自然语言处理的量子语义框架

Christopher Agostino*
NPC Worldwide, Bloomington, Indiana 47403, USA

Quan Le Thien
Department of Physics, Indiana University, Bloomington, Indiana 47405, USA and
Quantum Science and Engineering Center (QSEC),
Indiana University, Bloomington, Indiana 47405, USA

Molly Apsel
Cognitive Science Program, Indiana University, Bloomington, Indiana 47405, USA and
Department of Psychological and Brain Sciences,
Indiana University, Bloomington, Indiana 47405, USA

Denizhan Pak
Cognitive Science Program, Indiana University, Bloomington, Indiana 47405, USA and
Luddy School of Informatics, Computing, and Engineering,
Indiana University, Bloomington, Indiana 47405, USA

Elina Lesyk
Independent Consultant, Munich 81549, Germany†

Ashabari Majumdar
Department of Physics, University of Notre Dame, Notre Dame, IN 46556, USA and
NPC Worldwide, Bloomington, Indiana 47403, USA

(Spontaneity Collaboration)

语义退化代表了自然语言的一种基本属性，这种属性不仅仅是简单的多义性，而是包括随着语义表达复杂性增加而出现的潜在解释的组合爆炸。大型语言模型（LLMs）和其他现代自然语言处理系统面临的固有限制正是因为它们在自然语言中运作，使其受到语义退化所施加的相同解释约束。在这项工作中，我们使用 Kolmogorov 复杂性论证，随着表达的复杂性增加，任何解释代理（无论是人类还是由 LLM 驱动的 AI）恢复单一预期含义的可能性消失。这种计算的不可解决性表明，传统的认为语言形式自身拥有意义的观点是有缺陷的。我们提出相反的观点，认为意义实际上通过观察者依赖的解释行为而实现。为此，我们使用各种 LLM 代理作为“计算认知系统”，在不同的上下文设置下解释模糊词对，从而进行语义贝尔不等式测试。在几项独立实验中，我们发现平均 CHSH 期望值范围从 1.2 到 2.8，其中几次运行产生的值（例如，2.3-2.4）显著违反了经典边界（$|S| \leq 2$）。这表明在模糊性下的语言解释可以表现出非经典上下文性，与人类认知实验的结果一致。这些结果内在地意味着使用经典频率论为基础的方法进行自然语言分析本质上是有损的。相反，我们提出采用贝叶斯风格的重复抽样方法可以提供更实用且适当的语言意义的上下文特征。

在 2010 年代深度学习革命之前，从 1980 年代中期到 2000 年代初经历了一个漫长的人工智能寒冬。在这段时间里，大多数自然语言处理和人工智能的研究人员认为——基于 1960 年代和 1970 年代未能实现的人工智能承诺——计算机本身永远无法产生任何形式的通用智能或模拟被描述为自然语言的现象。尝试的人工智能研究人员通常发现自己陷入了在极端情况下会失效的启发式中，通常会放弃这些任务，而转向更易解决的问题。参见 Dreyfus 的《计算机仍然不能做的事情》，以更好地理解研究人员在广泛使用大规模神经网络（如谷歌翻译和后续出现的其他成功工具如 transformers、大型语言模型）之前所感知的人工智能推理的逻辑和语义限制。这些现在已成为消费产品中常见的大规模神经网络通常被称为分布式语义模型（DSMs）。实际上，DSMs 从统计共现中推断相关性和意义，它们已经支持各种高度实用的自然语言处理（NLP）应用，包括主题模型、情感分类和大型语言模型。根据构造，像 BERT 风格的主题模型这样的 DSMs 假定文档具有单一的、内在的语义组成。

由于动态系统方法（DSM）在许多领域的实际成功，一些研究人员已经忘记或未考虑到人工智能最初受到的许多批评，而这些批评至今仍然成立。他们假设这些限制也将通过更多的计算能力或更多的数据得到解决。然而，请考虑以下在当前自然语言处理领域中的突出问题：

1. 基于 DSM 的方法仍然存在局限性，特别是在处理复杂、模糊或语境丰富的文本时 [1, 2] 。

2. 与之前模型发布时的飞跃相比，最新一代前沿大型语言模型（LLMs）似乎明显缺乏进展，这似乎揭示了一个根本性的障碍，即更多的计算/数据并未缓解语义问题解决的可靠性问题（例如，GPT-4.5、Llama 4、Gemini 2.5、Claude 3.7 Sonnet）。同样，许多前沿研究人员已转向优先考虑更具动态能力的

* info@npcworldwi.de; http://www.cjagostino.com
† info@elinalesyk.com

模型，而不是在 LLMs 上的进展（例如，JEPA [3] 、MAMBA [4] 、CTM [5] ，或者对 LLMs 的修改使其能够自我提升 [6, 7] ）。

这些问题和不断发展的研究环境表明，基于 DSM 的方法所面临的挑战可能源于语言学中的一个更基本的问题：语义退化问题。这个概念超越了简单的一词多义，指的是在处理复杂的语言表达时，所出现的潜在解释的固有多样性 [8–10] 。

确实，这一概念直观上是合理的：经验观察向任何观察者揭示，自然语言的意义不是固定不变或绝对的 [11, 12] ，而解释本身在根本上是依赖于上下文的 [13–15] 。因此，任何代理在解释自然语言表达时实现的语义意义都在很大程度上取决于一个可能因素的组合爆炸式增长的集合。这些因素包括但不限于：

- 周围的句子和话语上下文 [16, 17]
- 代理的当前注意力集中和任务需求，可以选择性地突出特定的概念特征，[18–23]
- 代理的背景知识、文化环境 [24–26] ，
- 瞬时心理状态 [27] [1]
- 使用的特定语言，因为不同语言对语义空间的划分方式不同 [28] 。

这种深刻的语境依赖性意味着意义不仅仅是从具体单词中独立解码的，而是由一个解释的行为者在特定情境中主动构建或实现的— [29–31] 。相同的表达展现给不同的行为者---或者在不同条件下展现给同一行为者---可以产生不同的解释 [32, 33] 。关键的是，这突显了通过直接解释实现语义意义的观察者依赖性，提供了一种作用于状态的可观察量的量子力学类比，我们将在第 II 节更详细地探讨其影响。这种协同创造过程被认为是通过一种称为相关性实现（RR）的过程调节的——这是一种核心认知能力，使个人能够通过运用对语境敏感的注意机制来识别相关信息并过滤掉不相关信息，来有效地导航广阔的语义空间 [34–36] 。至关重要的是，RR 在标准计算意义上是非算法的 [36] ，因此更善于处理像问题定位和不确定搜索空间这样的挑战，而这些挑战对于纯粹在“小世界”中运作的形式系统而言是难以驾驭的 [37] 。这种具身的、观察者依赖的、语境化的意义构建观点挑战了语义表达拥有内在的、独立于语境或观察者的意义的任何概念，符合动态语义记忆模型 [13, 18, 29, 32, 38–41] 。

这一动态过程的固有不确定性和深层次的上下文依赖性表明，经典的概率和逻辑框架是不足的。因此，研究人员转向了非经典框架——例如那些采用量子理论原理的框架——以寻找更合适的数学工具。此类方法已被用于建模广泛的认知现象，包括概念组合、决策制定和记忆 [42–48] 。这些受量子启发的模型的实用性不仅仅是理论上的，还在实证研究中得到了证明。例如，Aerts *et al.* [49] 在分析关于“宠物鱼问题（Pet-Fish problem）”的概念共现统计时，识别出了量子似的上下文效应，表明在此规模上的意义构建偏离了经典概率假设。同样地，其他研究将贝尔定理 [50] 应用于人类认知，揭示了非经典相关性，这种相关性在信息检索判断 [51] 和认知决策任务 [52–54] 中违反了经典界限。

最近的研究通过区分两种类型的情境影响提供了一个关键的概念改进。在面部特征判断的实验中，Bruza *et al.* [55] 区分了情境敏感性，即情境的标准因果影响，以及真正的情境性。后者被定义为一种非因果形式的情境依赖，其中某个特性在测量之前可能是真正不确定的。他们认为，如果发现认知现象是情境性的，那么潜在的认知特性就不具有明确定义的、先存的值。相反，该特性在判断时刻被实际化——一种非经典模型独特地能够形式化的现象。

鉴于当前的状况，我们在这项工作中旨在实现两个主要任务：（1）识别阻碍前沿大型语言模型表观进展的组合问题；（2）提供一种非经典框架以理解和研究自然语言的实用路径。为此，在第 I 节中，我们从信息理论的角度探讨语义简化在单回合问题解决任务中的作用。然后，在第 II 节中，我们为解释行为制定量子语义理论框架。接下来，我们在第 III 节中详细介绍我们的实验方法，以测试自然语言解释是否可能表现出非经典行为，并利用大型语言模型进行解释替代。第 IV 节展示了我们的研究结果，第 V 节讨论了它们对计算语言学未来及认知科学理解的广泛影响。

I. KOLMOGOROV 复杂性、语义退化和解释的挑战

Kolmogorov 复杂性（KC）[56] 的概念提供了一个强大的视角来理解自然语言解析的基本限制，这与期望能够可靠地为用户解决问题的大规模语言模型（LLM）特别相关。Kolmogorov 复杂性，$K(s)$ ，是一串 s 的一个计算机程序（在一个固定的通用描述语言中）的最短长度，该程序可以输出 s 。虽然 KC 严格适用于有限字符串，但其潜在原则——所需的最小描述信息量——可以扩展到语义领域。对于某些语义表达 S_E ，我们可以将 $K(M(S_E))$ 概念化为明确指定给定语义表达 S_E 的预期含义 $M(S_E)$ 所需的最少位数。此规范不仅要捕捉所涉及概念的身份，还要捕捉其精确的上下文细微差别以及将它们绑定为一个连贯整体的复杂关系网络。例如，指定" 我刚去动物收容所并带回了一只狗" 的单一连贯解释的 KC 很低，而对于像詹姆斯 · 乔伊斯的《芬尼根的守灵夜》这样复杂且相互关联的作品中的一段来说，

> 在拉图什的地方，我会执掌任何被拥有的敏感钱币，我会把它全部砸进成本价的次等私酒投资中。我用我那件幸运的旧外套和你背上那半盎司赌一把（如果夫人们剥去我的衣服可能会招来不列颠的冷弹），我是那个一定会让它像收银机一样赚钱的实干家，确信有锅在杆上。还有，凭着一人之鱼和十二人之雍，我会播下我的狂野梅树以收获丰盈的麦酒，给大量的麦草和蜂蜜酒，以及甜麦芽酒施肥，我会在紧要关头用我的魔术偶然成功，公正、自由和愉快地，以一个因素的身份在市场上腾飞。

[1] 虽然许多人认为大型语言模型只是完全与物理世界的物质现实脱节的统计机器，但现实世界中它们所处环境的物理变化或像宇宙射线这样的随机现象确实可能引发瞬态状态。此外，当一个像 Anthropic 的 Claude 这样的模型拥有年份时间的知识时，其法语名字以及 Claude 的训练和系统提示会将法国思维方式强化到模型上，以至于它的响应模式和投入精力的意愿会根据年份时间是否与法国假期重合而有所不同。

的 KC 会非常高，以至于需要大量来自上下文的约束才能消除该段落的" 预期含义" 的不明确性，可能需要一个比原始表达式长几个数量级的程序，反映了将其多方面模糊性解析为特定读法所需的大量信息。

因此，为了限制解释到与作者原意密切对齐的那些所需的信息负载 $K(M(S_E))$ ，会随着表达式 S_E 内概念和关系的数量和相互关联性而大幅增加。尽管组成概念 c_i 构成了基础，每个概念需要一定数量的比特进行情境实例化。这个线性关系本身可能是可处理的，但正是概念间的关系 $r_{i,j}$ 促成了预期意义的实现，而指定相互关系的约束需要指定 $O(N^2_{Concepts})$ 数量的比特，以便尝试消除潜在解读歧义并重现预期的解读。因此，一个表达式的 KC 可以写为一个不等式，其下限基于数量

$$K(M(S_E)) \geq \sum_{i}^{N} c_i + \sum_{i,j}^{N} \cdot c_{i,j}^2 \tag{1}$$

可见，$K(M(S_E))$ 至少随着表达式复杂性的增加而超线性增加。每个 $K(M(S_E))$ 比特代表一种信息特异性或语义选择点（例如，决定将 “bat” 解释为一种动物，而不是用于棒球的木棍），这些是解释者必须成功重构的。关键的是，在每个这样的决策点，都存在退化解决方案的潜在机遇，因为许多词汇有多重含义，因此我们为每个比特 d_b 指定了一些退化（注意，这实质上起到了错误率的作用，但我们以这种方式使用它以维持概念性）。因此，解释者正确地实际化所有 $K(M(S_E))$ 比特以达到精确预期意义的概率是正确获得每个比特的概率的乘积，

$$P(\text{perfect interpretation}) = \frac{1}{N!} \prod_{k=1}^{K(M(S_E))} (\frac{1}{d}) \tag{2}$$

。

假设每个位的平均简并度为 $\bar{d}_b$ ，我们在图 1 中用图形来说明这一关系。随着 $K(M(S_E))$ 在图 1 中增长，对于复杂性适中表达的完美（或足够相似）的解释概率呈指数级减少，迅速趋近于零。这个结果清晰地展示了语义退化的实际作用：围绕任何表达式 S_E 的备选和合理解释的组合爆炸。此时，有必要提及这种情况与统计力学中的概念有类似之处，其中 S_E 类似于微观状态的集合。在推断中甚至一个比特（自由度的约束）的错误会导致不同的微观状态，并且由于高维（$K(M(S_E))$），解释者几乎不可能复现构成该集合的特定微观状态（我们将在第 II 节中论证这些状态本身是先验不可知的），从而导致高 “语义熵”。这种基于 KC 的分析凸显了 NLP 系统的一个根本性限制，并解释了 LLM 辅助任务中需要对语义退化表达进行深入而明确理解或翻译时持续存在的困难：LLM 生成一个合理的意义——众多易访问的微观状态之一——但几乎从未生成唯一意图的那个。这个结果本身强调了超越训练人工智能系统以重点关注单次响应成功的必要性，并优先研究能够成功模拟自然语言同时能够动态更新和自我调整的替代模型。我们希望这里的这些教训以及量子语义方法能够提供一个更清晰的基础，以便未来的方法和模型能够在此基础上进行训练、测试和评估。

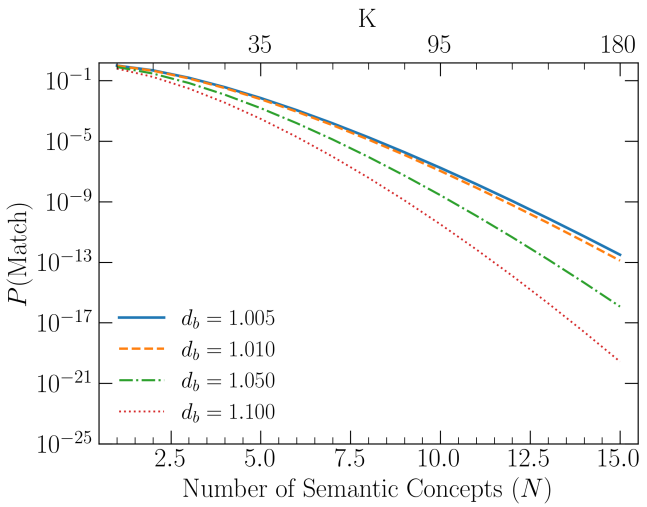

FIG. 1. 完美语义解释的概率与表达式中核心语义概念数量（$N_{concepts}$ ）之间的关系，说明了语义复杂度超线性增长的影响。总语义比特（K ，显示在上方 x 轴）被建模为单个概念所需信息和它们成对关系的信息总和，这里假设每个概念需要 $c_{concept} = 5$ 比特，每个关系需要 $c_{relationship} = 1$ 比特。不同曲线代表每比特平均错误概率（p_e ）的变化。随着 $N_{concepts}$ 的增加，K 的超线性增长导致完美解释概率以更快的指数速度下降，这凸显了随着概念间依赖关系的增加，语义退化带来的巨大挑战。

II. 量子语义学理论

这项工作的基本前提是，意义不是语义表达的内在静态属性，而是通过表达与位于特定背景中的解释代理之间的动态交互而实现的一种涌现现象。如所描述，这个框架自然挑战了在经典物理学中历史形成的现实主义假设。为了正式建模这种依赖观察者和上下文的意义性质，我们提出了一个量子语义框架，该框架反映了量子物理学与经典物理学之间的差异。语义表达被从语法中分离出来，并被视为可观测量，这与量子系统中物理测量结果与系统分离的方式相似，而不是假设经典系统中的现实主义。我们希望通过这种量子逻辑结构，更清楚地阐明诠释和意义实现的过程。

我们首先假设一个语义表达，记为 S_E ，没有预定义的固有意义。相反，当被代理接触时，它作为一个符号提供了一系列可能的解释。为了表示这种互动的能力，S_E 被关联到复杂希尔伯特空间 $\mathcal{H}_S$ 中的一个状态向量 $|\psi_{S_E}\rangle$ ，即语义状态空间。这个向量是基态 $|\psi_{S_E}\rangle = \sum_i c_i |e_i\rangle$ 的线性叠加。该提案的一个关键方面是，这些基态 $\{|e_i\rangle\}$ 本身并未被假定为已知或事先固定，也不一定可以解释为一组预定义的语义原语。相反，它们可能代表潜在语义差异化的抽象维度，仅通过特定的解释行为（测量）才变得操作上相关或部分可识别，其中 $\{|e_i\rangle\}$ 是可观察量 UTF8min ъ 的本征基集合——这里用一个日语字符表示，发音为 “mo” ——这由执行测量的代理设定，即解释语法。因此，希尔伯特空间 $\mathcal{H}_S$ 是一个形式构造，其基础由表达式可以独特涉及的所有可能方式的集合定义，而不是预先列举的基本意义列表。系数 c_i 是复数（$c_i \in \mathbb{C}$ ），由语义表达 S_E 如何可以作为语义基底 $|e_i\rangle$ 的叠加分解决定。虽然平方模 $|c_i|^2$ 关系到关于所选可观察量的测

量结果的概率，这些系数的复相位包含额外的信息，在经典概率模型中没有直接的对应，因此不可能从这些系数中推导出任何比喻的具体意义，就像在语言模型中考虑权重一样。实际上，复数性质对于在语义处理中建模潜在干涉和纠缠效应是基本的。c_i 的完整解释意义只有通过代理设定的相关解释轴 $|e_i\rangle$ 才能实现，即没有预存和孤立的意义概率的概念。

代理 A 通过一个解释性可观察量 $\hat{\text{UTF8min}}\ \text{ъ}_A(t)$ 来与 S_E 互动，这个可观察量是由其语义记忆（包括目标、人格、知识、注意状态）在当前上下文 $C_A(t)$ 的激活下动态构成的。通过应用一个厄米算子记为 $\hat{\text{UTF8min}}\ \text{ъ}_A(t)$ 到 $|\psi_{S_E}\rangle$，来表示为了确定特定语义方面的解释行为。此算子 $\hat{\text{UTF8min}}\ \text{ъ}_A(t)$ 体现了特定的语义探测。$\hat{\text{UTF8min}}\ \text{ъ}_A(t)$ 的特征值 $\{m_i\}$ 包含了代理 A 的语义测量的所有可能结果，具体来说，它们代表了所有可能的由代理实现的不同解释，即使这些解释不能被明确列举。实现解释 m_i 的概率因此由

$$P(m_i) = |\langle c_i|\psi_{S_E}\rangle|^2 = \langle\psi_{S_E}|\hat{P}_i|\psi_{S_E}\rangle \tag{3}$$

给出，其中 $\hat{P}_i = |e_i\rangle\langle e_i|$ 是到 m_i 特征空间的投影算子。这种互动类似于量子测量。

通过这种表述，不同主体的语义度量现在可以具有非交换性，类似于量子可观察量。如果两个不同的解释操作，$\hat{\text{UTF8min}}\ \text{ъ}_1$ 和 $\hat{\text{UTF8min}}\ \text{ъ}_2$，不交换，即 $[\hat{\text{UTF8min}}\ \text{ъ}_1, \hat{\text{UTF8min}}\ \text{ъ}_2] \neq 0$，那么它们探测到的语义方面通常不能同时拥有明确的、预先存在的值。

意义实现的动态性扩展到了解读过程的时间演化。代理的解读可观测量 $\hat{\text{UTF8min}}\ \text{ъ}(t)$ 和语义表达 $|\psi_{S_E}(t)\rangle$ 一般不是静态的。当 $\hat{\text{UTF8min}}\ \text{ъ}(t)$ 作为一个可观测量可以显式地随时间变化时，$|\psi_{S_E}(t)\rangle$ 的动态性可以类似于量子力学，由一个语义薛定谔方程来控制：

$$i\hbar_{sem}\frac{\partial}{\partial t}|\psi_{S_E}(t)\rangle = \hat{H}_{sem}(t)|\psi_{S_E}(t)\rangle \tag{4}$$

，其中语义哈密顿量 $\hat{H}_{sem}(t)$ 生成这种演化，封装了驱动因素，如语境中的变化 $C_A(t)$、来自 S_E 的顺序信息处理以及代理的内在认知动态。常数 $\hbar_{sem}$ 为这些语义动态设定了尺度，其中 $|e\rangle_i$ 之间的量子相干性很重要。这种形式主义允许代理的解释参与演化。$\hat{H}_{sem}(t)$ 对 t 的依赖性允许在代理浏览不同语义位置时，模拟这种演化的变化。然而，在这项工作中，我们将框架的时间演化部分搁置一旁，更加关注通过语义贝尔测试进行采样探索，从而验证语义表达 S_E 的量子态性质，因此采用了一种先前在认知心理学实验中使用的逻辑，以揭示人类判断中跨多个领域的非经典语境性，包括决策、信息检索和概念评估 [51–54] 。

III. 实验设计

本节概述了设计实验的方法，旨在测试语义理解中的非经典关联，类似于量子物理中的 CHSH 类型贝尔测试 [50, 57] 。特别是，该实验重点关注大型语言模型（LLM）代理在不同上下文（基于人物）的设置下如何解释简单句子结构中的模糊词对。

A. 大型语言模型作为观察者

在这项工作中，LLM 代理在该语义贝尔测试中充当“观察者”。为了减轻潜在的模型特定偏见并增强我们发现的稳健性，每个代理实例都是从一个预定义的多样化、最先进的基础模型和供应商池中随机选取的。该池包括诸如 Gemini 的各种版本（例如，‘gemini-1.5-flash’、‘gemini-2.0-flash-lite’、‘gemini-2.0-flash’）、Anthropic 的 Claude 系列（例如，‘claude-3-5-sonnet-latest’、‘claude-3-5-haiku-latest’ 和 ‘claude-3-7-sonnet-latest’）、DeepSeek 的 ‘deepseek-chat’，以及来自 OpenAI 的各种模型（例如，‘gpt-4o‘、‘gpt-4o-mini‘、‘gpt-4.1-mini‘、‘gpt-4.1-nano‘、‘gpt-4.1-nano’）。该方法符合多模型三角验证的建议，以确保结果的公平性和普遍性 [58–61] ，并确保我们发现的任何潜在相关性不仅限于单个模型的权重。在每个实验试验中，生成两个主要基础角色，“Alice” 和 “Bob”。这些角色的特点是随机分配的属性，如年龄（例：25-70 岁）和地点（例：印第安纳州布卢明顿；密歇根州底特律），这些属性隐式定义了他们的主要语言（在此设置中为英语）。这些属性为试验中的代理基础语义记忆配置文件提供信息。

值得注意的是，尽管大型语言模型因语义退化对复杂任务表现出局限性，其内部机制——尤其是注意力架构——却作为一种黑箱运行，这种黑箱类似于大脑将潜在解释状态坍缩为特定的状态并在回应用户询问时使用。虽然基础机制与人类语言解释的生物和认知基础当然是不同的，两者观察特定状态的相似性表明大型语言模型确实有效地重现了语言理解和认知的这一功能。因此，这些模型可以作为自然语言任务中的实验解释探针。此外，大型语言模型还被证明可以生成反应，这些反应可以在第一序模仿人类在各种情境中的语言行为（例如，调查例子 [62–64] [2] Kitadai *et al.* [62] ，另注意角色在改善回应真实性方面的力量，这是我们也利用的重要方面。因此，我们主张，人们可以并且应该使用大型语言模型来探索语义解释的统计模式在多种情况下（例如角色和上下文变化）下的表现。通过观察大型语言模型如何处理语义歧义和依赖上下文的意义，我们可以深入了解解释机制以及在处理具有高度语义退化的任务时哪些计算策略更成功，哪些则较不成功 [65, 66] 。

B. 贝尔测试

我们的语义贝尔测试的核心涉及向大型语言模型代理呈现包含两个模糊词语的句子。然后要求这些代理为句子中的每个词语选择一个单一且明确的意义。通过将模糊的词对（例如，将 “trunk” 与 “存储/树” 的意义 ‘A’ 和与船头/结的 “bow” 结合起来）嵌入简单的句子模板

[2] 这些参考文献中的最后一个，Tjuatja *et al.* [64] ，实际上将大语言模型在这方面的表现描述为较差，因为他们指出，这些模型容易受到干扰，而人类在类似的调查中不会受到如此强烈的影响。我们在此注意到，他们的发现很好地契合了这项工作的框架，并且如果我们进一步考虑人类主体（由于身体嵌入而具有丰富的上下文）和大语言模型（上下文贫乏且非身体化）之间相关性实现的作用，那么显然，他们实验中大语言模型回应的主要 “问题” 是由于严重的上下文劣势所导致的。

Experiment	N Trials	|S|
1	5	2.8
2	5	1.2
3	10	2.0
4	10	2.44
5	20	2.32
6	20	2.0
7	50	2.33
8	200	1.83

TABLE I. 根据不同实验运行计算的 CHSH 不等式的 $|S|$ 值，在人物配置（年龄，位置）以及每个实验的总试验次数（N）之间有轻微的差异。值 $|S| > 2$ 表示违反了经典的 Bell-CHSH 不等式。

中（例如，“The word1 was settled near the word2 ”）来构建刺激。为艾丽斯（A, A’）和鲍勃（B, B’）代理定义了四种不同的解释设置，这些设置通过给他们的基础角色提供附加的、不同的简短文本提示来创建，旨在为语义上正交的不同背景视角做准备（例如，“你是一名外科医生...”与“你是一名公交司机...”）。对于这四种情况中的每一种，相应的 LLM 代理为两个模糊词语提供一个单一的同时解释。艾丽斯和鲍勃并未被告知待考虑的定义，以避免影响他们或限制语义搜索空间，因此需要单独的 LLM 调用来确定每个解释是否与预定义的意义‘α’或‘β’对齐，如果选择不明确，则会触发重新解释（例如，如果对“chair”的意义选择是“团体领导者”或“用于坐着的家具”，而解释为“家具或领导者”）或超出这些选择范围（例如，对于该椅子的例子，如果认为句子指的是处决用的“电椅”，则被认为超出了定义范围，并进行重新解释）。这些分类解释被映射到数值（A $\to +1$, B $\to -1$ ），为每个设置生成一个 2 元素的结果向量。最后，结果向量（$\widehat{\text{UTF8min}}_{A}$, $\widehat{\text{UTF8min}}_{A'}$, $\widehat{\text{UTF8min}}_{B}$, $\widehat{\text{UTF8min}}_{B'}$ ）被规范化，期望值 $E(XY)$ 通过相应对（α , β ）的平均点积计算得出，然后用于计算 CHSH S 值 $S = E(A, B) - E(A, B') + E(A', B) + E(A', B')$ 。然而，此标准 CHSH 公式中的一个关键假设是边缘一致性，也称为“无信号传递”[67, 68]，这假定一个代理的结果的边缘概率与另一个代理的测量设置无关。正如 Dzhafarov *et al.* [67] 广泛讨论的那样，这一假设在认知和行为实验中经常被违反。此类违反称为“不一致的连通性”，可能会使贝尔型不等式的解释复杂化。因此，虽然 $|S| > 2$ 的结果表明局部现实主义的违背，但在解释时必须谨慎，因为过度的相关性可能源于这些直接的上下文影响，而不是来源于真正的上下文性。

所有代理定义和 LLM 响应处理都是使用开源的 `npcpy` 软件包 [3] 进行的。

IV. 结果

在这个简短的部分中，我们重点介绍并讨论了语义贝尔测试的结果。

在我们的实验中，我们进行了 8 次实验，其中我们将单个实验中的试验次数从 5 到 200 进行变化，并实例化不同的人格组合。我们在表 I 中展示了实验的各种 $|S|$ 值。为了说明，我们在图 2 中展示了试验 # 7 中 |S| 值随试验次数的演变，其中有 $N = 50$ 次试验。需要明确的是，$|S|$ 随着试验次数的变化并没有什么特别的意义，因为 $|S|$ 值本身应该从无限数量的试验中推导出来。该示例仅作为一种视觉辅助，展示了对于相当数量的试验，$|S|$ 值的稳定过程。

重要的是，在这些结果中，我们发现了一种丰富多样的相关结构，范围从经典行为到非经典量子逻辑。理论上，在贝尔实验中量子系统的上限为 $|S| \leq 2\sqrt{2}$ 。有趣的是，我们的一次实验（尽管是 $N = 5$ ），$|S|$ 值达到了 2.8，这为这些实验提供了一个与量子计算框架一致的经验上限。

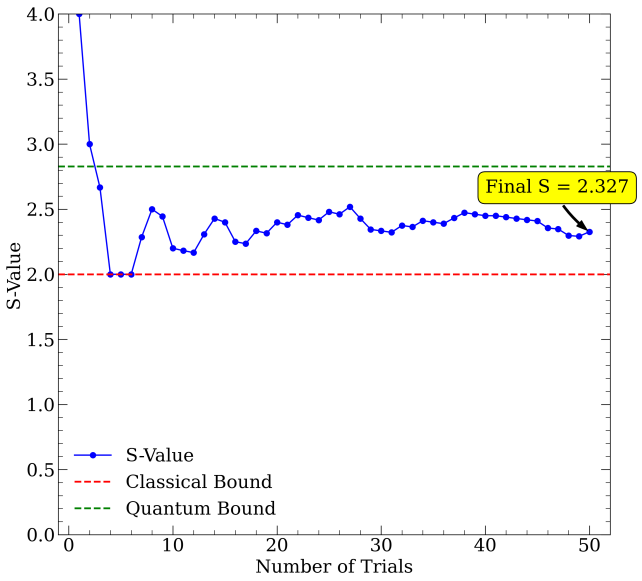

FIG. 2. 对于一项总共进行了 50 次试验的实验，CHSH S 值计算的示例演变随着试验次数的进展而变化，最终稳定在 2.327（表 I 中的实验 # 7）。虚线红线在 $S = \pm 2$ 处表示 CHSH 不等式的经典界限，虚线绿线表示在 S= $2\sqrt{2}$ 处的提出的量子界限。

V. 讨论

本研究中提出的论点，关于语义退化所施加的基本限制以及量子语义框架的潜力，为重新评估自然语言处理和认知科学中若干核心争论和现行方法提供了新的视角。本文将首先讨论语义退化的信息论挑战（由 Kolmogorov 复杂性量化）如何影响分布式语义模型（DSMs）和大型语言模型（LLMs）在需要深度情境理解的任务中的能力（第 V A 节）。随后，我们将探讨量子语义框架如何强调观察者依赖的意义实现，并如何与经典意义和解释理论交叉，以及提供替代方案（第 V B 节）。然后，我们将这些思路与具体的认知现象和心理语言学研究结果联系起来（第 ?? 节），最后考虑其对 NLP 研究（第 ?? 节）和实际工业应用的广泛影响（第 ?? 节）。

[3] https://github.com/NPC-Worldwide/npcpy

A. 语义退化：对 DSM 和 LLM 的影响

分布式语义模型，包括当前的大型语言模型（LLMs），通过从庞大的文本语料库中学习统计共现，实现了显著的成功 [11, 69–71] 。这些模型在理论上假设，足够的统计曝光可以导致稳健的语义表示。然而，语义退化原则，特别是通过第 I 节中详述的 Kolmogorov 复杂性（KC）进行分析时，揭示了对于需要精确、上下文特定解释的任务而言，这一假设存在根本性挑战，因为我们已经表明，随着语义表达的复杂性及其所需的上下文消歧的增加，指定单一预期意义的 KC 增长超线性。这意味着任何基于有限数据集和固定权重训练的 DSM 或 LLM——在某些关键方面更有可能提供与“预期”不一致的解决方案和解释，这导致了理解的完全崩溃。这些信息理论上的限制源于潜在解释的组合爆炸，能够充分解释 LLM 在复杂推理任务上的性能停滞，以及尽管模型大小和数据增加，DSM 面对模糊或上下文丰富文本所面临的持续困难 [1, 2] 。事实上，似乎在处理高模糊性任务时，DSM 的限制可能阻止它们达到“强”人工智能的地位。很可能替代方法或集成方法（例如，LLM 被用作推理引擎以及替代、更动态的模型）将占上风。

B. 量子语义与经典意义理论

语义退化所突显的局限性激发了对意义替代框架的探索。我们提出的量子语义框架（见第 II 节），认为意义不是文本的内在属性，而是通过观察者依赖的解读行为实现的，这直接挑战了传统的、常隐含于 DSM 和传统语言理论中的语义实在和局部性假设。经典实在主义假设意义是预先存在且明确的，而局部性则假设语义成分可以独立地确定。相比之下，量子语义方法将表达视为提供了一系列潜在的解释（即叠加态），具体的意义通过在特定背景下代理的解读“测量”而“坍缩”或实现。这与认知科学中的建构主义理论相一致，这些理论认为理解是一个主动的、情境化的过程 [13, 17, 72] 。对意义本质主义观点的哲学批判 [例如，见后 **?**] 也与我们的结论相呼应，因为意义与使用和互动相关，而不是静态地存在于符号中。解读操作的非交换性——量子语义模型的一个特点——意味着情境探查的顺序可能影响实现的意义，这种现象在纯粹的经典加性模型中难以捕捉，但在人的认知中被观察到，例如判断顺序效应 [73] 和类似于量子态的情境效应，如“宠物鱼”问题 [49] 。

依赖于观察者的实际化在量子语义中是核心，这为一系列既定的认知和心理语言学发现提供了令人信服的解释框架。如果意义不是固定的，而是在互动中实现的，那么代理人的当前上下文、目标和被启动的语义记忆对解释的深刻影响就能自然地被容纳。这种量子“测量上下文性”的概念与动态语义记忆的模型、注意力集中度的解释、关联实现的非算法性质，甚至语义分类的跨语言差异相一致。

这种测量上下文性的概念在研究人类判断的过程中找到了强有力的实证类比。例如，Bruza *et al.* [46, 55] 在面部判断方面的研究将这种上下文性从语言扩展到感知，认为这些属性在判断的瞬间是未定的且是构建出来的。这强化了我们的核心前提：如果具体的感知特征是未定的，那么抽象的语义意义更有可能需要通过解释行为来实现。我们的实验结果提供了直接的定量支持来证明这一观点。我们观察到的 CHSH 不等式的违反在量级上与一些人类认知实验的结果相当，例如 Aerts *et al.* [53] 关于概念纠缠的研究。这一相似之处表明，LLMs 利用了人类语义处理中固有的同样非经典概率结构。

这些研究结果的汇聚将我们的 LLM 实验定位为一种新的方法，用于调查语义上下文性本身的基本性质。如此看来，LLM 作为复杂的“计算认知系统”，能够在集体人类数据中的宏观统计现象与单个主体的微观解释行为之间架起桥梁。因此，我们观察到的上下文性表明，这种非经典行为不是人类心理或某种特定 LLM 架构的怪癖，而是任何复杂、互联系统中语义意义的结构和处理的普遍特征。

双重论点——语义简并对经典系统施加的限制以及量子语义框架的潜力——强烈建议自然语言处理方法应向非经典的、贝叶斯信息的方法转变。与其追求单一、明确的解释，不如采用在不同语境条件下的解释蒙特卡罗采样技术，结合语义空间的动态探索（例如，通过马尔可夫随机游走），可能提供更实用和稳健的文本特征。这特别适用于当前大语言模型面临本质难题的任务，如细微的翻译、创意发现和复杂的单回合完成，因其受到柯尔莫哥洛夫复杂性挑战的影响；一个旨在探索广泛的可能实现分布的系统，能够有效地从问题到解决方案的多种潜在路径，尽管较慢，但在近似理解语义表达需求方面可能更有效。采用这样一个贝叶斯启发的视角使得系统可以将歧义视为语义景观中固有且有信息特征，而不是要消除的错误。这种方法直接解决了语义简并造成的计算难题，提供了一条实用路径，以构建更加具有弹性和情境感知的语言技术，更好地反映意义本身的概率性质。

将这些概念转变为实际应用，尤其是在诸如多代理系统或需要动态文档理解的大规模企业环境等复杂场景中，需要开发新颖且可适应的系统架构，以帮助模型接近智能。这些架构必须能够管理和利用情境变化，而不是试图消除它。关键的是，量子语义框架提出的意义实现的观察者依赖性，强调了人类参与系统（HITL）持续且根本的重要性。HITL 远非在 AI 实现完美自主性之前的临时措施，而是在导航固有的语义模糊性、验证解释、确保系统输出与人类目标和伦理考虑保持一致方面，将始终是不可或缺的一部分，特别是在诸如医疗保健、国防和金融等安全关键领域。明确认识到纯自动解释在开放环境中的根本局限性，以及基于情境感知框架的替代方案潜力，可以指导开发更现实、更健壮、最终更具能力的语言技术。

VI. 结论

在这项工作中，我们使用了基于 Kolmogorov 复杂性理论框架和一种新颖的实验设计，采用大型语言模型代理来研究语义意义的基本、非经典本质。我们通过识别任何语言解释行为中固有的信息理论极限来进行分析，并提供了对不同 LLM 驱动的 AI 代理解释行为中的上下文性的首次已知测试。我们的主要结论如下：

1. 语义退化是自然语言的一个基本特性，它在解释中施加了信息论的限制；我们使用 Kolmogorov 复杂性（I 节）的分析形式化地表明，这使得从复杂表达式中恢复单一意图的意义对于任何系统来说在计算上都是不可行的，从而为在 LLM 中观察到的性能平台提供了明确的解释。

2. 在我们的语义贝尔测试实验中使用 LLM 代理时，由于频繁且显著违反 CHSH 不等式（$|S| > 2$），在歧义下的语言解释表现出了非经典的上下文性（见章节 III，IV）。

3. 在大型语言模型（LLMs）诠释行为中测量的语境性与人类认知科学中一系列非经典发现的更广泛模式一致，表明观察者依赖性和不确定性是信息处理的一般原则，而不仅仅是人类心理的产物。

4. 由我们的实验验证的、意义的观察者依赖性质表明，没有绝对的、根本的意义可以发现，只有情境化的解释。因此，唯一可行的科学方法论是从寻找任何单一的“正确”答案转变为使用重复的贝叶斯风格的采样来描绘这些条件解释如何在一个可能性空间中互相连接。

5. 当考虑到在人类认知中具有相似发现的情况下，在各种非生物大语言模型代理中一致出现的非经典情境性，表明这些统计特性并不是任何特定解释系统的产物，而是自然语言本身的客观、结构性特征。

[1] M. N. Jones, When does abstraction occur in semantic memory: Insights from distributional models, Language, Cognition and Neuroscience **34**, 1338 (2018).

[2] P. Hoffman, M. A. Lambon Ralph, and T. T. Rogers, Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words, Behavior Research Methods **45**, 718 (2013).

[3] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, Self-supervised learning from images with a joint-embedding predictive architecture, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023) pp. 15619–15629.

[4] A. Gu and T. Dao, Mamba: Linear-Time Sequence Modeling with Selective State Spaces, arXiv e-prints , arXiv:2312.00752 (2023), arXiv:2312.00752 [cs.LG].

[5] L. Darlow, C. Regan, S. Risi, J. Seely, and L. Jones, Continuous Thought Machines, arXiv e-prints , arXiv:2505.05522 (2025), arXiv:2505.05522 [stat.ML].

[6] Q. Sun, E. Cetin, and Y. Tang, Transformer-Squared: Self-adaptive LLMs, arXiv e-prints , arXiv:2501.06252 (2025), arXiv:2501.06252 [cs.LG].

[7] T. Simonds and A. Yoshiyama, LADDER: Self-Improving LLMs Through Recursive Problem Decomposition, arXiv e-prints , arXiv:2503.00735 (2025), arXiv:2503.00735 [cs.LG].

[8] P. Tabossi, L. Colombo, and R. Job, Accessing lexical ambiguity: Effects of context and dominance, Psychological Research **49**, 161 (1987).

[9] K. S. Binder and K. Rayner, Contextual strength does not modulate the subordinate bias effect: Evidence from eye fixations and self-paced reading, Psychonomic Bulletin & Review **5**, 271 (1998).

[10] K. Rayner, A. E. Cook, B. J. Juhasz, and L. Frazier, Immediate disambiguation of lexically ambiguous words during reading: Evidence from eye movements, British Journal of Psychology **97**, 467 (2006).

[11] Z. S. Harris, Distributional Structure, WORD **10**, 146 (1954).

[12] S. A. McDonald and R. C. Shillcock, Rethinking the Word Frequency Effect: The Neglected Role of Distributional Information in Lexical Processing, Language and Speech **44**, 295 (2001).

[13] L. W. Barsalou, Situated simulation in the human conceptual system, Language and Cognitive Processes **18**, 513 (2003).

[14] W. Yeh and L. W. Barsalou, The Situated Nature of Concepts, The American Journal of Psychology **119**, 349 (2006).

[15] E. T. Higgins, Knowledge activation: Accessibility, applicability, and salience, in *Social Psychology: Handbook of Basic Principles*, edited by E. T. Higgins and A. W. Kruglanski (Guilford Press, 1996) pp. 133–168.

[16] K. Bicknell, J. L. Elman, M. Hare, K. McRae, and M. Kutas, Effects of event knowledge in processing verbal arguments, Journal of Memory and Language **63**, 489 (2010).

[17] W. Kintsch, Predication, Cognitive Science **25**, 173 (2001).

[18] L. A. M. Lebois, C. D. Wilson-Mendenhall, and L. W. Barsalou, Are Automatic Conceptual Cores the Gold Standard of Semantic Processing? The Context-Dependence of Spatial Meaning in Grounded Congruency Effects, Cognitive Science **39**, 1764 (2015).

[19] E. Yee, S. Z. Ahmed, and S. L. Thompson-Schill, Colorless green ideas (can) prime furiously, Psychological Science **23**, 364 (2012).

[20] K. Hoenig, E.-J. Sim, V. Bochev, B. Herrnberger, and M. Kiefer, Conceptual Flexibility in the Human Brain: Dynamic Recruitment of Semantic Maps from Visual, Motor, and Motion-related Areas, Journal of Cognitive Neuroscience **20**, 1799 (2008).

[21] S. Van Dantzig, D. Pecher, R. Zeelenberg, and L. W. Barsalou, Perceptual Processing Affects Conceptual Processing, Cognitive Science **32**, 579 (2008).

[22] C. Bermeitinger, D. Wentura, and C. Frings, How to switch on and switch off semantic priming effects: Activation processes in category memory depend on focusing specific feature dimensions, Psychonomic Bulletin & Review **18**, 579 (2011).

[23] N. S. Hsu, D. J. M. Kraemer, R. T. Oliver, M. L. Schlichting, and S. L. Thompson-Schill, Color, Context, and Cognitive Style: Variations in Color Knowledge Retrieval as a Function of Task and Subject Variables, Journal of Cognitive Neuroscience **23**, 2544 (2011).

[24] P. Athanasopoulos, Cognitive representation of colour in bilinguals: The case of Greek blues, Bilingualism: Language and Cognition **12**, 83 (2009).

[25] B. Thompson, S. G. Roberts, and G. Lupyan, Cultural influences on word meanings revealed through large-scale semantic alignment, Nature Human Behaviour **4**, 1029 (2020).

[26] B. T. Johns, Distributional social semantics: Inferring word meanings from communication patterns, Cognitive Psychology **131**, 101441 (2021).

[27] J. Cesario, J. E. Plaks, N. Hagiwara, C. D. Navarrete, and E. T. Higgins, The Ecology of Automaticity: How

Situational Contingencies Shape Action Semantics and Social Behavior, Psychological Science **21**, 1311 (2010).
[28] V. Marian and M. Kaushanskaya, Language context guides memory content, Psychonomic Bulletin & Review **14**, 925 (2007).
[29] L. W. Barsalou, Simulation, situated conceptualization, and prediction, Philosophical Transactions of the Royal Society B: Biological Sciences **364**, 1281 (2009).
[30] D. Pecher, R. Zeelenberg, and J. G. W. Raaijmakers, Does Pizza Prime Coin? Perceptual Priming in Lexical Decision and Pronunciation, Journal of Memory and Language **38**, 401 (1998).
[31] A. M. Borghi, A. M. Glenberg, and M. P. Kaschak, Putting words in perspective, Memory & Cognition **32**, 863 (2004).
[32] L. Connell and D. Lynott, Principles of Representation: Why You Can't Represent the Same Concept Twice, Topics in Cognitive Science **6**, 390 (2014).
[33] E. Musz and S. L. Thompson-Schill, Semantic variability predicts neural variability of object concepts, Neuropsychologia **76**, 41 (2015).
[34] J. Vervaeke and L. Ferraro, Relevance, meaning and the cognitive science of wisdom (2013).
[35] B. P. Andersen, M. Miller, and J. Vervaeke, Predictive processing and relevance realization: Exploring convergent solutions to the frame problem, Phenomenology and the Cognitive Sciences , 1 (forthcoming).
[36] J. Jaeger, A. Riedl, A. Djedovic, J. Vervaeke, and D. Walsh, Naturalizing relevance realization: Why agency and cognition are fundamentally not computational (2023).
[37] J. Vervaeke and L. Ferraro, Relevance realization and the neurodynamics and neuroconnectivity of general intelligence, in *SmartData*, edited by I. Harvey, A. Cavoukian, G. Tomko, D. Borrett, H. Kwan, and D. Hatzinakos (Springer New York, New York, NY, 2013) pp. 57–68.
[38] D. L. Hintzman, MINERVA 2: A simulation model of human memory, Behavior Research Methods, Instruments, & Computers **16**, 96 (1984).
[39] R. K. Jamieson, J. E. Avery, B. T. Johns, and M. N. Jones, An Instance Theory of Semantic Memory, Computational Brain & Behavior **1**, 119 (2018).
[40] W. O. Van Dam, S.-A. Rueschemeyer, O. Lindemann, and H. Bekkering, Context Effects in Embodied Lexical-Semantic Processing, Frontiers in Psychology **1**, 150 (2010).
[41] W. O. Van Dam, M. Van Dijk, H. Bekkering, and S.-A. Rueschemeyer, Flexibility in embodied lexical-semantic representations, Human Brain Mapping **33**, 2322 (2012).
[42] D. Aerts, Quantum structure in cognition, Journal of Mathematical Psychology **53**, 314 (2009), special Issue: Quantum Cognition.
[43] P. Bruza, J. R. Busemeyer, and L. Gabora, Introduction to the special issue on quantum cognition, Journal of Mathematical Psychology **53**, 303 (2009), special Issue: Quantum Cognition.
[44] L. Gabora and D. A. and, Contextualizing concepts using a mathematical generalization of the quantum formalism, Journal of Experimental & Theoretical Artificial Intelligence **14**, 327 (2002), https://doi.org/10.1080/09528130210162253.
[45] P. Bruza, K. Kitto, B. Ramm, L. Sitbon, and D. Song, Quantum-like non-separability of concept combinations, emergent associates and abduction, Logic Journal of the IGPL **20**, 445 (2012).
[46] P. Bruza, K. Kitto, B. Ramm, and L. Sitbon, A probabilistic framework for analysing the compositionality of conceptual combinations, Journal of Mathematical Psychology **67**, 26 (2015).
[47] J. M. Yearsley and J. R. Busemeyer, Quantum cognition and decision theories: A tutorial, Journal of Mathematical Psychology **74**, 99 (2016), foundations of Probability Theory in Psychology and Beyond.
[48] E. M. Pothos and J. R. Busemeyer, Quantum cognition, Annual Review of Psychology **73**, 749 (2022).
[49] D. Aerts, M. Czachor, B. D'Hooghe, S. Sozzo, *et al.*, The pet-fish problem on the world-wide web., in *AAAI Fall Symposium: Quantum Informatics for Cognitive, Social, and Semantic Processes* (2010).
[50] J. S. Bell, On the einstein podolsky rosen paradox, Physics Physique Fizika **1**, 195 (1964).
[51] S. Uprety, Investigation and modelling of quantum-like user cognitive behaviour in information access and retrieval (2020).
[52] D. Aerts, S. Aerts, J. Broekaert, and L. Gabora, The violation of bell inequalities in the macroworld, Foundations of Physics **30**, 1387 (2000).
[53] D. Aerts, J. A. Arguëlles, L. Beltran, S. Geriente, M. S. de Bianchi, S. Sozzo, and T. Veloz, Spin and wind directions i: Identifying entanglement in nature and cognition, Foundations of Science **23**, 323 (2018).
[54] D. Aerts, J. A. Arguëlles, L. Beltran, S. Geriente, M. Sassoli de Bianchi, S. Sozzo, and T. Veloz, Spin and wind directions ii: A bell state quantum model, Foundations of Science **23**, 337 (2018).
[55] P. Bruza, L. Fell, P. Hoyte, S. Dehdashti, A. Obeid, A. Gibson, and C. Moreira, Contextuality and context-sensitivity in probabilistic models of cognition, Cognitive Psychology **140**, 101529 (2023).
[56] A. N. Kolmogorov, Three approaches to the quantitative definition of information, Problems of Information Transmission **1**, 1 (1965).
[57] J. F. Clauser, M. A. Horne, A. Shimony, and R. A. Holt, Proposed experiment to test local hidden-variable theories, Phys. Rev. Lett. **23**, 880 (1969).
[58] W. Guo and A. Caliskan, Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases, in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (2021) pp. 122–133.
[59] B. K. Payne, H. A. Vuletich, and K. B. Lundberg, The Bias of Crowds: How Implicit Bias Bridges Personal and Systemic Prejudice, Psychological Inquiry **28**, 233 (2017).
[60] Z. Siddique, I. Khalid, L. D. Turner, and L. Espinosa-Anke, Shifting perspectives: Steering vector ensembles for robust bias mitigation in llms, arXiv preprint arXiv:2503.05371 (2025).
[61] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, and N. K. Ahmed, Bias and fairness in large language models: A survey, arXiv preprint arXiv:2309.00770 (2023).
[62] A. Kitadai, K. Ogawa, and N. Nishino, Examining the feasibility of large language models as survey respondents, in *2024 IEEE International Conference on Big Data (BigData)* (2024) pp. 3858–3864.
[63] A. Salecha, M. E. Ireland, S. Subrahmanya, J. Sedoc, L. H. Ungar, and J. C. Eichstaedt, Large language

models display human-like social desirability biases in big five personality surveys, PNAS Nexus **3**, pgae533 (2024), https://academic.oup.com/pnasnexus/article-pdf/3/12/pgae533/61188312/pgae533.pdf.

[64] L. Tjuatja, V. Chen, T. Wu, A. Talwalkwar, and G. Neubig, Do llms exhibit human-like response biases? a case study in survey design, Transactions of the Association for Computational Linguistics **12**, 1011 (2024), https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00685/2468689/tacl_a_00685.pdf.

[65] S. T. Piantadosi and F. Hill, Meaning without reference in large language models, arXiv preprint arXiv:2208.02957 (2022).

[66] A. Lampinen, I. Dasgupta, S. Chan, K. Mathewson, M. Tessler, A. Creswell, J. McClelland, J. Wang, and F. Hill, Can language models learn from explanations in context?, in *Findings of the Association for Computational Linguistics: EMNLP 2022*, edited by Y. Goldberg, Z. Kozareva, and Y. Zhang (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022) pp. 537–563.

[67] E. N. Dzhafarov, J. V. Kujala, and V. H. Cervantes, Contextuality-by-default: a brief overview of ideas, concepts, and terminology, in *Quantum Interaction: 9th International Conference, QI 2015, Filzbach, Switzerland, July 15-17, 2015, Revised Selected Papers 9* (Springer, 2016) pp. 12–23.

[68] V. H. Cervantes and E. N. Dzhafarov, Advanced analysis of quantum contextuality in a psychophysical double-detection experiment, Journal of Mathematical Psychology **79**, 77 (2017).

[69] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, Distributed Representations of Words and Phrases and their Compositionality, arXiv preprint arXiv:1310.4546 (2013), arXiv:1310.4546 [cs.CL].

[70] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv Preprint arXiv:1810.04805 (2018), arXiv:1810.04805.

[71] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, Language models are few-shot learners, in *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Curran Associates, Inc., 2020) pp. 1877–1901.

[72] J. L. Elman, Finding Structure in Time, Cognitive Science **14**, 179 (1990).

[73] R. M. Hogarth and H. J. Einhorn, Order effects in belief updating: The belief-adjustment model, Cognitive Psychology **24**, 1 (1992).