ the set of all probability measures on $(\Omega^{(n)}, \sigma(\{D_i\}_{i=1}^n))$. Also, for $\vec{I} = ((X_1, D_1), \dots, (X_n, D_n))$, we write $\sigma(\vec{I}) = \sigma(\{D_i\}_{i=1}^n)$. Given an input $\alpha \in S(E)$, for an inference set $\vec{I} = ((X_1, D_1), \dots, (X_n, D_n))$ we denote by $P_{\alpha,\vec{I}} \in P(\Omega^{(n)}, \sigma(\vec{I}))$ the probability measure that for $A_1 \times \cdots \times A_n \in \sigma(\{D_i\}_{i=1}^n)$, $P_{\alpha,\vec{I}}(A_1 \times \cdots \times A_n)$ is the probability within the inference set $\vec{I}$ that the event $X_1(A_1)$ occurs first, $X_2(A_2)$ occurs second, ..., $X_n(A_n)$ occurs last. We call $P_{\alpha,\vec{I}} \in P(\Omega^{(n)}, \sigma(\vec{I}))$ the joint probability distribution of an inference set $\vec{I}$ under the input $\alpha \in S(E)$.

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

**Definition 3.2.** With the above notations, a model $\mathcal{M}$ for $S = (E, \Omega, \mathcal{F})$ is defined to be a family of joint probability distributions of inference sets

$$\{P_{\alpha,\vec{I}} \in P(\Omega^{(n)}, \sigma(\vec{I})) : \alpha \in S(E), \vec{I} \in \mathcal{I}^{(n)}\},$$

that satisfies the following axioms:

(P1) For $\vec{I}_1 = (X, D), \vec{I}_2 = (Y, J) \in \mathcal{I}^{(1)}$ and $A \in \sigma(D)$, $B \in \sigma(J)$, if $P_{\alpha,\vec{I}1}(A) = P_{\alpha,\vec{I}2}(B)$ for all $\alpha \in S(E)$, then $X(A) = Y(B)$.

(P2) For $\vec{I} \in \mathcal{I}^*$, $\vec{I}_i = (X, D_i)$, $i = 1, 2$, if $A \in \sigma(\vec{I})$ and $B \in \sigma(D_1) \cap \sigma(D_2)$, then $P_{\alpha,\vec{I}\vec{I}_1}(A \times B) = P_{\alpha,\vec{I}\vec{I}_2}(A \times B)$ for every $\alpha \in S(E)$.

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

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

(P4) If $\vec{I}_1 = ((X_1, D_1), \dots, (X_n, D_n))$, $\vec{I}_2 = ((X_1, J_1), \dots, (X_n, J_n))$, and $A_i \in D_i \cap J_i$ for $i \in [n]$,

$$P_{\alpha,\vec{I}1}(A_1 \times \cdots \times A_n) = P_{\alpha,\vec{I}2}(A_1 \times \cdots \times A_n)$$

for every $\alpha \in S(E)$.

For interpreting physical meanings for 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-event does not depend on the partitions in general and hence is unambiguous. However, for $B \in \sigma(\vec{I}_1) \cap \sigma(\vec{I}_2)$ in (P4), $P_{\alpha,\vec{I}1}(B) \neq P_{\alpha,\vec{I}2}(B)$ in general and hence is unambiguous. However, for $B \in \sigma(\vec{I}_1) \cap \sigma(\vec{I}_2)$ in (P4), $P_{\alpha,\vec{I}1}(B) \neq P_{\alpha,\vec{I}2}(B)$ if $X_i$'s are quantum observables, due to quantum interference.

If $\vec{I}_1 = ((X_1, D_1), \dots, (X_n, D_n))$ and $\vec{I}_2 = ((Y_1, J_1), \dots, (Y_m, J_m))$ are two inference sets, $A \in \sigma(\{D_i\}_{i=1}^n)$, and if $\alpha$ is an input such that $P_{\alpha,\vec{I}1}(A) \neq 0$, then we define the conditional probability of $B$ given $A$ within $\vec{I}_1\vec{I}_2$ under the input $\alpha$ as follows:

$$P_{\alpha,\vec{I}2|\vec{I}_1}(B|A) = \frac{P_{\alpha,\vec{I}1\vec{I}_2}(A \times B)}{P_{\alpha,\vec{I}1}(A)}.$$

Since $P_{\alpha,\vec{I}1\vec{I}_2}$ is a probability measure on $(\Omega^{(n+m)}, \sigma(\{D'_i\}_{i=1}^{n+m}))$, where $D'_i = D_i$ for $i \in [n]$ and $D'_{n+j} = J_j$ for $j \in [m]$, $P_{\alpha,\vec{I}2|\vec{I}_1}(-|A)$ is a probability measure on $(\Omega^m, \sigma(\{J\}_{i=1}^m))$, which is called a conditional sequential joint probability distribution.

**Proposition 3.1.** Given $\alpha \in S(E)$, $\vec{I} \in \mathcal{J}^*$ and $A \in \sigma(\vec{I})$, if $P_{\alpha,\vec{I}}(A) \neq 0$, then the conditional sequential joint probability distribution $P_{\alpha,-|\vec{I}}(-|A)$ satisfies the axioms (P2)-(P4) in Definition 3.2.

**Proof.** By the axiom (P2), we have

$$P_{\alpha,\vec{I}1\vec{I}_2|\vec{I}}(B\times C|A) = \frac{P_{\alpha,\vec{I}\vec{I}_1\vec{I}_2}(A \times B \times C)}{P_{\alpha,\vec{I}}(A)} = \frac{P_{\alpha,\vec{I}\vec{I}_1\vec{I}'_2}(A \times B \times C)}{P_{\alpha,\vec{I}}(A)} = P_{\alpha,\vec{I}2'|\vec{I}}(B\times C|A),$$

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

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

$$P_{\alpha,\vec{I}2|\vec{I}_1}(B|A)P_{\alpha,\vec{I}1}(A) \neq P_{\alpha,\vec{I}1|\vec{I}_2}(A|B)P_{\alpha,\vec{I}2}(B).$$

This is because that the left-hand side is $P_{\alpha,\vec{I}1\vec{I}_2}(A \times B)$, 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

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

and the right-hand side is $P_{\alpha,\vec{I}2\vec{I}_1}(B\times A) = \mathrm{Tr}[\alpha(X_1(A)^{1/2}X_2(B)^{1/2})^\dagger X_1(A)^{1/2}X_2(B)^{1/2}] = \mathrm{Tr}[\alpha X_2(B)^{1/2}X_1(A)X_2(B)^{1/2}]$, and so, $P_{\alpha,\vec{I}1\vec{I}_2}(A \times B) \neq P_{\alpha,\vec{I}2\vec{I}_1}(B \times A)$. 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 mechanism (cf. [**?**, **?**] 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 flow of many

electrons, so they should be regarded as quantum statistical ensembles from a physical perspective. Consequently, we need to model modern intelligence machines involving open quantum systems. To this end, combining the history theory of quantum systems (cf. [**?**]) and the theory of effect algebras (cf. [**?**, **?**]), we present physical models realizing generative AI systems as open quantum systems.

Let $H$ be a separable complex Hilbert space with the inner product $(-,-)$ being conjugate linear at the first variable and linear at second variable. We denote by $L(H)$ the set of all bounded linear operators on $H$, by $O(H)$ the set of all bounded self-adjoint operators, and by $P(H)$ the set of all orthogonal projection operators. We denote by $I$ the identity operator on $H$. Without stated otherwise, an operator means a bounded linear operator in the sequel. An operator $T$ is positive if $(Tx, x) \geq 0$ for all $x \in H$, and in this case we write $T \geq 0$. We define $\mathrm{Tr}[T] = \sum_i (x_i, Tx_i)$ for a positive operator $T$, where $\{x_i\}$ is an orthogonal basis of $H$. 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 $\rho$ is a density operator if $\mathrm{Tr}[\rho] = 1$, and the set of all density operators on $H$ is denoted by $S(H)$. 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 [**?**, **?**, **?**] for more details on the theory of operator 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 $H$ is denote by $\mathcal{E}(H)$. For $E, F \in \mathcal{E}(H)$, 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. [**?**, **?**]) that $(\mathcal{E}(H), 0, I, \oplus)$ is a $\sigma$-effect algebra, and each state $\alpha$ on $\mathcal{E}(H)$ has the form $\alpha(E) = \mathrm{Tr}[\rho E]$ for every $E \in \mathcal{E}(H)$, where $\rho$ is a unique density operator on $H$, and vice versa. Thus, we identity $S(\mathcal{E}(H)) = S(H)$. Let $(\Omega, \mathcal{F})$ be a measurable space. An observable $X \in O(\mathcal{E}(H), \Omega, \mathcal{F})$ is a positive operator valued (POV for short) measure on $(\Omega, \mathcal{F})$, i.e.: 1. $X(F)$ is an effect on $\mathcal{E}(H)$ for any $F \in \mathcal{F}$; 2. $X(\emptyset) = 0$ and $X(\Omega) = I$; 3. for an orthogonal decomposition $\{F_j\}$ in $\mathcal{F}$, $X(\bigcup_j F_j) = \sum_j X(F_j)$, where the series on the right hand is convergent in the strong operator topology on $L(H)$, i.e., $X(\bigcup_j F_j)h = \sum_j X(F_j)h$, $\forall h \in H$.

To understand the inference process, let us exhibit 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 $E(t, s) = \{K_m(t, s)\}$ denote the time-evolution operator from time $s$ to $t$, where $K_m(t, s)$'s are usually called the Kraus operators, such that

$$\sum_m K_m(t, s)^\dagger K_m(t, s) = I.$$

That is, $E(t, s)$ are quantum operations (cf. [**?**]) such that for every state $\rho \in S(H)$,

$$E(t, s)\rho = \sum_m K_m(t, s)\rho K_m(t, s)^\dagger,$$

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

$$E(t,s)X = \sum_m K_m(t,s)^\dagger X K_m(t,s),$$

in the Heisenberg picture. We refer to [**?**] 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

$$\rho(t_1) = E(t_1,t_0)\rho(t_0) = \sum_m K_m(t_1,t_0)\rho(t_0)K_m(t_1,t_0)^\dagger.$$

Suppose that a measurement $(X_1, D_1)$ is made at time $t_1$, where $X_1 \in O(\mathcal{E}(H), \Omega, \mathcal{F})$ and $D_1 \in D(\Omega, \mathcal{F})$. Then the probability that an event $X_1(A_1)$ with $A_1 \in D_1$ occurs is

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

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

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

Next suppose a measurement $(X_2, D_2)$ is performed at time $t_2 > t_1$. Then according to the above, the conditional probability of an event $X_2(A_2)$ for $A_2 \in D_2$ 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

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

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

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

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

$$P((X_1(A_1), X_2(A_2)), \rho(t_0)) = P(X_1(A_1), \rho(t_1))P(X_2(A_2)|X_1(A_1), \rho(t_0)) = \mathrm{Tr}[X_1(A_1)\rho(t_1)]\,\mathrm{Tr}[X_2(A_2)\rho(t_2)].$$

Generalizing to a sequence of measurements $(X_1, D_1), (X_2, D_2), \ldots, (X_n, D_n)$ at times $t_1 < t_2 < \cdots < t_n$, where $X_i \in O(\mathcal{E}(H), \Omega, \mathcal{F})$ and $D_i \in D(\Omega, \mathcal{F})$ for

$i \in [n]$, the sequential joint probability of associated events $X_i(A_i)$ with $A_i \in D_i$ occurring at $t_i$ for $i \in [n]$ is

$$P((X_1(A_1), X_2(A_2), \dots, X_n(A_n)), \rho(t_0)) = \mathrm{Tr}[X_1(A_1)\rho(t_1)]\,\mathrm{Tr}[X_2(A_2)\rho(t_2)]\cdots\mathrm{Tr}[X_n(A_n)\rho(t_n)],$$

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

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

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

Therefore, given an inference set $\vec{I} = ((X_1, D_1), \dots, (X_n, D_n))$, for an input $\rho(t_0) \in S(E)$, the sequential joint probability within the inference set $\vec{I}$ that the event $X_1(A_1)$ occurs at $t_1$, $X_2(A_2)$ occurs at $t_2$, ..., $X_n(A_n)$ occurs at $t_n$, where $A_i \in D_i$ for $i \in [n]$ and $t_0 < t_1 < t_2 < \cdots < t_n$, is given by

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

where $\rho(t_i) = E(t_i, t_0)\rho_{\mathrm{red}}(t_{i-1})$ for $i \in [n]$, and $E(t_i, t_0)\rho_{\mathrm{red}}(t_{i-1}) = \sum_m K_m(t_i, t_0)\rho_{\mathrm{red}}(t_{i-1})K_m(t_i, t_0)^\dagger$ in the Schrödinger picture operator defined with respect to the fiducial time $t_0$.

**Proposition 4.1.** Let $H$ be a separable complex Hilbert space and let $(\Omega, \mathcal{F})$ be a measurable space. A physical model associated with $(\mathcal{E}(H), \Omega, \mathcal{F})$ defined by

$$\{P_{\rho,\vec{I}} \in P(\Omega^{(n)}, \sigma(\vec{I})) : \rho \in S(H), \vec{I} \in \mathcal{I}^{(n)}\},$$

where $P_{\rho,\vec{I}}$'s are given by (14), satisfies the axioms in Definition 3.2.

**Proof.** For $\vec{I}_1 = (X, D), \vec{I}_2 = (Y, J) \in \mathcal{I}^{(1)}$ and $A \in \sigma(D)$, $B \in \sigma(J)$, by (14) we have

$$P_{\rho,\vec{I}1}(A) = \mathrm{Tr}[\rho X(A)], \quad P_{\rho,\vec{I}2}(B) = \mathrm{Tr}[\rho Y(B)].$$

If $P_{\rho,\vec{I}1}(A) = P_{\rho,\vec{I}2}(B)$ for all $\rho \in S(H)$, then $X(A) = Y(B)$, i.e., the axiom (P1) holds.

For $\vec{I} = ((X_1, D_1), \dots, (X_n, D_n)) \in \mathcal{I}^*$, $\vec{I}_1 = (Y, J_1)$, if $A_i \in D_i$ for $i \in [n]$ and $B \in J_1$, by (14) we have

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

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

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

for $i \in [n-1]$. Also, for $\vec{I}_2 = (Y, J_2)$ and $B \in J_2$, by (14) we have

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

Hence, we have $P_{\rho,\vec{I}\vec{I}_1}(A_1 \times \cdots \times A_n \times B) = P_{\alpha,\vec{I}\vec{I}_2}(A_1 \times \cdots \times A_n \times B)$ for $B \in \sigma(J_1) \cap \sigma(J_2)$. Since $\sigma(I)$ is generated by $D_i$'s, this concludes the axiom (P2). Similarly, we can check the axioms (P3) and (P4) and omit the details.

Note that the probability family $P_{\rho(t_0),\vec{I}}$'s are determined by the time-evolution operator $E(t,s)$ consisting of Kraus operators, which define physical models for generative AI systems.

## 5. Large Language Models

In this section, we construct physical models in the Fock space over the Hilbert space of tokens for large language models based on a transformer architecture. Consider a large language model $S$ with the set $T$ of $N$ tokens. A finite sequence $\{x_i\}_{i=1}^n$ of tokens is called a text for $S$, 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 $h$ be the Hilbert space with $\{|x\rangle : x \in T\}$ being an orthogonal basis, and we identity $x = |x\rangle$ for $x \in T$. Let $H = \mathcal{F}(h)$ be the Fock space over $h$, that is,

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

where $h^{\otimes n}$ is the $n$-fold tensor product of $h$. We refer to [**?**] for the details of Fock spaces. In what follows, for the sake of convenience, we involve the finite Fock space $H = \mathcal{F}^{(M)}(h) = \mathbb{C} \oplus \bigoplus_{n=1}^{M} h^{\otimes n}$ for a large integer $M \gg N$. Note that an operator $A^{(n)} = A_1 \otimes \cdots \otimes A_n \in L(h^{\otimes n})$ for $A_j \in L(h)$ satisfies that for all $h^{(n)} = h_1 \otimes ... \otimes h_n \in h^{\otimes n}$, $Ah^{(n)} = (A_{1h}1) \otimes \cdots \otimes (A_{nh}n) \in h^{\otimes n}$, and in particular, if $\rho_i \in S(h)$ for $i \in [n]$, then $\rho^{(n)} = \rho_1 \otimes \cdots \otimes \rho_n \in S(h^{\otimes n})$. Given $\alpha \in \mathbb{C}$ and a sequence $A^{(n)} \in L(h^{\otimes n})$ for $n \in [M]$, the operator $\mathrm{diag}(\alpha, A^{(1)}, \cdots, A^{(M)}) \in L(H)$ is defined by

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

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

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

where $0^{(i)}$ denotes the zero operator in $h^{\otimes i}$ 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 $H = \mathcal{F}^{(M)}(h)$ $(M \gg L)$ for a transformer $\mathrm{Transf}_L$ (9) 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 $T = \{x_i\}_{i=1}^n$ and

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

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

$$s_i^{(\ell)} = \langle W_\ell^Q x_{n+\ell-1}, W_\ell^K x_i \rangle, \quad \forall i \in [n+\ell-1].$$

Then, a physical model for $\mathrm{Transf}_L$ consist of an input $\rho(t_0)$ and a sequence of quantum operations $\{E(t_\ell, t_0)\}_{\ell=1}^L$ in the Fock space $H$ 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 $\Omega = \{\natural\} \cup T$ and write $D = (\{\omega\} : \omega \in \Omega)$. At first, the input state $\rho_T$ is defined by

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

Then there is a physical operation $E(t_1, t_0)$ in $H$ depending only on the attention mechanism ($W^Q$ and $W^K$) and $\mathrm{FFN}_1$ such that

$$E(t_1, t_0)\rho(t_0) = \sum_{i=1}^n \mathrm{softmax}(S^{(n)})_i \, \mathrm{diag}(0, 0^{(1)} \cdots, 0^{(n)}, |x_1\rangle\langle x_1| \otimes \ldots \otimes |x_n\rangle\langle x_n| \otimes |y_i^{(1)}\rangle\langle y_i^{(1)}|, 0^{(n+2)}, \cdots),$$

where $y_i^{(1)} = \mathrm{FFN}_1(W_1^V x_i)$ and $\{y_i^{(1)}\}_{i=1}^n \subset \{|x\rangle : x \in T\}$. Define $X_1 : 2^\Omega \to \mathcal{E}(H)$ by

$$X_1(\{\natural\}) = \mathrm{diag}(1, I_h, \ldots, I_h^{\otimes n}, 0^{(n+1)}, I_h^{\otimes(n+2)}, \cdots),$$

and for every $x \in T$,

$$X_1(\{x\}) = \mathrm{diag}(0, 0^{(1)}, \ldots, 0^{(n)}, \underbrace{I_h \otimes \cdots \otimes I_h}_{n \text{ times}} \otimes |x\rangle\langle x|, 0^{(n+2)}, \cdots).$$

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

$$\rho_{\mathrm{red}}(t_1)_i = \frac{E_i^{(1)}\rho(t_1)E_i^{(1)}}{\mathrm{Tr}[E_i^{(1)}\rho(t_1)]} = \mathrm{diag}(0, 0^{(1)} \cdots, 0^{(n)}, |x_1\rangle\langle x_1| \otimes \ldots \otimes |x_n\rangle\langle x_n| \otimes |y_i^{(1)}\rangle\langle y_i^{(1)}|, 0^{(n+2)}, \cdots),$$

for every $i \in [n]$, where $\rho(t_1) = E(t_1, t_0)\rho(t_0)$, and

$$E_i^{(1)} = \mathrm{diag}(0, 0^{(1)}, \ldots, 0^{(n)}, \underbrace{I_h \otimes \ldots \otimes I_h}_{n \text{ times}} \otimes |y_i^{(1)}\rangle\langle y_i^{(1)}|, 0^{(n+2)}, \cdots).$$

Next, there is a physical operation $E(t_2, t_0)$ in $H$ depending only on the attention mechanism ($W^Q$ and $W^K$) and $\mathrm{FFN}_2$ at time $t_2$ such that

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

for $i_1 \in [n]$, where $y^{(2)}_{i_2} = \mathrm{FFN}_2(W_2^V x_{i_2})$ (with $x_{n+1} = y^{(1)}_{i_1}$) and $\{y^{(2)}_{i_2}\}_{i_2=1}^{n+1} \subset \{|x\rangle : x \in T\}$. Define $X_2 : 2^\Omega \to \mathcal{E}(H)$ by

$$X_2(\{\natural\}) = \mathrm{diag}(1, I_h, \dots, I_h^{\otimes(n+1)}, 0^{(n+2)}, I_h^{\otimes(n+3)}, \cdots),$$

and for every $x \in T$,

$$X_2(\{x\}) = \mathrm{diag}(0, 0^{(1)}, \dots, 0^{(n+1)}, \underbrace{I_h \otimes \cdots \otimes I_h}_{(n+1)\text{ times}} \otimes |x\rangle\langle x|, 0^{(n+3)}, \cdots).$$

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

$$\rho_{\mathrm{red}}(t_2)_{i_1,i_2} = \frac{E^{(2)}_{i_2}\rho(t_2)_{i_1}E^{(2)}_{i_2}}{\mathrm{Tr}[E^{(2)}_{i_2}\rho(t_2)_{i_1}]} = \mathrm{diag}(0, 0^{(1)}, \cdots, 0^{(n+1)}, |x_1\rangle\langle x_1|\otimes\dots\otimes|x_n\rangle\langle x_n|\otimes|y^{(1)}_{i_1}\rangle\langle y^{(1)}_{i_1}|\otimes|y^{(2)}_{i_2}\rangle\langle y^{(2)}_{i_2}|, 0^{(n+3)}, \cdots),$$

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

$$E^{(2)}_{i_2} = \mathrm{diag}(0, 0^{(1)}, \dots, 0^{(n+1)}, \underbrace{I_h \otimes \dots \otimes I_h}_{(n+1)\text{ times}} \otimes |y^{(2)}_{i_2}\rangle\langle y^{(2)}_{i_2}|, 0^{(n+3)}, \cdots).$$

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

$$P_T(y^{(1)}_{i_1}, \dots, y^{(L)}_{i_L}) = \mathrm{softmax}(S^{(n)})_{i_1} \cdots \mathrm{softmax}(S_L^{(n+L-1)})_{i_L}$$

within the inference $((X_1, D), \dots, (X_L, D))$. Thus, we can construct a physical model for $\mathrm{Transf}_L$ if we prove that $E(t_\ell, t_0)$'s exist.

**Proposition 5.1.** With the above notations, there exists a physical model $\{E(t_\ell, t_0)\}_{\ell=1}^L$ in $H = \mathcal{F}^{(M)}(h)$ $(M \gg L)$ for a transformer $\mathrm{Transf}_L$ (9) such that given an input text $T = \{x_i\}_{i=1}^n$, a text $(y^{(1)}_{i_1}, \dots, y^{(L)}_{i_L})$ is generated with the probability $P_T(y^{(1)}_{i_1}, \dots, y^{(L)}_{i_L}) = \mathrm{softmax}(S^{(n)})_{i_1} \cdots \mathrm{softmax}(S_L^{(n+L-1)})_{i_L}$ within the inference $((X_1, D), \dots, (X_L, D))$.

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

$$1 \cong \mathrm{diag}(1, 0^{(1)}, 0^{(2)}, \cdots), \quad |x\rangle\langle x| \cong \mathrm{diag}(0, |x\rangle\langle x|, 0^{(2)}, \cdots),$$

$$|x_1\rangle\langle x_1|\otimes\dots\otimes|x_n\rangle\langle x_n| \cong \mathrm{diag}(0, 0^{(1)}, \cdots, 0^{(n-1)}, |x_1\rangle\langle x_1|\otimes\dots\otimes|x_n\rangle\langle x_n|, 0^{(n+1)}, \cdots),$$

for $n \geq 1$. We need to construct $E(t_1, t_0)$ to satisfy (16). We first define

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

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

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

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

$$\Phi(|x_1\rangle\langle x_1| \otimes ... \otimes |x_n\rangle\langle x_n|) = \sum_{i=1}^{n} \mathrm{softmax}(S^{(n)})_i \,\mathrm{diag}(0, 0^{(1)} \cdots, 0^{(n)}, |x_1\rangle\langle x_1| \otimes ... \otimes |x_n\rangle\langle x_n| \otimes |y_i^{(1)}\rangle\langle y_i^{(1)}|, 0^{(n+2)}, \cdots)$$

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

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

Then $\Phi$ extends uniquely to a positive map $E_\Phi$ from $S$ into $L(H)$, that is,

$$E_\Phi\left(a_0 1 + \sum_{x\in T} a_x |x\rangle\langle x| + \cdots + \sum_{x_1,\ldots,x_n \in T} a_{x_1,\ldots,x_n} |x_1\rangle\langle x_1| \otimes ... \otimes |x_n\rangle\langle x_n| + \cdots\right)$$

$$= a_0 |x_0\rangle\langle x_0| + \sum_{x\in T} a_x \Phi(|x\rangle\langle x|) + \cdots + \sum_{x_1,\ldots,x_n \in T} a_{x_1,\ldots,x_n} \Phi(|x_1\rangle\langle x_1| \otimes ... \otimes |x_n\rangle\langle x_n|) + \cdots,$$

where $a_0, a_x, a_{x_1,\ldots,x_n}$ are any complex numbers for $n \geq 1$. Since $S$ is a commutative $C^*$-algebra, by Stinespring's theorem (cf. [**?**, **?**]), we follows that $E_\Phi : S \to L(H)$ is completely positive.

Hence, by Arveson's extension theorem (cf. [**?**, **?**]), $E_\Phi$ extends to a completely positive operator $E(t_1, t_0)$ in $L(H)$ (note that $E(t_1, t_0)$ is not necessarily unique), i.e., a quantum operation in $H$. By the construction, $E(t_1, t_0)$ satisfies (16). Also, by Kraus's theorem (cf. [**?**]) we conclude that $E(t_1, t_0)$ has the Kraus decomposition (12).

By the same way, we can prove that $E(t_2, t_0)$ exists and satisfies (17). Step by step, we thus can obtain a physical model $\{E(t_\ell, t_0)\}_{\ell=1}^L$ as required.

**Remark 5.1.** A physical model for the transformer with a multi-headed attention (7) can be constructed in a similar way. Also, we can construct physical models for the transformer of the form (10), even for the transformer of more complex structure (cf. [**?**] and reference therein). We omit the details.

Physical models satisfying the above joint probability distributions associated with a transformer $\mathrm{Transf}_L$ are not necessarily unique. However, a physical model $\{E(t_\ell, t_0)\}_{\ell=1}^L$ uniquely determines the joint probability distributions, that is, it defines a unique physical process for operating the large language model based on $\mathrm{Transf}_L$. Therefore, in a physical model $\{E(t_\ell, t_0)\}_{\ell=1}^L$ for $\mathrm{Transf}_L$, training for $\mathrm{Transf}_L$ corresponds to training for the Kraus operators $E(t_\ell, t_0) = \{K_j^{(\ell)}(t_\ell, t_0)\}$, which are adjustable and learned during the training process, determining the physical model, as corresponding to the parameters $W_\ell^Q$, $W_\ell^K$ and $W_\ell^V$ in $\mathrm{Transf}_L$. From a physical perspective, training for

a large language model is just to determine the Kraus operators $E(t_\ell, t_0) = \{K_j^{(\ell)}(t_\ell, t_0)\}$ associated with the corresponding physical system (cf. [**?**]).

**Example 5.1.** Let $T = \{e_0, e_1\}$ be the set of two tokens embedded in $\mathbb{R}^2$ such that $e_0 = (1, 0)$ and $e_1 = (0, 1)$. Then $h = \mathbb{C}^2$ with the standard basis $|0\rangle = |e_0\rangle$ and $|1\rangle = |e_1\rangle$. Let $H = \mathcal{F}^{(3)}(\mathbb{C}^2) = \mathbb{C} \oplus \mathbb{C}^2 \oplus [\mathbb{C}^2 \otimes \mathbb{C}^2] \oplus [\mathbb{C}^2 \otimes \mathbb{C}^2 \otimes \mathbb{C}^2]$.

Suppose that $W^Q = W^K = \mathrm{FFN} = I$ in $\mathbb{R}^2$ and let $W^V = \sigma_x$, i.e., $W^V e_0 = e_1$ and $W^V e_1 = e_0$. In the sequel, we construct a quantum operation $E$ associated with $\mathrm{SelfAtt} = (I, I, \sigma)$ and $\mathrm{FFN} = I$ in $\mathbb{R}^2$. To this end, define

$$\Phi(1) = |0\rangle\langle 0|,$$

$$\Phi(|0\rangle\langle 0|) = |0\rangle\langle 0| \times |W^V e_0\rangle\langle W^V e_0| = |0\rangle\langle 0| \times |1\rangle\langle 1|,$$

$$\Phi(|1\rangle\langle 1|) = |1\rangle\langle 1| \times |W^V e_1\rangle\langle W^V e_1| = |1\rangle\langle 1| \times |0\rangle\langle 0|;$$

$$\Phi(|0\rangle\langle 0| \otimes |0\rangle\langle 0|) = |0\rangle\langle 0| \times |0\rangle\langle 0| \times |1\rangle\langle 1|,$$

$$\Phi(|0\rangle\langle 0| \times |1\rangle\langle 1|) = \Phi(|1\rangle\langle 1| \otimes |0\rangle\langle 0|) = \frac{1}{1+e}|0\rangle\langle 0| \times |1\rangle\langle 1| \times |0\rangle\langle 0| + \frac{e}{1+e}|0\rangle\langle 0| \times |1\rangle\langle 1| \times |1\rangle\langle 1|,$$

$$\Phi(|1\rangle\langle 1| \otimes |1\rangle\langle 1|) = |1\rangle\langle 1| \times |1\rangle\langle 1| \times |0\rangle\langle 0|.$$

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

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

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

$$E\left(a|0\rangle\langle 0| + \sum_{i=0,1} b_i |e_i\rangle\langle e_i| + \sum_{j,k=0,1} c_{j,k} |e_j\rangle\langle e_j| \otimes |e_k\rangle\langle e_k|\right)$$

$$= a|0\rangle\langle 0| + \sum_{i=0,1} b_i \Phi(|e_i\rangle\langle e_i|) + \sum_{j,k=0,1} c_{j,k} \Phi(|e_j\rangle\langle e_j| \otimes |e_k\rangle\langle e_k|),$$

for any $a, b_i, c_{j,k} \in \mathbb{C}$. As shown in Proposition 5.1, $E$ can extend to a completely positive operator in $L(\mathcal{F}^{(3)}(\mathbb{C}^2))$, which is a quantum operation in $H = \mathcal{F}^{(3)}(\mathbb{C}^2)$ associated with $\mathrm{SelfAtt} = (I, I, \sigma)$ and $\mathrm{FFN} = I$ in $\mathbb{R}^2$. Note that $E$ is not necessarily unique.

**Example 5.2.** As in Example 5.1, $T = \{x_0, x_1\}$ is the set of two tokens embedded in $\mathbb{R}^2$ such that $x_0 = (1, 0)$ and $x_1 = (0, 1)$. Then $h = \mathbb{C}^2$ with the standard basis $|0\rangle = |x_0\rangle$ and $|1\rangle = |x_1\rangle$. Let $H = \mathcal{F}^{(6)}(\mathbb{C}^2)$. Assume an input text $T = (x_0, x_1, x_0)$. The input state $\rho_T$ is then given by

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

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

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

$$+\frac{2e}{2e+1}\,\mathrm{diag}(0, 0^{(1)}, 0^{(2)}, 0^{(3)}, |x_0\rangle\langle x_0|\otimes|x_1\rangle\langle x_1|\otimes|x_0\rangle\langle x_0|\otimes|x_1\rangle\langle x_1|, 0^{(5)}, 0^{(6)}).$$

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

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

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

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

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

$$+\frac{e}{e+3}\,\mathrm{diag}(0, 0^{(1)}, 0^{(2)}, 0^{(3)}, 0^{(4)}, |x_0\rangle\langle x_0|\otimes|x_1\rangle\langle x_1|\otimes|x_0\rangle\langle x_0|\otimes|x_0\rangle\langle x_0|\otimes|x_1\rangle\langle x_1|, 0^{(6)}).$$

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

$$+\frac{e}{e+1}\,\mathrm{diag}(0, 0^{(1)}, 0^{(2)}, 0^{(3)}, 0^{(4)}, |x_0\rangle\langle x_0|\otimes|x_1\rangle\langle x_1|\otimes|x_0\rangle\langle x_0|\otimes|x_1\rangle\langle x_1|\otimes|x_1\rangle\langle x_1|, 0^{(6)}).$$

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

Therefore, we obtain the joint probability distributions:

$$P_T(x_0,x_0) = \frac{1}{(2e+1)(e+3)}, \quad P_T(x_0,x_1) = \frac{e}{(2e+1)(e+3)},$$

$$P_T(x_1,x_0) = \frac{1}{(2e+1)(e+1)}, \quad P_T(x_1,x_1) = \frac{e}{(2e+1)(e+1)}.$$

## 6. Conclusion

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 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 [**?**] for an argument on the physical principle of generic AI and to [**?**] for a mathematical foundation of general AI, including quantum AI.

## References

[1] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv: 1409.0473, 2014.

[2] H.P. Breuer, F. Petruccione, The Theory of Open Quantum Systems, Oxford University Press, Oxford, 2002.

[3] Z. Chen, Turing's thinking machine and 't Hooft's principle of superposition of states, ChinaXiv: 202405.00127.

[4] Z. Chen, L. Ding, H. Liu, J. Yu, A topos-theoretic formalism of quantum artificial intelligence (in Chinese), Scientia Sinica Mathematica 55 (2025), online: www.sciengine.com/SSM/doi/10.1360/SSM-2024-0126.

[5] A. Dvurečenskij, S. Pulmannová, New Trends in Quantum Structures, Springer, Berlin, 2000.

[6] D.J. Foulis, M.K. Bennett, Effect algebras and unsharp quantum logics, Foundations of Physics 24 (1994), 1331-1352.

[7] B. Geshkovski, C. Letrouit, Y. Polyyanskiy, P. Rigollet, A mathematical perspective on transformers, Bulletin of the American Mathematical Society, 2025, in press.

[8] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.

[9] S. Gudder, A histories approach to quantum mechanics, Journal of Mathematical Physics 39 (1998), 5772-5788.

[10] C.J. Isham, Quantum logic and the histories approach to quantum theory, Journal of Mathematical Physics 35 (1994), 2157-2185.

[11] S. Minaee, et al., Large language models: A Survey, arXiv: 2402.06196.

[12] M.A. Nielsen and I.L. Chuang, Quantum computation and quantum information, Cambridge University Press, Cambridge, England, 2001.

[13] V. Paulsen, Completely Bounded Maps and Operator Algebras, Cambridge University Press, Cambridge, 2002.

[14] M. Reed, B. Simon, Method of Modern Mathematical Physics, Vol. I, Academic Press, San Diego, 1980.

[15] M. Reed, B. Simon, Method of Modern Mathematical Physics, Vol. II, Academic Press, Cambridge, 1980.

[16] W. Rudin, Functional Analysis, Second Edition, The McGraw-Hill Companies, Inc., New York, 1991.

[17] P. Petersen, J. Zech, Mathematical Theory of Deep Learning, arXiv:2407.18384v3, 2025.

[18] K. Sharma, M. Cerezo, L. Cincio, P.J. Coles, Trainability of dissipative perceptron-based quantum neural networks, Physical Review Letters 128 (2022), 180505: 1-7.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems, 30 (2017), 5998-6008.

[20] J. Vuckovic, A. Baratin, R.T. Combes, A mathematical theory of attention, arXiv: 2007.02876.

[21] Y. Zhang, et al., Tensor product attention is all you need, arXiv: 2501.06425.

[22] W.X. Zhao, et al., A survey of large language models, arXiv: 2303.18223.

**Wuhan Institute of Physics and Mathematics, IAPM, Chinese Academy of Sciences, 30 West District, Xiao-Hong-Shan, Wuhan 430071, China.**

**E-mail address:** chenzeqian@hotmail.com

*Note: Figure translations are in progress. See original paper for figures.*