Physical models realizing the transformer architecture of large language models

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

The introduction of the transformer architecture in 2017 (cf.[11]) marked the most striking advancement in natural language processing. The transformer is a model architecture relying entirely on an attention mechanism to draw global dependencies between input and output. However, we believe there is a gap in our theoretical understanding of what the transformer is, and why it works physically. In this paper, from a physical perspective on modern chips, 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. Our physical models underlie the transformer architecture for large language models.

PACS numbers:

Large language models (LLMs for short) are based on deep neural networks (DNNs) (cf.[5]), and a common characteristic of DNNs is their compositional nature: data is processed sequentially, layer by layer, resulting in a discrete-time dynamical system. The introduction of the transformer architecture in 2017 marked the most striking advancement in terms of DNNs (cf.[11]). 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.[15] and references therein).

The transformer has a modularization framework and is constructed by two main building blocks: self-attention and feed-forward neutral networks. Self-attention is an attention mechanism (cf.[1]) relating different positions of a single sequence in order to compute a representation of the sequence. In line with successful large language models, one often focuses 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. 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]).

To the best of our knowledge, physical models for the transformer architecture of large language models are usually described by using systems of mean-field interacting particles (cf. [4, 13] and references therein), i.e., large language models are regarded as classical statistical systems. However, since modern chips process data through controlling the flow of electric current, i.e., the dynamics of largely many electrons, so they should be regarded as quantum statistical ensembles and open quantum systems from a physical perspective (cf.[2, 12]). In this paper, 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.

In the transformer architecture of a large language model $\mathfrak{G}$, we assume the finite set $\mathbf{T}$ of tokens in $\mathfrak{G}$ 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. A finite sequence $\{x_i\}_{i=1}^n$ of tokens is called a text for $\mathfrak{G}$, simply denoted by $T = x_1 x_2 \cdots x_n$ or $(x_1, x_2, \cdots, x_n)$, where n is called the length of the text T. We write $[n] = \{1, 2, \ldots, n\}$ for an integer n. For a Hilbert space $\mathbb{K}$, we use $\mathcal{L}(\mathbb{K})$ and $\mathcal{S}(\mathbb{K})$ respectively to denote the set of all linear bounded operators and the set of all density operators in $\mathbb{K}$.

Recall that a self-attention layer SelfAtt with an attention mechanism (W^Q, W^K, W^V) in the transformer architecture is the only layer that combines different tokens, where W^Q and W^K are two $d' \times d$ real matrixes (i.e., the query and key matrixes) and W^V is the $d \times d$ real matrix (called the value matrix) such that $W^V x \in \mathbf{T}$ for $x \in \mathbf{T}$. Let us denote the input text to the layer by $X = \{x_i\}_{i=1}^n$. For each $i \in [n]$, letting

$$s_i = \frac{1}{\sqrt{d}} \langle W^Q x_n, W^K x_i \rangle, \quad \forall i \in [n],$$

we can interpret $S^{(n)} = \{s_i\}_{i=1}^n$ as similarities between the n-th token x_n (i.e., the query) and the other tokens (i.e., keys). The softmax layer is given by

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

which can be interpreted as the probability for the n-th query to "attend" to the i-th key. Then the self-attention

*Electronic address: chenzeqian@hotmail.com

layer SelfAtt is defined by

$$\mathrm{SelfAtt}(X)_n = \sum_{i=1}^{n} \mathrm{softmax}(S^{(n)})_i W^V x_i, \qquad (1)$$

indicating that the output $W^V x_i$ occurs with the probability $\mathrm{softmax}(S^{(n)})_i$, which is often referred to as the values of the token x_i. Note that (W^Q, W^K, W^V) are trainable parameters in the transformer architecture.

In the same block, a feed-forward neural network FFN is then applied to $W^V x_i$'s such that $y_i = \mathrm{FFN}(W^V x_i)$ with the probability $\mathrm{softmax}(S^{(n)})_i$ for each $i \in [n]$, and so the output is $x_{n+1} = y_i = \mathrm{FFN} \circ \mathrm{SelfAtt}(\{x_i\}_{i=1}^n)$ for some $i \in [n]$. One can then apply the same operations to the extended sequence $x_1 x_2 \cdots x_n x_{n+1}$ in a next block, obtaining $x_{n+2} = \mathrm{FFN}' \circ \mathrm{SelfAtt}'(\{x_i\}_{i=1}^{n+1})$, to iteratively compute further tokens (there is usually a stopping criterion based on a special token).

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 $\{\mathrm{FFN}_\ell\}_{\ell=1}^L$, i.e.,

$$\mathrm{Transf}_L = (\mathrm{FFN}_L \circ \mathrm{SelfAtt}_L) \circ \cdots \circ (\mathrm{FFN}_1 \circ \mathrm{SelfAtt}_1) \quad (2)$$

where the indices of the layers SelfAtt and FFN in (2) indicate the use of different trainable parameters in each of the block. Then, given an input text $T = x_1 \cdots x_n$, Transf_L generates a text $y_{i_1} \cdots y_{i_L}$ with the joint probability

$$P_T(y_{i_1}, \cdots, y_{i_L}) = \mathrm{softmax}(S_1^{(n)})_{i_1} \cdots \mathrm{softmax}(S_L^{(n+L-1)})_{i_L}$$

where $\mathrm{softmax}(S^{(n+\ell-1)_\ell})_{i_\ell}$ is given by the attention mechanism $(W_\ell^Q, W_\ell^K, W_\ell^V)$ in the ℓ-th building block for each $\ell = 1, \ldots, L$.

Now, we are ready to construct a physical model realizing the transformer architecture (2) for LLMs. To this end, consider a large language model $\mathfrak{S}$ with the set $\mathbf{T}$ of N tokens embedded in $\mathbb{R}^d$. Let $\mathbf{h}$ be the Hilbert space with $\{|x\rangle : x \in \mathbf{T}\}$ being an orthogonal basis, and we identity $x = |x\rangle$ for $x \in \mathbf{T}$. Let $\mathbb{H} = \mathcal{F}(\mathbf{h})$ be the Fock space over $\mathbf{h}$, that is,

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

where $\mathbf{h}^{\otimes n}$ is the n-fold tensor product of $\mathbf{h}$ (cf.[9]) In what follows, for the sake of convenience, we involve the finite Fock space

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

where M is an integer such that $M \gg N$. 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 $h^{(n)} = h_1 \otimes \ldots \otimes h_n \in \mathbf{h}^{\otimes n}$,

$$A h^{(n)} = (A_1 h_1) \otimes \ldots \otimes (A_n h_n) \in \mathbf{h}^{\otimes n},$$

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

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

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

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

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

It suffices to construct a physical model in the Fock space $\mathbb{H} = \mathcal{F}^{(M)}(\mathbf{h})$ $(M \gg L)$ for a transformer Transf_L (2) with a composition of L blocks, consisting of L self-attention maps $\{\mathrm{SelfAtt}_\ell = (W_\ell^Q, W_\ell^K, W_\ell^V)\}_{\ell=1}^L$ 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, for $\ell = 1, \ldots, L$, one has

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

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

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

A physical model needed to construct for Transf_L consist 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 $\mathbb{H}$ (cf.[7]), 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 by 2^Ω the set of all subsets of Ω, and write $\mathbf{D} = (\{\omega\} : \omega \in \Omega)$. At first, for an input text $T = x_1 \cdots x_n$, the input state ρ_T is given as

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

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

$$\begin{aligned}\mathcal{E}(t_1,&t_0)\rho(t_0)\\ &= \sum_{i=1}^{n} \mathrm{softmax}(S_1^{(n)})_i \mathrm{diag}(0, 0^{(1)} \cdots, 0^{(n)}, |x_1\rangle\langle x_1| \otimes \ldots \otimes |x_n\rangle\langle x_n| \otimes |y_i^{(1)}\rangle\langle y_i^{(1)}|, 0^{(n+2)}, \cdots),\end{aligned} \tag{3}$$

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

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

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

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

Making a measurement $(X_1, \mathbf{D})$ at time t_1, we obtain an output $y_i^{(1)}$ with probability $\mathrm{softmax}(S_1^{(n)})_i$ and, according to the von Neumann-Lüders reduction postulate (cf.[7]), the appropriate density operator to use for any further calculation is

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

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

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

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

$$\begin{aligned}\mathcal{E}(t_2, t_0)\rho_{\mathrm{red}}(t_1)_{i_1} =& \sum_{i_2=1}^{n+1} \mathrm{softmax}(S_2^{(n+1)})_{i_2}\\ &\times \mathrm{diag}(0,0^{(1)}, \cdots, 0^{(n+1)}, |x_1\rangle\langle x_1| \otimes \ldots \otimes |x_n\rangle\langle x_n| \otimes |y_{i_1}^{(1)}\rangle\langle y_{i_1}^{(1)}| \otimes |y_{i_2}^{(2)}\rangle\langle y_{i_2}^{(2)}|, 0^{(n+3)}, \cdots)\end{aligned} \tag{4}$$

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

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

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

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

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

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

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

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

Step by step, we obtain a physical model $\{\mathcal{E}(t_\ell, t_0)\}_{\ell=1}^{L}$ with the input state $\rho_T = \rho(t_0)$ as given an input text $T = x_1 \cdots x_n$, such that a text $(y^{(1)}_{i_1}, y^{(2)}_{i_2}, \ldots, y^{(L)}_{i_L})$ is generated with the probability

$$\begin{aligned}\mathrm{P}_T(y^{(1)}_{i_1}, &y^{(2)}_{i_2}, \ldots, y^{(L)}_{i_L})\\ &= \mathrm{softmax}(S^{(n)}_1)_{i_1} \cdots \mathrm{softmax}(S^{(n+L-1)}_L)_{i_L},\end{aligned}$$

within the sequential measurement $(X_1, \mathbf{D}), \ldots, (X_L, \mathbf{D})$. Thus, the physical model so constructed realizes the transformer architecture Transf_L.

A physical model for the transformer with a multi-headed attention (cf.[11]) can be constructed in a similar way. Also, we can construct physical models for the transformer of more complex structure (cf.[14] and reference therein). We omit the details.

Example. Let $\mathbf{T} = \{x_0, x_1\}$ be 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

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

If $W^Q_1 = W^V_1 = \mathrm{FFN}_1 = I$ and $W^K_1 = \sigma_x$ in $\mathbb{R}^2$, an associated physical operation $\mathcal{E}(t_1, t_0)$ at time t_1 satisfies

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

By measurement, we obtain x_0 with probability $\frac{2}{e+2}$ and obtain x_1 with probability $\frac{e}{e+2}$, 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 $W^Q_2 = W^K_2 = \mathrm{FFN}_2 = I$ and $W^V_2 = \sigma_x$ in $\mathbb{R}^2$, an associated quantum operation $\mathcal{E}(t_2, t_0)$ at time t_2 satisfies

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

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

Hence, we obtain the joint probability distributions:

$$P_T(x_0,x_0) = \frac{2}{e+2}\frac{1}{3e+1} = \frac{2}{(e+2)(3e+1)},$$
$$P_T(x_0,x_1) = \frac{2}{e+2}\frac{3e}{3e+1} = \frac{6e}{(e+2)(3e+1)},$$
$$P_T(x_1,x_0) = \frac{e}{e+2}\frac{e}{e+1} = \frac{e^2}{(e+2)(e+1)},$$
$$P_T(x_1,x_1) = \frac{e}{e+2}\frac{1}{e+1} = \frac{e}{(e+2)(e+1)}.$$

In conclusion, 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. Note that physical models satisfying the 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. Therefor, 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 quantum operations $\{\mathcal{E}(t_\ell,t_0)\}_{\ell=1}^L$, which are adjustable and learned during the training process, determining the physical model, as corresponding to the parameter matrixes $\{(W_\ell^Q,W_\ell^K,W_\ell^V)\}_{\ell=1}^L$ in Transf_L. From a physical perspective, training for a large language model is just to determine the quantum operations $\{\mathcal{E}(t_\ell,t_0)\}_{\ell=1}^L$ associated with the corresponding open quantum system (cf.[10]). This means that our physical models underlie the transformer architecture for large language models. We refer to [3] for a mathematical foundation of general AI, including quantum AI.

Appendix. In this appendix, we show that if a building block consists of an attention mechanism (W^Q,W^K,W^V) and FFN in a transformer architecture, then there is a quantum operation $\mathcal{E}$ in $\mathbb{H}$ depending only on (W^Q,W^K,W^V) and FFN such that given an input text $T=\{x_i\}_{i=1}^n$, for the input state

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

the quantum operation $\mathcal{E}$ satisfies

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

where $y_i=\mathrm{FFN}(W^Vx_i)$'s $(i\in[n])$ are given by (W^Q,W^K,W^V) and FFN.

To this end, 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.,

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

for $n\geq 1$. We first define

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

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

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

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

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

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,

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

where $a_0, a_x, a_{x_1,\cdots,x_n}$ are any complex numbers for $n\geq 1$. Note that $\mathbf{S}$ is a commutative C^*-algebra. By Stinespring's theorem (cf.[8, Theorem 3.11]), $\mathcal{E}_\Phi:\mathbf{S}\mapsto\mathcal{L}(\mathbb{H})$ is completely positive. Hence, by Arveson's extension theorem (cf.[8, Theorem 7.5]), $\mathcal{E}_\Phi$ extends to a completely positive operator $\mathcal{E}$ in $\mathcal{L}(\mathbb{H})$, i.e., a quantum operation in $\mathbb{H}$ (note that $\mathcal{E}$ is not necessarily unique). By the construction, $\mathcal{E}$ satisfies the required condition.

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

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

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

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

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

[6] S. Minaee, *et al.,* Large language models: A survey, *arXiv* (2025), 2402.06196v3.

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

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

[9] M. Reed, B. Simon, *Method of Mordern Mathematical Physics,* Vol. I, Academic Press, San Diego, 1980.

[10] 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.

[11] 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.

[12] C.J. Villas-Boas, C.E. Máximo, P.J. Paulino, R.P. Bachelard, G. Rempe, Bright and dark states of light: The quantum origin of classical interference, *Physical Review Letters* **134** (2025), 133603: 1-6.

[13] J. Vuckovic, A. Baratin, R.T. Combes, A mathematical theory of attention, *arXiv* (2020), 2007.02876.

[14] Y. Zhang, *et al.*, Tensor product attention is all you need, *arXiv* (2025), 2501.06425.

[15] W.X. Zhao, *et al.,* A survey of large language models, *arXiv* (2025), 2303.18223v16.