我们假设自然语言具有类似于量子力学的特性，采用这种类比的理由不仅来自于注意到自然语言的意义可以用状态叠加的形式来描述，还因为我们注意到自然语言符号的表示具有一种符号-意义对应的二元性。

同时，基于统计实现的自然语言处理（NLP）方法已经流行了数十年，并不断发展。这让我们思考是否能够超越统计理论，考虑将量子力学的相关理论引入自然语言建模中。在过去的十年中，NLP 普遍采用将自然语言符号转换为某种数学表示进行处理的方法，即所谓的词向量化，该方法在通用语言处理任务中取得了惊人的良好表现。我们发现，将构成自然语言的符号完全转换为数值表示进行处理以构建通用语言模型（ULM）的技术思想是高度可行且可能的。这两个真实世界应用反馈的实验进展激励我们探索基于量子理论构建自然语言模型。

在过去的一个世纪里，随着人类对物理世界的深入研究，量子理论变得越来越完善。人类使用自然语言进行交流时，词义通常具有模糊性。这种现象激发了一群研究人员尝试通过量子力学同样“难以捉摸”的性质来解释这种现象 [2] 。我们应该注意到，逻辑上严格的量子力学可以为我们提供形式化的数学工具来建模可用的一组量子语言（QLM）。

我们集合了一系列为建模提供基础的预设。这些假设是通过观察人类长期以来对自然语言的使用而得出的，尚需进一步验证。

假设 I：自然语言中存在符号-意义的二元性。同时，意义表现为潜在概念的叠加。自然语言的字符集描述了许多不同意义的同时存在，其形式类似于叠加态。我们假设自然语言符号是某种量子系统的参照物。

假设二：自然语言递归地由多个级别的子系统组合而成，这些子系统在组合时相互作用。这种相互作用导致量子系统状态的改变，符号代表了量子系统的状态，这使得符号在特定语言环境中展现出特定的意义。

假设 III：根据量子理论，量子系统的状态函数提供的信息包含了该系统的所有属性。这是基于实验反馈提出的一个更为基础的假设。

在本文的上半部分，基于这些假设的完整模型似乎可以相当合理地解释许多现有的语言建模方法。同时，它使我们能够通过该模型结合量子理论对自然语言的属性进行一些推测。

本文进一步探讨了构建通用自然语言模型的关键步骤，并指出使用量子态来表示自然语言符号的序列可以自然地包含上下文信息对词语的影响。在进一步的尝试中，本文利用量子统计学研究了大规模符号表示的量子系统聚合时的宏观性质。这引导我们提出一个更为根本的猜想来解释：“为什么自然语言可以用这种量子理论形式建模?”以进行更本质的推测：

信息的物理性：信息存在于某种物理结构上。显然，这种物理结构不仅可以用经典力学或经典统计理论来描述，还可以用量子理论来描述。这使得应用严格的量子力学理论来描述自然语言的状态、演化和统计属性成为可能。

类似的实现也在 Randall 的研究 [4] 中提到。本文尝试将这些想法更深入地研究到 QLM 中。

值得一提的是，在构建模型的过程中，我们发现了与当前词嵌入的 NLP 基础技术相关的模型。一方面，现有模型应该能够说明我们建模思想的可行性；另一方面，这些模型为主流方法提供了理论解释。反过来，主流模型也可以从这些原则中得到补充。

在应用方面，由于我们仍然没有实际操作和测量完整量子态的相关技术。在本文中，我们尝试在神经网络的基础上对模型进行一些验证。我们可以利用神经网络的原理来模拟人类学习语言的过程。与其深入讨论计算机科学中具体神经网络的构建，本文将神经网络作为一种工具来构建算法，并尽可能避免工程细节对模型的影响。

在将语言视为量子统计系统的背景下，我们在一个简单的含时间的语料库上验证了语言量子统计性质随时间的演变，并将结果与社会现象进行了相关性分析。

总之，本文的结论在四个方面显示了促进作用：为 NLP 处理中的词嵌入技术提供了解释，以及为机器学习提供了新的训练理念，尝试将信息物理学的研究扩展到量子领域，最后提供在数学语言学中使用量子统计方法的可能性。

自然语言处理领域正在经历另一个快速发展的时期。由 GPT、Deepseek 等代表的大型语言模型展现了具有里程碑意义的性能。

我们首先快速了解自然语言领域主流模型的发展，然后集中研究与尚不成熟的量子语言建模概念相关的研究。

2.1 自然语言处理的主流方法

回顾自然语言处理领域的过去，随着计算机和人工智能领域中 NLP 问题的诞生，在过去的七十多年里出现了许多不同的实现。我们在此列出了一些被认可的发展节点。

在上个世纪，自然语言处理领域的早期理念是基于手动编写的语言处理规则，包括概念依赖理论、专家系统和其他方法。从九十年代开始，诸如隐马尔可夫模型、条件随机场等方法使得统计工具进入了研究人员的视野。当 N-gram 模型被广泛研究时，统计方法已经成为机器学习的重要工具。

在过去的二十年中，随着神经网络作为工具的改进，诸如 RNN、LSTM 和 GRU 之类的自然语言处理架构在深度学习的概念下被展开。在经过两个重要模型的过渡之后，即 word2vec [5] 和 seq2seq [6] ，Transformer 模型架构引入了编码器-解码器结构和自注意力机制 [11] ，取代了循环神经网络（RNN），成为 NLP 的新主流模型。在使用预训练机制的 Transformer 模型机制上，出现了大量的半监督学习 Gpt 系列模型 [12] [13] [14] [15] ，值得单独提及的是在 InstructGPT [16] 的 RLHF 机制 [17] 的训练过程中引入了人类反馈。

大型语言建模领域后续的整体发展显示出商业激励和科学研究交织在一起的趋势，促使模型快速迭代。每个 NLP 模型可以等同于一个高度任务特定的神经网络架构，其伴随的训练方法，必要时，还包括用于训练的数据集。

该领域当前的前沿问题包括多模态使得大语言模型能够跨自然语言处理边界处理图像、声音、视频等数据。参数微调的方法能够提高模型性能、模型可解释性以及其他问题。推理建模的出现使得大语言模型展现出某些机器思维能力。

值得注意的是，这样的持续迭代已经促使一些带有前代模型技术的新模型的出现。如果我们将整个技术发展

的传承比作一棵大树，我们可以说，这棵树的主干，即当前自然语言处理的主导思想，是基于统计性质的语言模型。显然，统计方法已经取得了惊人的成果，同时探索其他分支上的可能理论模型也同样重要。这样的想法引导我们关注已经非常成熟的量子理论。

在量子语言建模的命题下，早期研究的关键词指向量子信息检索的问题。虽然信息检索所研究的问题与本文的初衷略有不同。信息检索关注的是从非结构化数据中检索所需的信息，其核心是文档和查询的匹配。而自然语言处理领域致力于让计算机能够理解、解释和生成人类语言。

在这一领域中，值得提及的是 Sordoni 等人的研究结果 [9] 。该研究受到经典 N 元模型的启发，其中词被视为量子事件，并将量子概率引入统计语言建模中。为尚不清晰的 QLM 概念提供了一套逻辑上完整的设计架构。后续由张等人进行的工作 [10] ，推广了 Sordoni 的模型，通过改进使其能够与复杂的神经网络兼容，以取代略显过时的最大适应性估计来估计参数。有趣的是，张的模型中包括了一个使用复杂相位来编码词序的机制。这样做赋予了 Sordoni 模型中未使用的复数虚部实际意义。这些模型底层的技术思想基于词袋模型，其目标是通过量子统计理论来推广早期的统计语言模型。

信息检索问题和自然语言处理任务都有一些共同的技术基础。例如，本论文重点讨论的语言建模是两者领域共有的一个基础概念，其关注点在于如何数学地建模自然语言。基于这种联系，我们很高兴将相关工作列为参照结果。

也有一些零散的尝试是基于将不同技术结合在一起的目标，比如将量化与最新的 Transformer 模型结合起来。在量子神经网络上实现 NLP 的方法。除此之外，还有大量通过微调迭代传统技术的尝试，以优化现有模型的收益，这里不胜枚举。

如前一节所述，近年来自然语言处理领域出现了大量的重要发展。就考虑本文结果对技术的影响而言，本文中的模型大致可以看作是在 RNN 系列方法之后、Transformer 模型出现之前，对这一发展的解释和修改的一个分支。

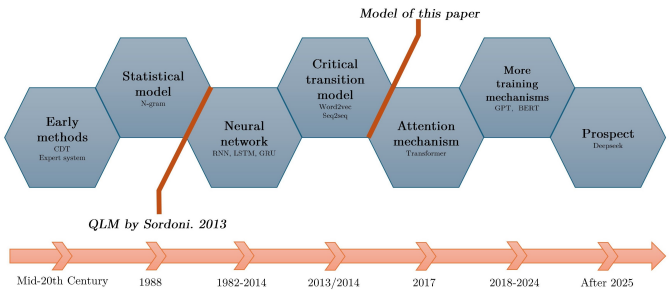

FIG. 1: 自然语言处理发展脉络

三、量子理论的介入

3.1 量子语言模型

本节构建了自然语言和量子态的映射模型。在自然语言研究中，我们最感兴趣的是由一系列语言或名义符号表示的意义。通过选择适当的基质，语义可以用希尔伯特空间及其上的代数来表示。自然语言由诸如单词和句子之类的基本语言元素组成。我们根据量子统计形式来定义上面提到的这些语言元素。每个语言元素被视为一个独立的量子系统，语言元素具有以下一般形式。

$$\rho_{elem} = \sum_i p_i \left|elem\right\rangle_i \left\langle elem\right|_i \tag{1}$$

上述定义将语言元素描述为混合态，可以直观地认为这些元素以概率 p_i 处于量子态 $\left|elem\right\rangle_i$ 。这种形式允许任何语言元素表示为有限维度的矩阵。其数学形式满足密度矩阵的所有约束条件。

UTF8gkai Specifically, the so-called linguistic elements can be character states ρ_{char} , word states ρ_{word} , sentence states ρ_{sente} or even paragraph states ρ_{para} or even longer natural language arrangements. Or let us step outside of time and space constraints, such as the Chinese character radicals $\rho_{日}$, All symbols of the entire language family $\rho_{Niger-Congo}$, A summary of all the languages used in the period $\rho_{La\ Troisième\ Rèpublique}$. Attention that these linguistic elements may not contain information about serialized structures. 密度矩阵具有容易描述混合状态量子系统的优势，并且在许多场景中，密度矩阵提供了一种更方便的处理方式。然而，本文使用密度矩阵来描述语言元素的根本动机是基于经典信息理论，该理论指出任何语言符号或符号集都必须具有信息熵。如果将语言元素视为纯态，那么整个系统的状态是完全已知的，这只能代表一些不具有普遍性的特殊情况。所有在人类日常使用中发展的自然语言都存在不确定性。这种不确定性既体现在决定语言元素量子态的概率幅度上，也体现在表现系统特定量子态的统计算符的不确定性上。经典熵描述关注的是后者。这是一种定性描述，关于信息熵的相关讨论可以见下文。

根据假设 III ，我们假设描述语言元素的量子态可以在由其自身属性组成的基底上展开。使用狄拉克符号描述是为了回避表征问题的处理。对于使用 ket 符号约束的语言元素 $\left|elem\right\rangle$ ，它是语言元素的量子叠加态。它的状态向量可以在一个特定属性基上展开，如下所示：

$$\left|elem\right\rangle = \sum_j c_j \left|attr\right\rangle_i \tag{2}$$

这意味着语言元素 $\left|elem\right\rangle$ 处于这些底物 $\left|attr\right\rangle$ 的叠加态。根据选择的底物，$\left|attr\right\rangle_i$ 在量子力学中起着基态的作用。这样做的好处是，类似于量子力学的形式，量子态可以根据不同的表示展示不同的信息。

自然语言具有外部特性，包括词汇特性如单词特性和发音，以及语法特性，这些是更复杂的语法概念，如词形和词位。这些内在特性可以统称为外在特性，以便与描述语言元素的量子系统进行比较。

采用这样一种更广泛的定义，一方面允许语言学家同时研究语言的许多其他外在属性。自然语言的一个有趣属性是，当我们研究例如词汇的各种属性时，仍然需要用自然语言本身来描述。事实上，通用的语言模型已经展示了能够以对话形式解决各种自然语言处理子任务的能力。很难不想象这些语言的属性是“自描述”的。这为研究这些外在属性提供了一种巧妙的方法。当我们表示量子态时，它使得总是可以扩展一组空间上完备基态的结合。

在本文中，我们关注于主流自然语言处理方法中产生结果的“语义”属性。事实上，以上描述已经表明，外部属性也可以被分解为一组更基本的基底集合。可以说，上述所有属性都可以表示为“语义空间”的某种子空间。语义的属性本质上是量子系统中所包含的内在属性。

3.2 语义表示

3.2.1 基本示例

对于自然语言，我们最感兴趣的属性是其序列所表示的意义。我们首先构建一个简单情况的描述，然后推广到一个更一般的理论。

首先假设一个语言元素 w 具有 i 个确定的语义。在这种最简单的情况下，我们可能希望查阅词典，使词语 w 具有 i 个确定的词典释义。这些词典释义中的每一个都由一些句子表达。根据假设 II，这些含义是由使用 w 产生的含义，而在解释中选择哪种特定的含义显然取决于上下文环境。

同时，w 的每个词汇意义都可以被视作为 j 个底层意义片段的概率幅度叠加。在这个例子中，我们借用语言学术语”义素”来指代这些基础意义片段。义素是构成一个词的意义的最小单位。

用 $|sema\rangle_i$ 表示第 i 个语义的语义向量，一组语义向量 $s = \{|seme\rangle_1, |seme\rangle_2 \dots |seme\rangle_j\}$ 构成单词 w 本身的语义基质。

基于其自身的语义基质的量子态扩展到密度矩阵得出：

$$\begin{aligned} \rho_{word} &= \sum_i p_i \left|sema\right\rangle_i \left\langle sema\right|_i \\ &= \sum_{n,m} \sigma_{nm} \left|seme\right\rangle_n \left\langle seme\right|_m \\ \sigma_{nm} &= \sum_i p_i c_n^{(i)} c_m^{(i)\dagger} \end{aligned} \tag{3}$$

这里 c_j 描述了量子态的内在概率幅度，而上述密度矩阵中的 p_i 描述了系统的统计属性。σ 是相应矩阵元素的系数。我们可以解释为，每个语言元素 w 的意义同时处于多重语义的叠加态，使其同时具有这些语义的意义。

根据量子力学的惯例，示例中的底物集合被命名为本征语义底物，意味着它是语义“可观测量”的“特征值”。我们将其以矩阵的形式量化。

示例 3-1：

我们选择词典 $\{computer, vector\}$ 进行说明。词向量在本文的发布版本中总共出现了 40 次，其中 6 次用作量子力学中状态向量意义的替代，剩下的 25 次指的是计算机技术词语中嵌入的词，9 次作为例子中的代词出现。

其他选定词语的示例含义及其在本文中的分布如下表所示：

各个特征语义基质由相互正交的基构成。根据定义（3），我们计算每个词汇的矩阵表示如下。

$$\rho_{vector} = \begin{bmatrix} 0.15 & 0 & 0 \\ 0 & 0.625 & 0 \\ 0 & 0 & 0.225 \end{bmatrix} \quad \rho_{computer} = \begin{bmatrix} 0.285 & 0 & 0 \\ 0 & 0.25 & 0 \\ 0 & 0 & 0.465 \end{bmatrix}$$

	p_i	$substrate_v$	p_i	$substrate_c$
$\|\text{Sema}\rangle_1$	$\frac{2}{7}$	$\|\text{quantum state}\rangle$	$\frac{1}{7}$	$\|\text{quantum computer}\rangle$
$\|\text{Sema}\rangle_2$	$\frac{5}{8}$	$\|\text{word mapping}\rangle$	$\frac{1}{4}$	$\|\text{classical computer}\rangle$
$\|\text{Sema}\rangle_3$	$\frac{9}{40}$	$\|\text{pronoun}\rangle$	$\frac{13}{28}$	$\|\text{pronoun}\rangle$

需要强调的是，本文中对模型的描述和使用都不涉及对量子态的“观察”操作。熟悉量子力学的读者可能会很容易联想到，“在使用（观察）词汇 w 时，以概率 c_i 解释（塌缩）为语义元素 $|seme\rangle_i$ ”这样的一种解释。在这个简单的示例中，这样的解释似乎是正确的。然而，我们建议阁下不要关注这种表述。优先考虑的是量子态的数学形式而不是“观察”，并且通过态的叠加原理强调某个特定的基态，其中语言元素同时承担这些元素的意义，而不是通过某些操作。这是因为接下来在模型的应用阶段，我们将在经典计算机上直接操作语言元素的数学形式。

3.2.2 常用语义底层

在前一节中，我们描述了一个易于理解的一般示例，我们可以看到其中的语言元素、意义和语义可以在一些自然语言中进行描述。然而，当研究对象的规模增加时，持续引入的新语义和底层结构的数量呈线性增长。

我们可以注意到，可能通过包含重复的语素叠加可以构成不同的语义。那么，是否可以找到一个基底集合，使得能够构造叠加态来表示研究对象集合中所有语言元素的量子态。

再进一步，似乎基底可通过自然语言进行解释这一特性同样不是必要的。在这一点上，基底不能再严格地被称为语义单元；它不能再被语言描述。显然，我们只需要知道语言元素量子态的公式化，并确保研究对象集都是在相同的公式化之下。这体现在寻找一组共同的语义基底，从中任何操作被形式化。

同样基于量子力学的惯例，我们可以将这样一组用于相同语言元素的基底称为共同语义基底。

在当前用于自然语言处理的词嵌入技术中，将自然语言符号映射到向量是很常见的。这种方法仅仅来自于实践反馈，并没有完全在理论上得到支持。事实上，到目前为止，我们为这种广泛使用的基本技术提供了一个结

合物理方法和语言学的良好的解释。可以认为在本文中，词向量可以被视为量子表示的简化近似。严格来说，这纯粹是本文尝试进行量子建模语言时的偶然结果。

例 3-2： 让我们继续前面的例子。

在这个例子中，两个词汇的语义被进一步分解为包含相同语义的基本单元。密度矩阵通过公式（3）计算，以展示基于共同语义基质的多重词汇的密度矩阵表示。

Computer for meaning quantum computer :

[0.2em] $|seme_{computer}\rangle_1 = |quantum\rangle + |state\rangle + |machine\rangle$

[0.2em] Computer for meaning classical computer :

[0.2em] $|seme_{computer}\rangle_2 = |classical\rangle + |state\rangle + |machine\rangle$

[0.2em] Vector for meaning quantum state :

[0.2em] $|seme_{vector}\rangle_1 = |quantum\rangle + |state\rangle$

[0.2em] Vector for meaning word embedded :

[0.2em] $|seme_{vector}\rangle_2 = |word\rangle + |mapping\rangle$

[0.2em]

$$\rho_{vector} = \begin{bmatrix} 0.083 & 0 & 0.083 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0.083 & 0 & 0.083 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.416 & 0.416 \\ 0 & 0 & 0 & 0 & 0.416 & 0.416 \end{bmatrix}$$

$$\rho_{computer} = \begin{bmatrix} 0.177 & 0.177 & 0.177 & 0 & 0 & 0 \\ 0.177 & 0.333 & 0.333 & 0.155 & 0 & 0 \\ 0.177 & 0.333 & 0.333 & 0.155 & 0 & 0 \\ 0 & 0.155 & 0.155 & 0.155 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

C-S SubStrate	vector		computer	
	p_{sema}	c_{seme}	p_{sema}	c_{seme}
$\|quantum\rangle$	$\frac{1}{6}$	$\frac{1}{\sqrt{2}}$	$\frac{8}{15}$	$\frac{1}{\sqrt{3}}$
$\|state\rangle$	$\frac{2}{6}$	$\frac{1}{\sqrt{2}}$	1	$\frac{1}{\sqrt{3}}$
$\|machine\rangle$	0	0	1	$\frac{1}{\sqrt{3}}$
$\|classical\rangle$	0	0	$\frac{7}{15}$	$\frac{1}{\sqrt{3}}$
$\|word\rangle$	$\frac{5}{6}$	$\frac{1}{\sqrt{2}}$	0	0
$\|mapping\rangle$	$\frac{5}{6}$	$\frac{1}{\sqrt{2}}$	0	0

3.2.3 量子态嵌入的优势

最初采用词向量的视角指出其对可以在语言意义之间进行的代数运算的关注。一个常被引用的例子是 $\overrightarrow{queen} = \overrightarrow{women} + \overrightarrow{king}$，这促使研究人员尝试用向量以代数方式表示自然语言。

在数学形式上，量子纯态确实可以表示为向量。本文中的模型指出，仅仅将嵌入方法视为一种向量属性，并不具有各种量子力学属性。在量子力学框架内，纯态的语义向量只能是其本征态的线性叠加。这为单词嵌入的数学形式和对象的新属性提供了新的约束。在向量空间中执行代数运算扩展到算子运算，而在希尔伯特空间中这些运算具有物理意义。

正如假设 II 中提到的，任何自然语言的意义都会因特定的语言环境而有所不同。与本节中促进底物的步骤相似，这种多义性不仅表现于多义词之中，还体现在语言使用背景中意义的细微变化，这些变化无法用自然语言严谨地描述。

对于一个可以视为混合态的语言元素来表示其确定语义中的任何一个，在特定语言环境中产生的细微语义变化反映在量子态的微小波动中。以密度矩阵形式的统计加权有助于描述这两种多义属性。

在下一章中，我们还将通过将它们描述为密度矩阵的形式，展示自然语言给归一化操作带来的许多好处。

四、语言元素的交互

4.1 结构化语言元素

4.1.1 语言构成

词组成句子，句子构成段落。自然语言通过潜在的构建规则产生更长且信息量更大的结构化序列。本节描述了这一过程。量子力学的框架本身包含了对系统间复杂性的描述。

对于两个不相互作用的量子系统的通用复合，其状态空间通常是子系统的张量积。对于状态在 ρ_j 的 j 个语言元素的复合体，其状态可以表示为：

$$\rho^{'} = \rho_1 \otimes \rho_2 \otimes \ldots \rho_j \tag{4}$$

组合产生的元素 $\rho^{'}$ 的上标用来表示基元素 ρ 比前者高一级。在本文中，组合的复合量子系统被称为语言组合。语言组合也可以被视为一个语言元素。

复合过程可以简化地描述为相同基本操作的重复进行的结果。一个元素进入系统的基本操作可以定义如下：

$$\rho^{'}_{end} = \rho^{'}_{start} \otimes \rho_{ins} \tag{5}$$

复合混合态矩阵可以被分解为组成子系统的底层纯态的混合统计表示：

$$\rho_{end} = \sum_{ij} p_i p_j \left|\phi_i\right\rangle_{start} \left|\phi_j\right\rangle_{ins} \left\langle\phi_i\right|_{start} \left\langle\phi_j\right|_{ins} \tag{6}$$

这种复合过程在实际语言使用中可以被相当直观地解释。句子中最有可能表现的意义来自每个词中最常见和最主要的语义的复合。

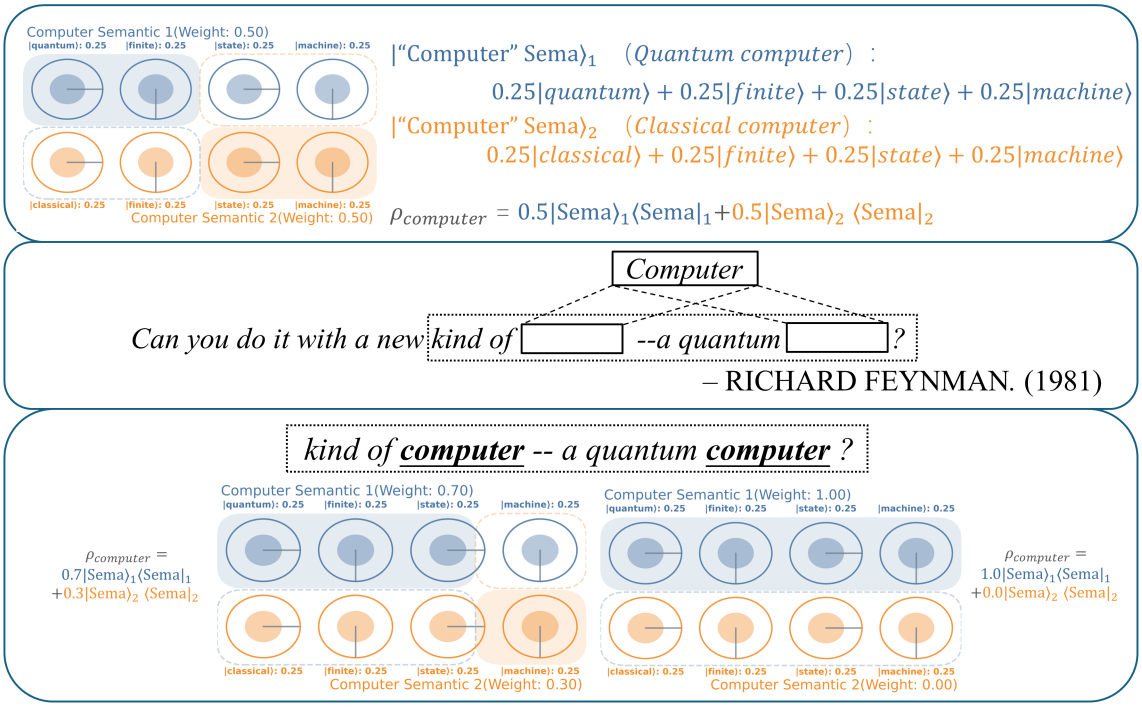

FIG. 2: 例子 4-1

显然，这种表示预设个体元素被视为独立的量子系统，它们之间没有相互作用。复合系统生成一个张量积状态，其基底是组合之前各元素基底的张量积。从我们上面列出的例子可以回忆，如果个体元素获得一个本征语义基底，那么复合后的语言组合将获得其构成元素的语义。但是，对于更一般的情况，我们为一组语言元素集合获取一个共同的语义基底时，在同一集合上将这些基底彼此叠加，会生成一些基底的幂等直积。

这似乎暗示着高阶语言元素必然包含更多信息，如较大的量子系统。它还引出了一个推论，即基于这种描述，对于一组有限的自然语言符号的共同内在基质，所有可能的自然语言序列的意义可以表示为该基质自身的幂张量乘积。这一假设尚需通过语言实验来验证。

值得一提的是，近年来 NLP 研究日益关注的问题之一是自然语言是一种语境相关的语言。继承自早期的模型，仅考虑语言的统计特性，在诸如生成任务等新的 NLP 任务中表现不佳。事实证明，在引入能够将语境有序的信息编码到模型中的机制之后，NLP 性能出现了质的飞跃。

量子态的数值表示可以将这种现象采用为量子系统之间的相互作用。在语言元素进入语境的过程中，它们的连续组合和分解过程中，语义空间的状态会实时地随着这些行为而变化，语义空间中自然语言符号的信息通过密度矩阵单个算符的统计聚合来呈现。

本文论证，这正是前一章所提到的使用量子态的基础，能够在特定的语言环境中表示语言元素的语义波动。

参见图 2 和示例 4-1，了解一个词在进入特定上下文环境时不同语义权重的变换的示例。图中的可视化方案使用面积来指示测量双量子位的特定基态的概率，在这个例子中，相位不是关键因素，但通过圆内的铅垂线来表示。每一行代表一个不同语义的基态，每一行底部的颜色面积表示密度矩阵中对应语义的比例。

图 3 中的例 4-2 表示另一种现象。当特定词语互相组合时，语言的语义基底发生变化。之前不存在的基底出现了，或者某个基底自行消失了。后者意味着在组合之前存在于子系统中的某个具有概率振幅的基底在组合之后变为零。这与上面直接乘积结果不同，这在量子力学中被称为纠缠态，通常是由于量子系统之间的相互作用。

回到语言使用，这种情况对应于繁琐的语法规则或特定的语言惯例。这些情况阻止了将复合过程用简单的量子系统复合来描述。

4.1.3 组合的数值近似

在本文中，提出了找到一种方法来专门表示后复合语言成分的密度矩阵仍然是一个极其重要的问题。这不仅涉及到模型的完整性，也是使模型可用的关键步骤。目前的量子理论仍然没有全面的纠缠形式理论。这已将问题从严格的理论问题转变为在有限时间内的工程问题。本节简要指出了一些值得考虑的方向，并提供了一些建设性的想法。同时也分析了当今丰富的自然语言处理方法作为实验结果的适用性。

从当前的自然语言处理方法应用来看，设计一个复合的词向量表示贯穿整个语句也是模型中的一个关键步骤。在 CBOW 方法中，简单地将词向量的平均值用作句子的词向量。[5] 在进一步的模型中，从序列的角度来看，设计了一系列称为编码器的神经网络架构。这使得序列中的符号可以通过网络的非线性参数按顺序统一到一个结果向量中。[6]

我们认为，由于密度矩阵的统计性质，这样如前所述的直接平均方法不会破坏模型的物理意义。这种训练方法的原因在于词向量的复合表示仅作为训练中的一个中间步骤使用，而不是专注于复合向量的具体意义。这个直接的方法是可行的，直到有更优的方法被发现。

对于进一步的序列模型，它充分利用了神经网络的非线性拟合能力。复合向量生成的方式过于不透明，以至于无法进行进一步讨论。当与本文内容结合时，我们推测这种复合过程可能与下一节中提到的语言集合在理论上有某些联系。此外，该模型催生了 Transformer 模型，该模型通过添加一个旨在限制复合向量生成的上下文窗口而表现良好，并且仍然值得进一步关注。

量子多体系统：如果我们继续从物理建模的角度来看问题，自然语言的序列化属性使得它可以被视为一维长量子系统的链条。例如，一个由几个不同原子组成的量子多体系统，我们的语言模型遇到了一定的相似性。已经存在一种称为矩阵积态的方法来表示这种量子系统的复合情况，该方法可以近似一维量子链的密度矩阵，并考虑诸如纠缠等现象。本文认为，基于这种方法的建模将是一个强有力的候选项。

马尔可夫链：更准确地说，在本文中默认是一个量子马尔可夫过程。马尔可夫链最初是在语言学研究中提出的，其自洽理论的发展已被广泛应用于自然语言处理任务中。马尔可夫性质表达了系统未来状态仅与当前状态相关的思想。我们可以认为语言元素的量子态仅在执行上述基本操作时发生变化，这使得以离散的方式研究整个系统的演变成为可能，而无需依赖时间变量。此外，将整个语言序列直接视为量子马尔可夫链是一种结合经典思想与本文开发的量子模型的泛化。

严格的量子建模：如上所述，整个序列可以被看作是基本操作的不断重复以获得结果。研究的对象被简化为大环境中的小型量子系统研究。从量子统计理论出发，通过将环境视为一个大型热库来获得小量子系统状态的变化。另一方面，如果我们进一步将对应自然语言的量子系统扩展为严格符合量子理论的量子力学系统，那么在知道具体相互作用的哈密顿量后，可以通过克劳斯算子和部分迹表示来描述操作对子系统和环境的量子态的影响。下一节将展示这种思路带来的进一步实用的结论。
本节的挫折在于，直接从语言元素的密度矩阵中获得较长的结构化序列表示似乎受到了阻碍。我们需要对暂时称为“语言元素之间的相互作用”进行一些研究。具体细节需要在理论推导的基础上经过实验验证。

4.2 语言演变

4.2.1 语言集成

现在我们想要了解关于语言元素的宏观性质。让我们尝试从量子统计物理的角度来看语言元素的演变。大量由相同基质集合表示的语言元素构成了一种可以视为语言集的东西。

如同文章最初对所给实例的警告，这种组合产生的语言元素也可以用密度矩阵来描述。必须仔细考虑系统内是否有结构化信息，即是否存在交互。这种语言集合合成可以对应于语言学中语料库的量子抽象表示。

对于一个未使用的语言元素，比如一个未被置于上下文考量中的句子，我们认为其状态是静态的。其下一个序列的语言元素的状态不会随时间改变。这种描述同样适用于语言集合，其中限制条件的描述应适应长时间不与其他系统交换的语言元素的情形。在这种情况下，系统内语言元素的状态不会随时间改变，隔离系统处于热力学平衡状态。

一个系统不与外界交换能量和粒子且其体积保持恒定的条件称为微正则系统。如果我们继续这个类似的量子理论。那么微正则量子系综的结论也可以应用于语言系综：

系统的熵 S 可以直接从系统的密度矩阵 ρ 推导出来，并且 ρ 与系统的微观状态数 $\Omega(E)$ 相关。k_B 是玻尔兹曼常数。这些系统函数有助于研究系综系统的性质。

上一节的讨论指出，当一个语言元素进入系统时，它引发的状态变化可以在研究中被视为瞬态。与时间无关，我们将系统状态的研究离散化为每个时刻。将系统近似为处于平衡状态，可以通过应用上述方程获得系统的熵。

但是，让我们从长远看。在自然语言的实际使用中，例如，在各种日常对话中使用的各种语言元素中，新词被创造，或旧词被赋予新的意义，或旧有意义被侵蚀。随着社会的进步，新概念被引入，人们需要找到新的词汇来表达它们。随着文化传播，外来词进入不同的语言。语言元素的量子状态一直在变化。

这种演变中是否也存在某种模式或趋势？例如，统计物理学假设系统演变为更加统计平均的状态，机械系统演变到能量最小值。在假设一中提到的对偶性中，符号被映射到语义空间，随着它们的演变是否也存在某种原则？有待实验验证。

此外，有必要提到与其他系统进行信息交换的情况。这对应于一个只有能量交换且没有与外部世界粒子交换的正则系综条件下的封闭系统，以及一个与外部世界进行能量和粒子交换的巨正则系综的开放系统。从上述讨论中我们注意到，能量在这里被类比为信息量，而粒子则是自然语言符号。

例如，常规条件可以用来描述在第 4.1 节中试图表示语言构成的密度矩阵。同样，可以使用规范或大规范条件来研究语言现象，如在特定语言家族中由于文化交流引起的语言使用变化。在本文的应用部分进行了初步尝试。

4.2.2 熵与信息能量

熵的概念早已被引入到语言学中，用于信息论的研究。在一些体现其本质的实际应用中，信息熵被用来计算完全编码一个字符集所需的比特数。或者通过计算信息熵来获得信息片段出现概率的信息，从而计算出数据压缩的最小体积。这里的信息熵取以 2 为底的对数，通常被认为表示研究中的系统状态数量所需的比特数。

包括 Shannon 在内的学者 [8] ，已经将信息熵应用于数学语言学，以及将语言视为马尔可夫链来研究其条件熵的方法。在这一理论中，符号集的大小与系统中包含的信息量相关。这为一些经验现象提供了初步解释，例如，中文象形文字通常较短，或中文读者阅读速度较快。[7]

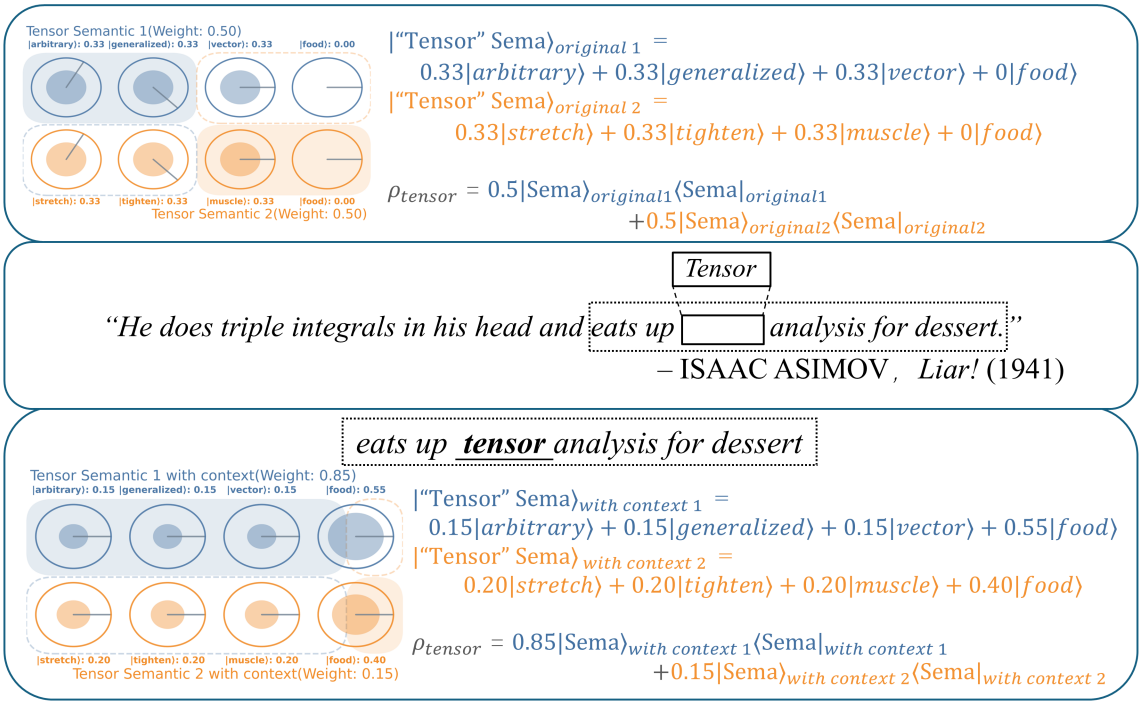

FIG. 3: 例子 4-2

这种解释性方法实际上强调在许多替代状态中存在系统不确定性，并且需要一定数量的符号来表示可能状态的完整范围。理论中符号集合的大小与系统中包含的信息相关。回顾我们假设 I 中提到的二元性，本文认为经典信息熵是对自然语言符号状态数的表征，涉及到这些符号出现的概率。

诚然，可以进一步证明，通过比较显示不同字符所需的二进制序列的长度，不同的符号携带的信息量不一致。这可以被视为对符号-意义二元性中表面符号的一个正在进行的研究。

经典信息熵理论的一个重要启示是，一个系统中存在可区分的状态是其包含信息的前提条件。

我们来看一个例子，这个例子指向这个特性。在生物信息学中，信息熵被用来研究特定的基因表达。这里存在类似符号-意义二元性的“基因名称（符号）-基因链（物理结构）”的二元性。不同的基因片段以不同的概率表达。当我们使用信息熵来研究基因表达时，信息熵可以指示每个基因片段的表达概率，但不能指示有关基因片段上碱基对数量的信息。我们无法从信息熵推断出频繁表达的基因具有更长的基因链。显然，碱基对的排列本身携带了一些信息。

这使我们意识到，也存在一些性质不包含在系统微观态数的概念中，因此无法用熵来表征它们。但是，这些性质与系统携带的信息量有关。

经典的香农熵在量子信息中被推广为冯 · 诺依曼熵：

$$S(\rho) = -\mathrm{Tr}(\rho \log \rho) \tag{7}$$

混合态系统自身带有不确定性，熵是它们的统计性质。在前一节的特定条件下，冯 · 诺依曼形式可以与混合态系统的熵相关联。

一个矛盾出现的领域是，量子熵在纯态下为零。这是因为虽然观察时表现出不确定性，但在物理意义上系统的纯态是完全确定性的。

一开始这引起了极大的困惑，如果我们继续采用经典熵的阐述，似乎纯态不携带信息。但当我们已经表明在语义空间中的任何量子态都对应于特定语义时，这种说法显然有问题，因为特定语义不含信息。

基于这些解释矛盾的表现，本论文认为，就像密度矩阵描述了混合态的统计特性和量子态的概率幅信息一样，这些信息同样可以良好地存储在量子态内部的概率幅内。在本文中，我们称这种上述提到的本质属性，属于语言符号的量子态表示为信息能量。

这里显然包含的一个先决假设是，每个自然语言量子态的基质都携带相同基本量的信息传递。只有这样，我们才能量化两个纯态的信息能量特性。如果我们更普遍地假设不同的基质携带不同数量的信息能量，那么我们就无法进行绝对的信息能量大小比较。

我们认为，经典信息理论只是暗示了表达所有符号所需的状态数量，而量子模型则可以进一步让我们了解到符号所携带的信息量。在统计分布中，根据系统统计性质的不确定性来表达信息携带，而在量子状态中，则基于量子性质展示了确定性的信息携带。

4.2.3 信息的物理性

我们谨慎地借用了大量量子理论来解释观察到的语言现象。到目前为止，这些异想天开的理论居然看起来真的切实可行。

我们想知道为什么这些抽象符号，源于人类生活长期实践的产物，会与物理系统共享相似性。这种相似性仅仅体现在代数结构上吗?

本文试图给出一个基本的解释。

首先是一种推理，源自科学本身，因为我们无法观察到宇宙存在于它之外，所以信息必定存在于可观测宇宙的某处。这个假设在本文中暂时被称为信息的物理性：信息必须以某种形式存在于宇宙中的物质排列中。这使得人类可以使用作为某些物理结构替代物的符号。我们可以假设，代表语言的信息必须存储在宇宙中以量子力学方式描述的某个系统之上。

这个假设使我们无需谨慎地称本文中提到的理论为“通过量子力学的代数形式的类比”，并努力使自然语言的现象符合严格的量子力学解释。

在这个解释性框架中，我们可以大胆地指出，信息能量应类似于哈密顿量。经典熵告诉我们，不同状态是赋予信息意义的关键。在本文中，我们认为这一方法的合理性正是源于这样一个原则性事实：不同状态是区分信息的基础。信息同样可以体现在例如量子系统的离散能级上。

目前，我们无法形成任何关于语义空间的基底如何映射到离散能级的量子系统的想法。仅观察到二者表现相似。不同的能级显然对应于不同的能量，使得我们在上一节中试图量化信息能量的尝试变得更加受限。

但是如果我们进一步并停止关注信息能量的大小，而只关注信息能量的状态数量会怎样呢? 语言的信息性与量子系统希尔伯特空间的大小有关。

通过这种解释，尽管量子系统可能具有不一致的哈密顿量，导致不同的结构和演化方式，但仍然可以强调系统的状态情形，并找到量子系统到自然语言符号的映射。而且最有希望的是，引入类比的哈密顿量使我们能够看到描述自然语言演化对应于量子系统的完整方法。

语言信息熵的研究已经存在很长时间了，但还没有进一步的研究来推导出其他有意义的语言“热力学函数”。在信息论中，关于其他“热力学量”的意义仍然没有确凿的证据。这些信息的热力学量的表征使我们能够快速获得语言集体的状态。这超出了本文的原始目的，因此不会进一步发展。

V. 评估

5.1 与旧技术的比较性能

将自然语言嵌入数值表示的方法在无需进一步验证的情况下在本文中取得了巨大成功。本文为这种技术提供了规范和理论支持。本节的主要目的是验证将语言嵌入密度矩阵的新方法的可行性，并在一组公平的测试中比较两者的性能。

然而，嵌入对象之间存在根本的差异。这使得如何确定旧模型和新模型是否具有相同的参数尺度成为一个值得质疑的问题。同样，在旧模型中，已经存在许多隐含的代码优化以及经过长期使用发展而来的算法迭代，这也影响了结果的性能。本节中的结果仅供参考。实验代码在附录的代码声明中可以找到，其中提供了运行环境的更详细描述。

当前对嵌入的验证不可避免地使用神经网络。在本文研究的问题中，参考人类理解语言的过程是有用的。所有语言都是在人生的后期学习，以便理解它们的使用。对于初学者来说，这些词可能看起来空白或者什么都没意义。然而，当一个人接触到越来越多的语言使用样本，最终会在大脑中形成符号与其意义的对应关系。神经网络的训练是模拟上述过程的尝试。

让我们关注本文中模型的具体实现。实现过程中的一个困难在于，密度矩阵的意义比简单的向量要复杂得多，随意调整向量的参数只会改变其属性，而随意调整密度矩阵的参数可能使其不再合法。从结果向后推导，密度矩阵必须满足几个条件，例如厄米矩阵、半正定，以及迹为 1。

我们使用成熟的数值方法 Cholesky 分解在训练过程中保持数据结构满足密度矩阵的约束。为了简化，我们在实验中使用所有实数。在这里，我们还假设下三角矩阵的对角元素是非负的和实数，这样密度矩阵就对应于一个唯一的分解矩阵。

很容易将保真度的使用与用于训练的向量相似性测量方法中的余弦相似性方法替换关联起来。

比较是使用 CBOW 嵌入训练方法进行的，该方法已在上文中进行了分析，其理论上的量子态替代对结果没有影响。CBOW 方法基于一个基本假设，即一个词的词向量应该与其所嵌入的上下文环境的总向量相同。同时参考论文 [3] 对几种词嵌入方法进行了绝对评估，我们将密度矩阵嵌入的方法与传统向量嵌入方法进行了并排比较。每个数据集的绝对结果如表 I 所示。

TABLE I: 绝对评估结果

Model	WS	WSS	WSR	MEN	RG65
Quantum(3Q)	0.0382	0.0146	0.0466	0.0495	0.0232
Classical(36D)	0.1160	0.0638	0.1822	0.0073	0.0651

Abbreviations: WS=WordSim, WSS=WordSim-Similarity, WSR=WordSim-Relatedness, RG65=Rubenstein-Goodenough.

每个数据集都是对词对相似性的人类评估结果。评估通过计算嵌入结果与人工结果之间的相似性来完成，并将两者的皮尔逊相关系数制成表格。密度矩阵嵌入的结果均为正值，并且可以看到新模型倾向于从人类结果中学习。

需要注意的是，在这个实验设置中，由 3 个量子比特组成的 8×8 的密度矩阵是由其下三角矩阵唯一确定的。下三角矩阵参数的数量是 36，这与用于比较的矢量方法的维数相同。严格来说，我们可以在相同的存储大小下比较新旧模型。考虑到前面提到的看不见的差异、代码库中的随机优化等，这里的比较只是定性的信息。

替换嵌入目标不仅重新设计了词嵌入技术的基础，还可以被视为在本文理论下对向量方法的理论精炼。因此，该技术仍有广阔的发展空间。本文的重点是验证新方法的可行性。

5.2 观察语言演变

模型的后半部分的一个重要推论是，许多人类语言现象将在其量子态表示中具有某种形式的规律性。鉴于我们没有找到一个基于时间分类的结构良好的语料库，本论文尝试设计一个简单的实验来展示这种方法的可行性。

我们将分析一个由中国历史组成的简单语料库，以尝试捕捉这一模式的尾迹。使用东方语言有几个优点：汉语虽然其字形经历了变化，但自从始皇统一文字以来，一直使用相同的符号系统；并且在历史上有许多民族融合和广泛的文化交流时期，这使得它适合进行上述研究；使用真实历史的理由是，这些历史是在一千多年的时间里编写的，继朝代的官方正字法对其语言进行了标点和语法校对。

我们首先使用上一节的方法训练词嵌入，并通过观察不同时期存在集合函数的变化。在我们对大多数热力学函数在语言中的作用缺乏认识以及对如何测量信息能量缺乏线索之间，我们首先研究冯 · 诺依曼熵，这个参数已经可以有一个定义。

如果这些预测得以实现，在像唐代这样的大规模民族融合时期，语言中将会涌入新的概念，这将导致汉语中某些系数的可观察变化。本文将从汉代到清代的时期分为九个历史时期，依据每个朝代集中修订历史的时期。分时期的结果显示在附录中。

需要简要注意的是，尽管《元史》和《明史》撰写时间非常接近，它们还是被划分为不同的时期。划分的理由不仅仅是直观上的王朝更迭，还包括对历史环境的理解，即在修订《元史》时有意去蒙古化所使用的语言的倾向。分析表明，这种划分有助于反映这一部分的结果。对于希望严格按照出版时期划分的读者，也计算了晚元和早明四部史书作为一个数据集的冯 · 诺伊曼熵，为 1.6336。供您参考。

同样，对于被归类为“中国古代”时期的《三国志》，其仍然是用文言文写成的而不是用白话文写成的，并且在一定程度上继承了清朝的语言。总之，本文并非严格的历史讨论。本文尽可能在有限的语料库上验证结果，希望强调历史的读者能包容我们。

将《三国志》归类为“汉”时期也是如此，因为它仍然遵循由太史公司马迁创建的历史书写风格，并且其文本的使用与汉代史书相近。总之，这篇论文不是严格的历史讨论。尽量在有限的语料库上验证结果，希望重视历史的读者能谅解。

每个训练阶段的语言集合的 Neumann 熵结果，使用结论 1 的模型及其参数，如图 4 所示。

在进一步分析结果之前，重要的是要注意。作者认为，尽管在本文中作为数学模型进行了研究，研究者仍必须对语言有清晰的理解。然而，研究者重要的是要清楚地理解这样一个事实，即“语言”在更一般的背景下与许多复杂的社会科学学科相关联。这种数学方法应仅作为研究某些学科（如人类学）的参考。

首先，该数据集中的某些区域会影响实验的精度。每个数据的大小各不相同，而现有的分词程序对古代汉语的支持不够好。除此之外，计算机学者还应考虑训练过程中各种参数的影响。

基于上述前提条件，我们可以说，在一个精度有限的实验中，本文观察到了模型推测结果。在中国，有一种传统，即在新朝建立之初修订前朝的历史。举例来说，本文认为“晚唐”和“宋”时期包含有唐朝的语言使用遗留物。因此，我们在唐朝整个时期和从明清到现代的时期有两次冯 · 诺依曼熵的增加。

本文中的模型表明，语言集合的冯 · 诺伊曼熵的减少意味着语言中的统计性质正在减弱，同一个词的歧义性也在减少。然而，人类生产中的概念总是随时间增加，与语言对应的信息量也应该在增长。因此，可以假设语言中的信息能量量必须随着时间增加。

设想一些现实生活中需要考虑的语言应用，语言的使用者在接触到新事物时，倾向于借用现有词汇来指代新概念，这增加了语言歧义的统计性质。而随着时间的推移，这些概念随着用于指代它们的词汇的使用而被规范化。词汇的歧义逐渐再次减少，而语言所指代的概念变得更复杂，信息能量正在增加。

唐朝前后的发展与初唐时期中国境内各民族的融合以及中唐时期一个与世界相连的帝国的出现有关。从明清时期到现代也是一个技术飞速发展的时期，甚至出现了现代工业革命，将人类世界的各个部分连为一体。

在过去的两千两百年里，语言的冯 · 诺依曼熵总体呈下降趋势，如上所述，我们可以看到“语言正在消化概念”。本文认为，信息逐渐“被纳入量子系统的内部状态”，因此整体的熵正在减少。

在本文的过程中，我们确实尝试解决分期语料库和古代汉字分词的问题。然而，很快认识到这些工作量大的项目远远超出了本文的研究主题。好消息是，从一个初步结果来看，利用本文设想的理论解释对进一步研究有影响。祝贺！

5.3 通用量子语言模型

古典计算机：我们已经描述了对现有广义自然语言模型的影响。本节探讨了利用本文内容构建量子模型，并探讨了两个构建理念的障碍。

一个问题是表示复合系统的问题，这在上文中已经探讨过，很明显，组合和分解任意大小的语言元素的能力是实现普遍性的捷径。本文模型对现有 NLP 技术影响的初步讨论也已在第 4.1.3 节提供。如前一章的分析所指出的那样，似乎在从语言元素精确构建语言复合的密度矩阵方面存在很大的困难，直到对纠缠现象有充分理解。暂且把这个恼人的绊脚石放在一边，看看我们能否为其他想法构建一些有意义的结果。

另一种途径是将量子态表示整合到现有的自然语言处理技术路线上。一个例子是尝试在规则系统条件下表示语言组合，这在第 4.2.1 节中提到。由于对温度、配分函数等关键热力学函数的理解仍然不足，因此无法应用。在获得这种瞬态响应的完整描述后，能否将其与编码器-解码器模型结合，我们推测，由于模型使用神经网络暴力法将元素拟合到整体序列数值关系上，本文中的模型可能已经存在于网络的非线性参数中。

在编码器-解码器的技术框架中，虽然我们在算法的中间步骤中生成了某种整体序列的表示，但这并不是整个任务的重点。[6] 我们不需要强制整体表示是序列的精确表示这一事实是非常有效的。或者，使用量子态表示允

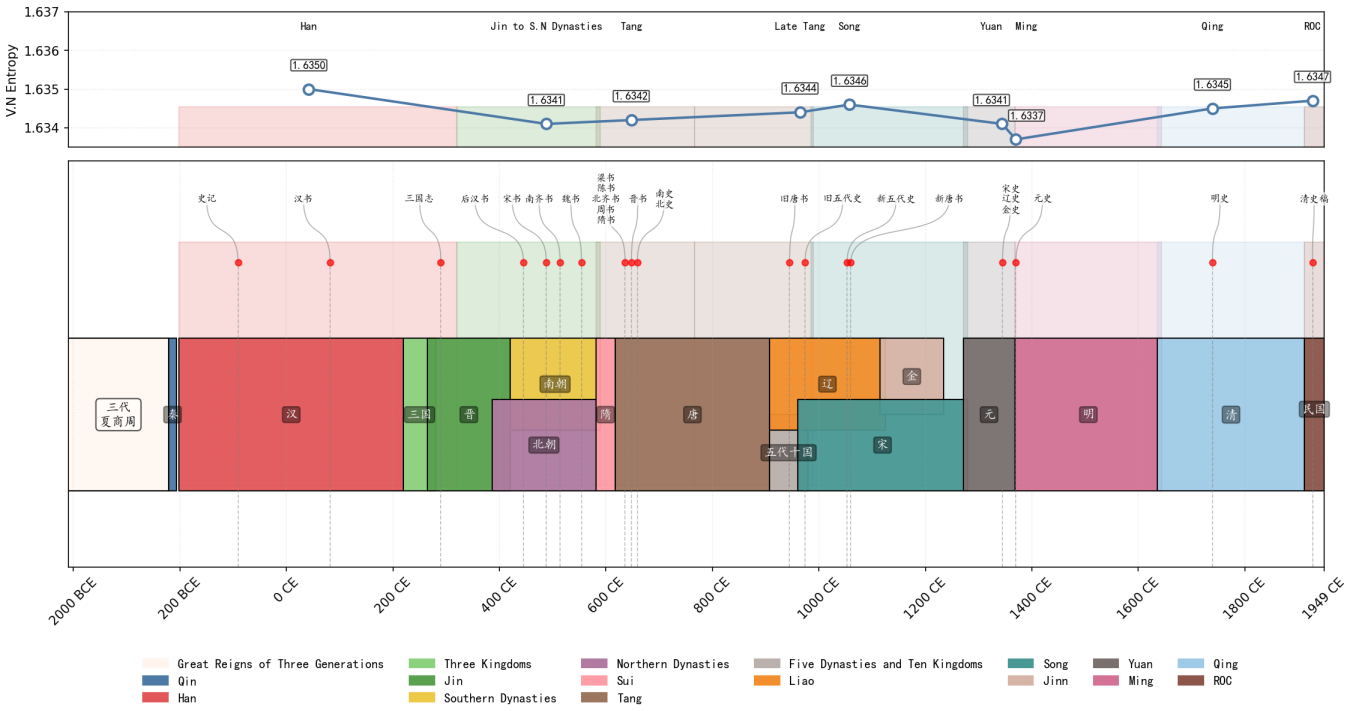

FIG. 4: 各种朝代历史的 V.N 熵

许我们重新配置编码器-解码器框架，例如，通过模拟量子随机游走。

量子计算机：现在让我们回到量子工程中的终极和基本问题，这个模型在量子计算机上能做什么？

在当前的量子电路中，对密度矩阵的编码是通过分离纯态和概率项来实现的。这意味着将密度矩阵分解为 n 个纯态的凸组合需要总共 $2 * log_2 n$ 个量子比特来实现。$log_2 n$ 个主要量子比特表示纯态，而 $log_2 n$ 个辅助量子比特通过旋转门表示概率项。由于这种概率叠加在“量子数据结构”中始终保持纯净，它总是维护密度矩阵的合法性。保真度的计算也通过交换门得到很好地解决。

负责编码概率的主要和次要量子比特都呈对数增长。图 5 显示了编码三个量子比特的密度矩阵的量子电路示意图。

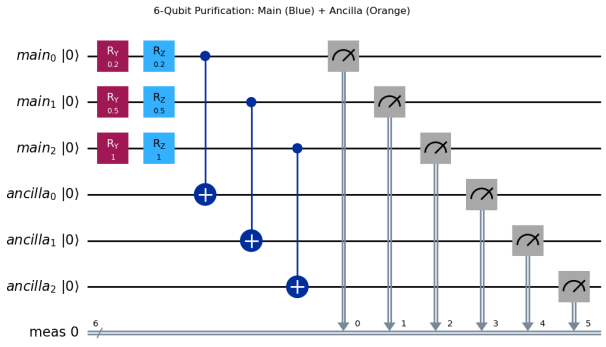

FIG. 5: 用于编码三比特密度矩阵的量子电路

由相似电路生成的量子态仍然需要在专为量子计算机设计的神经网络上进行训练。无论是编码密度矩阵还是将神经网络提升为量子电路，这种处理都可以被准确地描述为使用量子电路复制经典算法。从这个角度来看，这似乎是本末倒置。

我们设想通过量子力学过程，主动使一个量子系统处于混合态，而不是将其数学表示分解到由多个量子位组成的数据结构中，并直接对本文第三章中提到的语言元素之间的相互作用进行建模。

考虑到拥有十个量子比特的量子存储，我们可以在理论上编码一个 1024 维的叠加态。这相当于一个与之维数相同的语义空间，已经超出了目前 Chatgpt 使用的 768 维向量空间。就目前的技术而言，量子计算操作是直接在量子存储介质上进行的，我们将这种用于存储量子语言信息的内存简称为 Q-drive。

如果一个量子电路被设计成让量子计算机在影响 Q 驱动的状态时不断读取（输入）文本（信息），这使我们能够通过量子操作直接将语言元素的量子状态记录到这个 Q 驱动上。这使我们能够跳过神经网络方法，该方法相当于对系统状态的蛮力再现。这就像费曼提到的直接通过量子计算机模拟量子现象一样。

这导致了一种更纯净和更自然的方式来获取有关量子态的信息，并且由于通信的结果，内在具有随时间生成新状态的能力。无需区分预训练的数据集等，或设计机制来确保语言模型的数据随时间变化或由于用户交互而变化。

显然，在生成量子态、保持密度矩阵合法性、保真度计算和其他由物理架构决定的 Q 驱动问题中存在天然优势。

如果我们只关心信息是否能够被表示，而不是使用二

进制正则化信息。这似乎甚至有助于我们构建“非门基”的量子通用计算机。

这样做的缺点也很明显，基于我们无法读取这种 Q 驱动的确切状态这一公理。交互者只能通过提供信息来影响其内部量子状态，使其在交互者的认知之外自行演变。通过与其交互，我们可以根据其内部量子状态获得反馈。

有人说语言是思想的载体。前面的描述让我们觉得仿佛在与一位真实的智能进行互动。这种思路的可行性尚待观察。至少，“这听起来像是我们又向人工智能迈进了一步。”

六、结论

模型的起始点来自实验结果，即近年来自然语言处理技术的显著进展，并根据从语言使用中总结的一些假设建立了一个量子语言模型。我们回顾了文章的关键研究要素：

在第一部分中，介绍了通过量子力学代数对自然语言符号进行数值表示的方法，并指出现有的词嵌入技术是一种简化情形。

在第二部分中，讨论了语言元素的复合过程。同时探讨了通过量子统计研究语言学的宏观特性及其随时间演变的可能性。

在第三部分中讨论了应用，包括在经典计算机上的近似和现有嵌入技术的改进。还讨论了使用量子统计研究自然语言的可行性。此外，还包括了对未来应用的具体想法。

语言的抽象意义存在于一个高维希尔伯特空间中，由一组语义本征子层表示。任何明确的意义都可以在这个空间中表示为一个量子态。假设可能存在一组通用的本征子层，可以容纳所有人类语言，使得任何语义意义都可以由它表示。

在本文中，语言的基本组成部分，即语言元素，相当于指代上述空间中叠加态的统计混合的符号。这种表示法的优点在于它可以涵盖特定语境中语义的小波动，以及同时具有多重意义的情况。由于我们在现实中没有诸如无法观察叠加态的物理限制，我们认为使用语言元素等同于直接访问关于语言叠加态的原始信息。（这在测量公理下当然是被禁止的；我们关注其数学形式。）

我们认为可以基于这些模型从量子统计的角度研究语言的演化。对于孤立的语言系统，符合微正则系综条件。对于其他有信号和符号交换的系统，选择其他系统性条件。为了使两个系统彼此复杂，我们假设遵循正则系综条件。可以假设，两个语言系综相互作用的一些进化过程可以使用巨正则系综的外部条件来研究。本文没有进一步阐述这一点。

关于这一系列模型建立的基本解释，本文推测这是基于信息存在于物理系统这一原则事实。因此，建议量子系统的能量对应于系统信息能量的研究量。

从应用的角度来看，本文尝试对大语言模型背后的技术进行理论建模，并相应地对其进行微调和改进。尝试设计一些使用量子相关原则和算法的新计算技术。同时，这对于在量子计算机上进一步构建人工智能也有所帮助。

本文的理论扩展了有趣的问题：神经科学家可能会问：在神经元的微观水平上是否存在量子过程？意识是量子的吗？我们可能会问，本文的模型是否表面上暗示思想的概念基于量子结构？概念是否只是我们宇宙中的希尔伯特空间的一个子空间？

在技术层面：量子态的复数部分有任何意义吗？近期流行的推理语言模型是否与时间演化有关？理性是否与离散状态的演化有关？

文本 [4] 中提到了 Rolf Randall 的研究。过去，它对作者理解信息论起到了重要作用。同时，我们应该提到，在语言学领域，也曾试图寻找语言意义的基元 [1] 。虽然与本文中的建模思想没有直接关系，但这显示了不同学科为理解人类语言而作出的共同努力。

本文的应用向我们展示了，一方面人类语言包含了思想，另一方面人类语言通过特定符号来承载。语义空间复杂化为思维空间。这似乎解释了大型语言模型的多模态能力。我们可以建立这样的层级：物理结构 « 信息 « 语言 « 思想。这个解释应当为当前人工智能技术思维提供一些支持。

此外，一个有趣的想法是，在量化自然语言符号的信息能量后，形成语言系统的三个变量：符号（语言）–统计分布（多义性）–信息能量（基质选择）。这些变量可以在一组人工设计的语言量子结构设备上进行平衡，以设计人工语言。

符号-统计分布的对应关系可以在经典信息论中已经被表述出来。例如象形文字和警句的以空间换取时间的技巧，前者以较大的字符表换取较短的序列长度。如果我们基于这三者，或者仅调整符号-信息能量的关系以淘汰语言的多义性。使自然语言如同数学语言，利用数学符号具有高度具体的意义。构建一种和谐的语言，使符号-信息能量达到数学优化的最佳解。这种和谐语言的个别符号将对应于系统的唯一纯态。

量子力学继续在意想不到的领域影响我们对自然的认知。由于线性代数的数学语言，彼此相距甚远的研究领域展现出类似的数学结构。这就是本文标题中“干预”一词的用意。

[1] An, H. A Study of Interpretative Primitives in Modern Chinese (China Social Science Publishing House, Beijing, 2005).

[2] Meng, B. Master's thesis, East China Normal University (2020).

[3] Schnabel, T. et al. In Conference on Empirical Methods in Natural Language Processing (2015). https://api.semanticscholar.org/CorpusID:6197592

[4] Landauer, R. Phys. Lett. A 217 , 188–193 (1996). 10.1016/0375-9601(96)00453-7

[5] Mikolov, T. et al. In Proc. 27th Int. Conf. Neural Inf. Process. Syst. 2 , 3111–3119 (Curran Associates, 2013).

[6] Sutskever, I. et al. In Proc. 28th Int. Conf. Neural Inf. Process. Syst. 2 , 3104–3112 (MIT Press, 2014).

[7] Feng, Z. Reform Writ. Syst. 4 , (1984).

[8] Shannon, C. E. Bell Syst. Tech. J. 30 , 50–64 (1951). 10.1002/j.1538-7305.1951.tb01366.x

[9] Sordoni, A. et al. In Proc. 36th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. 653–662 (ACM, 2013). 10.1145/2484028.2484098

[10] Zhang, P. et al. ACM Trans. Inf. Syst. 40 , 1–31 (2022). 10.1145/3505138

[11] Vaswani, A. et al. In Proc. 31st Int. Conf. Neural Inf. Process. Syst. 6000–6010 (Curran Associates, 2017).

[12] Radford, A. et al. In Conf. Neural Inf. Process. Syst. (2018). https://api.semanticscholar.org/CorpusID:49313245

[13] Radford, A. et al. In Conf. Neural Inf. Process. Syst. (2019). https://api.semanticscholar.org/CorpusID:160025533

[14] Brown, T. B. et al. arXiv (2020). https://api.semanticscholar.org/CorpusID:218971783

[15] Achiam, J. et al. arXiv (2023). https://api.semanticscholar.org/CorpusID:257532815

[16] Ouyang, L. et al. arXiv (2022). https://api.semanticscholar.org/CorpusID:246426909

[17] Knox, W. B. In Proc. Int. Conf. Mach. Learn. (2011). https://api.semanticscholar.org/CorpusID:17123490

UTF8gkai

TABLE II: 《二十五史》语料库的时期分类说明

Period	Title	Author(s)	Publication era	Dynasty
Han	史记 Record of the Grand Historian	司马迁 Sima Qian	征和二年 91 BCE	西汉 Western Han
	汉书 Book of Han	班固等 Ban Gu, et al.	建初五年 80 CE	东汉 Eastern Han
	三国志 Records of Three Kingdoms	陈寿 Chen Shou	太康年间 280-290 CE	西晋 Western Jin
Jin to S.N Dynasties	后汉书 Book of Later Han	范晔 Fan Ye	元嘉九年至二十二年 432-445 CE	刘宋 Liu Song
	宋书 Book of Song	沈约 Shen Yue	永明五年至六年 487-488 CE	梁 Liang
	南齐书 Book of Northern Qi	萧子显 Xiao Zixian	天监十三年至大同三年间 514–537 CE	梁 Liang
	魏书 Book of Wei	魏收 Wei Shou	天保二年至五年 551–554 CE	北齐 Northern Qi
Tang	晋书 Book of Jin	房玄龄等 Fang Xuanling, et al.	贞观二十年至二十二年 646-648 CE	唐 Tang
	梁书 Book of Liang	姚思廉 Yao Silian	贞观三年至十年 629–636 CE	唐 Tang
	陈书 Book of Chen	姚思廉 Yao Silian	贞观三年至十年 629–636 CE	唐 Tang
	北齐书 Book of Northern Qi	李百药 Li Baiyao	贞观三年至十年 629–636 CE	唐 Tang
	周书 Book of Zhou	令狐德🄵等 Linghu Defen, et al.	贞观三年至十年 629–636 CE	唐 Tang
	隋书 Book of Sui	魏徵等 Wei Zheng, et al.	贞观十年 & 显庆元年 636 CE & 656 CE *	唐 Tang
	南史 History of Southern Dynasties	李延寿 Li Yanshou	显庆四年 659 CE	唐 Tang
	北史 History of Northern Dynasties	李延寿 Li Yanshou	显庆四年 659 CE	唐 Tang
Late Tang	旧唐书 Old Book of Tang	刘🄵等 Liu Xu, et al.	开运二年 945 CE	后晋 Later Jin
	旧五代史 Old History of Five Dynasties	薛居正等 Xue Juzheng, et al.	开宝七年 974 CE	北宋 N. Song
Song	新五代史 New History of Five Dynasties	欧阳修 Ouyang Xiu	皇🄵五年 1053 CE	北宋 N. Song
	新唐书 New Book of Tang	欧阳修等 Ouyang Xiu, et al.	嘉🄵五年 1060 CE	北宋 N. Song
Yuan	辽史 History of Liao	脱脱等 Toqto'a, et al.	至正四年 1344 CE	元 Yuan
	金史 History of Jin	脱脱等 Toqto'a, et al.	至正四年 1344 CE	元 Yuan
	宋史 History of Song	脱脱等 Toqto'a, et al.	至正五年 1345 CE	元 Yuan

TABLE III:

《二十五史》文集的时期分类解释（续页）

Period	Title	Author(s)	Publication era	Dynasty
Ming	元史 History of Yuan	宋濂等 Song Lian, et al.	洪武三年 1370 CE	明 Ming
Qing	明史 History of Ming	张廷玉等 Zhang Tingyu, et al.	乾隆四年 1739	清 Qing
ROC	清史稿 Draft History of Qing	赵尔巽等 Zhao Erxun, et al.	民国十六年 1927 CE	民国 ROC