arXiv:2304.01083v1 [cs.CL] 3 Apr 2023

大语言模型的推理逻辑能否解耦为符号概念?

Wen Shen Lei Cheng YuxiaoYang MingjieLi Quanshi Zhang*上海交通大学

摘要

本文将大语言模型（LLM）的推理逻辑解释为一组符号概念。近期的许多研究 [4, 9, 10] 发现，传统深度神经网络通常会编码稀疏符号概念。然而，由于大语言模型的参数远多于传统深度神经网络，它是否也会编码稀疏符号概念仍是一个开放问题。因此，本文提出将大语言模型在对话任务中的推理分数解耦为少量符号概念。我们验证了这些稀疏概念能够很好地估计大语言模型在输入句子的所有任意掩蔽状态下的全部推理分数。我们还评估了大语言模型所编码概念的可迁移性，并验证了符号概念通常在相似的输入句子之间具有很高的可迁移性。更重要的是，这些符号概念可用于解释导致大语言模型预测错误的确切原因。

1. 引言

近年来，大语言模型（LLM），例如，ChatGPT [8]和 GPT-3 [1], 展现出了卓越的性能。尽管大语言模型已被用于指导人们完成撰写论文、分析报告和咨询信息等重要任务，但人们仍无法充分理解大语言模型的推理逻辑。

因此，为大语言模型的推理分数生成可问责的 解释，对于高风险决策至关重要。一般来说，可问责的解释需要满足三个要求。

• **可计数的推理模式**。 将大语言模型隐含的推理逻辑转化为明确且可计数的推理模式，对于解释大语言模型至关重要。具体而言，研究发现，深度神经网络（DNN）对输入句子的推理分数可以分解为少量推理模式（例如，某些短语）的影响 [9, 7]。这些推理模式可以视为大语言模型记忆的概念，并且应当能够普遍解释其在各种句子上的推理分数。

• **经实践验证**。 当将大语言模型的推理逻辑解释为一组推理模式时，解释的忠实性或准确性必须既有理论保证，也经过实验验证。

• **错误问责**。 当大语言模型给出错误预测时，重要的是明确指出导致错误的确切原因，并建立该原因与错误预测之间清晰的数学联系，而不是提供似是而非的解释。

深度神经网络中概念的定义。考虑到上述三项要求，Ren 等人 [9] 尝试将神经网络的推理逻辑解释为对输入变量之间显著交互作用的检测。具体而言，他们发现，少量输入变量之间的显著交互作用就能很好地模拟网络的推理分数。给定一个包含 n 个词、词索引为 $N = \{1, 2, ..., n\}$ 的输入句子，深度神经网络不会直接使用单个输入词进行推理，而是将不同词之间的交互作用编码为推理模式（或概念），以生成推理分数 v。例如，在分类任务中，给定输入句子“他是个新手”，词语 $S = \{$新手，新手$\} \subseteq N$ 之间的交互作用会产生特定的交互效应 $I(S)$，使网络的推理分数倾向于将“他”推断为“新手”这一含义。对 S 中的任意词进行掩码都会消除这一交互效应，即，使 $I(S) = 0$。

*Quanshi Zhang 是通讯作者。他任职于中国上海交通大学约翰·霍普克罗夫特计算机科学中心的计算机科学与工程系。 < zqs1022@sjtu.edu.cn>

研究发现，在所有 2^n 交互作用中，深度神经网络通常只编码少量显著交互作用，其他交互作用的影响都可以忽略 $I(S) \approx 0$ [9, 7]。因此，我们可以将这些显著交互作用视为**概念**(即可计数的推理模式），并将其余影响可忽略的交互作用视为噪声模式。受此启发，我们可以将深度神经网络的推理逻辑解释为对少量概念的检测。换言之，**我们可以利用少量显著交互概念，准确估计深度神经网络在指数数量的随机掩码句子上的推理分数**。更重要的是，当深度神经网络做出错误预测时，将交互概念从网络中清晰地解耦出来，有助于我们确定网络所编码的、**导致错误的确切交互概念**。此外，这些概念在不同样本和不同深度神经网络之间也表现出很强的可迁移性，并在分类任务中展现出相当强的判别能力 [7]。

使用概念解释大语言模型。 然而，我们以上关于使用稀疏概念来模拟传统深度神经网络逻辑的发现，都是基于实验中的经验观察，尚无扎实的证明。另一方面，大语言模型通常比传统深度神经网络拥有更多参数，而且训练数据也多得多。因此，大语言模型能否被解释为对少数概念的检测，目前仍不得而知。

因此，本文旨在将大语言模型针对对话的推理分数拆解为少数显著的交互概念。给定一个输入句子，例如一个问题或一个未完成的句子，大语言模型会依次生成词语来回答问题或补全句子。考虑下一个词的生成过程，大语言模型将前面的 n 个词作为输入，然后估计生成第（$n+1$）个词的概率分布。大语言模型会采样概率最高的词作为目标词 y^{truth}。因此，我们将 $v(\boldsymbol{x}) = \log p(y = y^{\text{truth}}|\boldsymbol{x})/(1 - p(y = y^{\text{truth}}|\boldsymbol{x})) \in \mathbb{R}$ 设为大语言模型的推理分数，其中 $p(y = y^{\text{truth}}|\boldsymbol{x})$ 表示生成第（$n+1$）个目标词 y^{truth} 的概率。具体而言，我们得出以下三个结论。

1. 我们已验证，给定一个输入句子，可以使用少量概念，准确估计大语言模型在该输入句子的所有任意掩蔽状态下的全部推理分数。对于一个经过充分训练、用于对话的大语言模型，我们量化输入句子中不同词语之间的交互作用，并验证生成特定词语的输出分数[1]可以通过少数交互概念得到准确估计。此外，给定一个部分词语被随机掩码的输入句子，我们仍然可以使用这些交互概念，准确估计生成下一个词语的输出分数。
2. 我们已验证，大语言模型编码的交互概念在相似输入句子之间具有很高的可迁移性。换言之，面对相似的输入句子，大语言模型通常会使用相似的交互概念来生成相同的词语。3. 我们已验证，解耦交互概念可以帮助人们找出大语言模型产生预测错误的确切原因。当大语言模型生成了违反事实的错误词语时，我们可以找出导致该错误词语的交互概念。

2. 回顾博弈论交互研究以解释深度神经网络

本文是博弈论交互解释体系中的一项非典型成果。近年来，我们的研究团队从以下三个方面建立了利用博弈论交互解释黑盒人工智能模型的理论体系。

• 首先，我们利用博弈论交互来**解释深度神经网络所编码的知识/概念**。长期以来，在博弈论中定义深度神经网络输入变量之间的交互，为解释深度神经网络提供了一个新的方向 [6, 12]。我们的研究团队定义了多变量交互作用 [15, 17] 和多阶交互作用 [16]，以表示复杂程度各异的交互作用。我们还定义了用于表示变量缺失的最优基线值，从而确保基于掩蔽的归因方法具有可信度 [10]。

从交互作用拓展到概念。具体而言，我们研究团队近期的研究 [9, 7]发现，使用深度神经网络输入变量之间的Harsanyi交互时，深度神经网络通常会编码非常稀疏的交互作用。这些研究实际上表明，博弈论交互可以视为深度神经网络编码的概念。该结论基于以下三项发现得出。(1) 研究发现，训练良好的深度神经网络通常只编码少量显著交互，而且每种交互都会对深度神经网络的推理分数产生一定影响。(2) 研究发现，我们可以利用这些少量显著交互，准确估计任意掩蔽输入样本的推理分数。(3) 研究发现，这些显著交互通常在不同样本和不同深度神经网络之间具有很强的可迁移性，而且还表现出很强的判别能力。

上述三项发现可以视为基础，用以解释深度神经网络的推理逻辑，即检测输入变量之间的显著交互作用。例如，我们的研究小组使用多变量交互作用来发现

[1]在此，我们聚焦于大语言模型依次生成词语的语言生成任务。

深度神经网络编码的典型视觉概念 [3], ，并使用多阶交互作用分析深度神经网络编码形状和纹理时的具体信号处理行为 [2]。

• 其次，我们利用博弈论交互作用来**考察深度神经网络的表征能力**。例如，我们利用多阶交互作用解释随机失活操作 [17],为深度神经网络带来的泛化能力，以及深度神经网络的对抗鲁棒性和对抗可迁移性 [13, 14, 9]。Deng 等人 [4] 发现，深度神经网络难以表征涉及中等数量输入变量的交互作用。Zhang 等人 [20] 发现，复杂交互作用（即涉及大量输入变量的交互作用）更容易发生过拟合。此外，Ren 等人 [11] 发现，贝叶斯神经网络（贝叶斯神经网络）较不容易编码复杂交互作用，从而避免了过拟合问题。

• 第三，我们利用博弈论交互作用来**分析许多实证发现所共有的机制**。Deng 等人 [5] 证明，十四种归因方法的核心机制都可以重新表述为交互重分配，从而能够公平比较基于不同启发式方法构建的归因方法。Zhang 等人 [18] 证明了十二种既有可迁移性增强方法所共有的机制：这些方法都减少了区域对抗扰动之间的交互作用。

3. 解释大语言模型的推理逻辑

3.1. 预备知识：用交互概念解释推理逻辑

最近，一系列研究 [9, 7]发现，深度神经网络的输出分数可以分解为不同输入变量之间各种交互作用的数值效应，例如，输入图像中不同区域之间的交互作用，以及输入句子中不同单词之间的交互作用。具体而言，给定一个预训练深度神经网络 v： $\mathbb{R}^n \to \mathbb{R}$，令 $\boldsymbol{x}$ 表示输入样本，该样本包含 n 个变量，索引为 $N = \{1, 2, ..., n\}$。不失一般性，我们只关注深度神经网络的一个标量输出 $v(\boldsymbol{x}) \in \mathbb{R}$，例如，输出向量的一个维度。深度神经网络通常会对不同变量之间的交互作用进行编码，以用于推理。每种交互作用都会对网络输出产生数值贡献 $v(\boldsymbol{x})$。为此，Ren 等人 [9]定义了以下指标，用于衡量深度神经网络 $S \subseteq N$中已编码的输入变量之间的交互作用。

$$I(S|\boldsymbol{x}) \triangleq \sum\nolimits_{T\subseteq S}(-1)^{|S|-|T|} \cdot v(\boldsymbol{x}_T). \tag{1}$$

其中， $v(\boldsymbol{x}_T)$ 表示将 $N \setminus T$ 中的变量屏蔽并保持 T 中的变量不变时的网络输出。

• **理解大语言模型在对话任务中的交互作用**。给定一个经过训练、用于对话的大语言模型，让我们理解公式（1）中定义的交互作用。大语言模型会依次生成词语来补全输入句子。例如，给定输入句子“他是绘画方面的新手，这意味着”，ChatGPT [8] 会将输入句子补全为：“他是绘画方面的新手，这意味着他是绘画初学者或生手。”在这种情况下，大语言模型会**逐个**生成词语：“他”、“是”、“一名”、“新手”等。因此，当大语言模型已经生成 n 个词语（记为 $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^\top$）时，我们分析大语言模型生成第 $n+1$ 个词语的概率。例如，当大语言模型已生成“他是绘画方面的新手，这意味着他是一个”时，我们将这个句子作为输入 $\boldsymbol{x}$ 提供给大语言模型。随后，我们分析大语言模型生成目标词 $y_{\text{truth}} =$ [新手] 的概率。大语言模型生成目标词 y_{truth} 的推理分数可表示如下。请注意，对于不同的任务， $v(\boldsymbol{x})$ 可以采用不同的设置。

$$v(\boldsymbol{x}) = \log \frac{p(y = y^{\text{truth}}|\boldsymbol{x})}{1 - p(y = y^{\text{truth}}|\boldsymbol{x})} \tag{2}$$

因此， $v(\boldsymbol{x}_T)$ 在公式（1）中对应于在输入句子中的 $N \setminus T$屏蔽词语时的推理分数。大语言模型通常使用填充标记（例如，OPT-1.3b模型使用的“PAD”标记 [19]）作为占位符，它不包含任何语义含义。因此，我们通过保持 $\boldsymbol{x}_T$ 中的词语不变，并使用填充标记屏蔽 T中的词语，生成屏蔽后的句子 $N \setminus T$。

通过这种方式，可以将 S 中特定词语集合之间的交互作用 $I(S|\boldsymbol{x})$ 理解为这种交互作用对大语言模型生成目标词的影响。 > 0 表示 S 中的词语之间的交互作用对生成目标词有积极影响。 < 0 表示这种交互作用阻碍大语言模型生成目标词。 ≈ 0 表示这种交互作用对生成目标词几乎没有影响。

更准确地说，交互作用 $I(S|\boldsymbol{x})$ 表示 S 中词语之间的与关系。在上面的例子“he is a green hand in painting means that he is a beginner”中，我们来考虑 $S = \{green, hand\}$ 这两个具有较强交互效应的词。只有当“*green*”和“*hand*”这两个词同时出现在输入句子中时，{*green*,*hand*} 之间的交互作用才会被激活，并产生数值效应 $I(S|\boldsymbol{x})$，促使大语言模型生成“新手”一词。

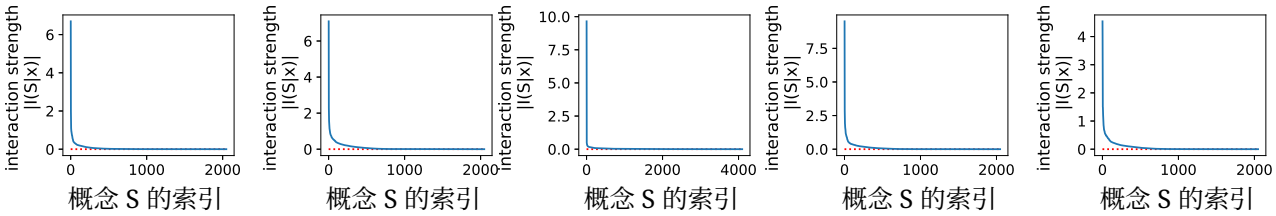


图 1. 所有交互作用按照 $|I(S|\boldsymbol{x})|$ 值降序排列。每个子图对应一个输入句子的交互作用。图中显示，只有少量交互作用对大语言模型的输出产生了显著影响，其他所有交互作用对大语言模型输出的影响都可以忽略不计。

否则，如果 S 中的任何词被屏蔽，那么大语言模型的输出就不会受到这种交互作用的影响，也就是说，使 $I(S|\boldsymbol{x}) = 0$。

• **稀疏且普适的匹配**。给定一个深度神经网络 v 和一个输入样本 $\boldsymbol{x}$，其中包含 n 个变量，我们可以生成 2^n 个不同的掩码样本 $\{\boldsymbol{x}_S|S \subseteq N\}$。定理 1 表明，我们可以使用方程（1）中定义的交互作用，在所有这些 2^n 掩码样本上普适地模拟深度神经网络的输出。为简便起见，我们使用 I_T 来表示 $I(T|\boldsymbol{x})$。

定理1。（忠实性，由 *Ren* 等人证明。 *[9]*）深度神经网络在所有掩码样本 $\{\boldsymbol{x}_S|S \subseteq N\}$ 上的输出，都可以普适地表示为被触发的交互效应之和，即 $\forall S \subseteq N,\ v(\boldsymbol{x}_S) = \sum_{T\subseteq S} I_T$。

此外，Ren 等人 [7] 发现，训练良好的深度神经网络通常会编码稀疏交互作用（见注释 2）。换言之，只有少数交互作用会对深度神经网络的输出产生显著影响，即，这些交互作用的 $|I(S|\boldsymbol{x})|$ 值相对较大。相比之下，其他所有交互作用对深度神经网络的输出影响都可忽略，即，$I(S|\boldsymbol{x}) \approx 0$，因此可以将其视为噪声模式。因此，**我们可以利用少数具有不可忽略效应的显著交互来准确估计深度神经网络在所有 2^n 不同掩码样本**上的输出，如下所示；从理论上保证了解释的忠实性。

备注2。（稀疏性）对于所有掩码样本 $\{\boldsymbol{x}_S|S \subseteq N\}$，深度神经网络的输出都可以通过$\Omega_{salient}$中的少数显著交互作用很好地估计，且满足 $|\Omega_{salient}| \ll 2^n$，即 $\forall S \subseteq N, v(\boldsymbol{x}_S) \approx \sum_{T\in\Omega_{salient},\ T\subseteq S} I_T$。

根据定理 1 和备注 2，我们可以将这些**稀疏的显著交互作用视为深度神经网络编码的符号概念**。例如，$S = \{$*green*，*hand*$\}$ 中两个词之间的显著交互作用可以视为一个概念，其语义含义是“新手。”

• **深度神经网络编码的符号概念的可迁移性和判别能力**。Li等人 [7] 发现，从一个样本中提取的许多显著交互作用，也可以在同一类别的另一个样本中找到，即，显著交互作用在同一类别的不同样本之间具有相当大的可迁移性。此外，这些稀疏的显著交互作用在分类任务中也表现出一定的判别能力。也就是说，对于同一类别中不同样本的分类，同一交互作用通常会持续产生正面影响或持续产生负面影响。

3.2. 探索大语言模型所表征的概念

在本小节中，我们进行了实验，以诊断大语言模型所编码的特征表示。

• **实验1：大语言模型的推理分数能否被解耦为符号概念？** 先前研究发现，传统深度神经网络通常编码稀疏符号概念。然而，大语言模型的参数远多于传统深度神经网络。对于大语言模型存在两种相互冲突的理解：1. 由于其参数数量庞大，大语言模型编码的概念比传统深度神经网络复杂得多；或者，2. 大语言模型编码了更多稀疏概念，因为大语言模型通常经过了更精细的训练，因此学习到了更清晰的特征。

因此，在本研究中，我们进一步考察参数远多于传统深度神经网络的大语言模型是否也编码稀疏符号概念。

首先，我们将可视化大语言模型所编码的不同概念的交互效应分布，以考察概念的稀疏性。此外，我们还将检验备注 2 中的论断：我们可以利用这些稀疏概念，准确估计大语言模型在所有随机掩码样本上的输出。具体而言，我们遵循 [9] 提取大语言模型编码的若干概念。然后，给定一个随机掩码句子 $\boldsymbol{x}_S$，$S \subseteq N$，我们比较真实输出 $v^{\text{real}}(\boldsymbol{x}_S)$ 与备注 $v^{\text{approx}}(\boldsymbol{x}_S)$中由提取出的概念近似得到的输出 2。如果提取出的概念能够很好地匹配大语言模型在所有掩码句子上的输出，那么我们就可以认为，大语言模型的推理分数能够由符号概念忠实地解释。

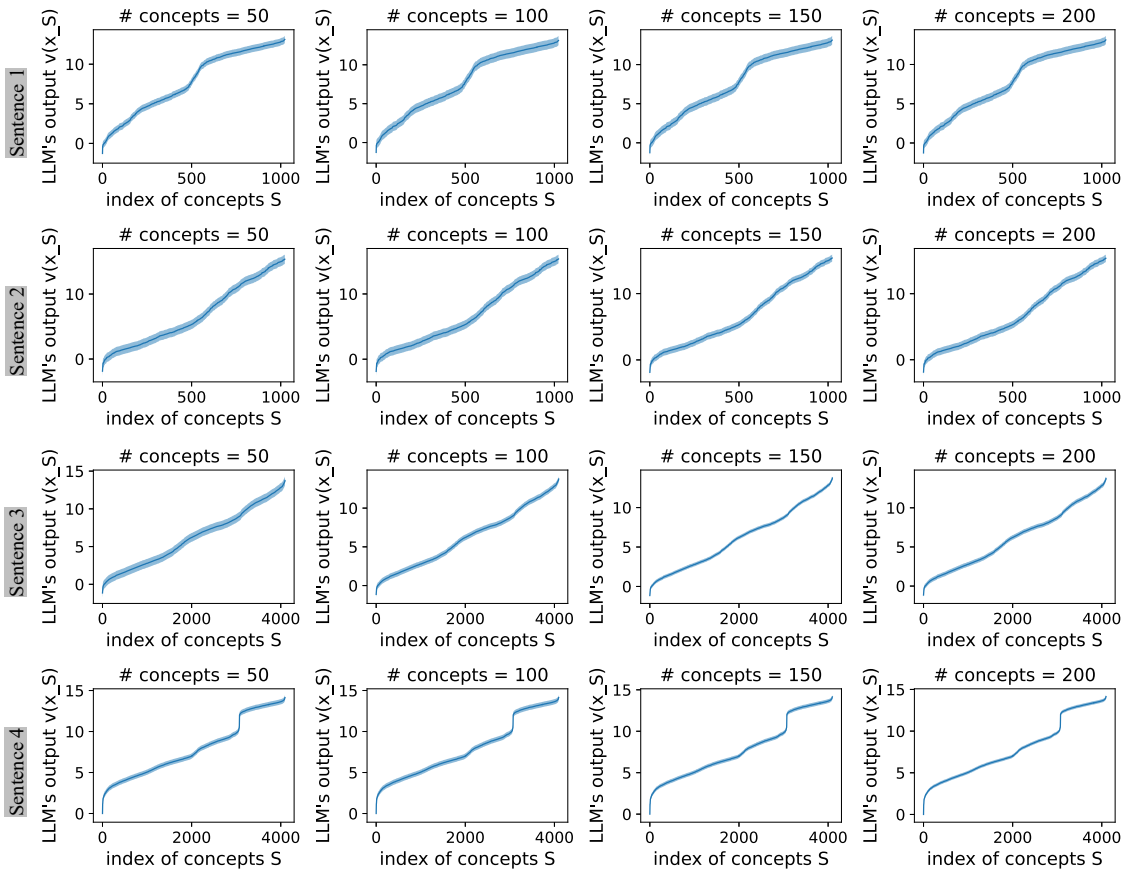


图2。不同掩码样本的模型输出（按升序排列）。阴影区域表示使用排名最高的概念匹配真实模型输出时的均方根误差（RMSE）。通常，使用更多概念匹配模型输出时，偏差会减小。

为此，我们以 OPT-1.3b模型 [19] 作为目标大语言模型开展实验，该模型包含 13 亿个参数。我们使用该大语言模型生成一组输入句子，以构建测试句子集，涵盖物理学、医学、艺术、机器学习等领域的事实知识。随后，对于每个测试句子，我们分析大语言模型在将句子 $(n+1)$以及已生成的前 y^{truth}个词 $\boldsymbol{x}$ 作为输入时，生成第n 个词的概率。具体而言，我们采用 [7] 中的方法，将交互 $I(S|\boldsymbol{x})$ 量化为概念$S \subseteq N$ 对目标（最优）词生成的影响 y^{truth}。图 1 展示了大语言模型针对不同输入句子的推理所产生的所有交互效应。结果表明，从一个输入句子中提取出的显著交互（即概念）只有少数，其余交互对大语言模型输出的影响都可以忽略不计。

此外，我们使用 [9, 7] 中的方法，分别生成排名最高的50、100、150和200个交互作用，将其作为四组不同的概念来解释大语言模型。然后，针对每个句子 $\boldsymbol{x}$，我们通过对不同词语集合进行采样 2^n 掩码句子 $\boldsymbol{x}_S$，构建了一个包含所有这些掩码句子的测试集 S。

为验证提取的概念是否与真实输出 $v^{\text{real}}(\boldsymbol{x}_S)$ 匹配良好，我们测量了近似误差的标准差。令 $\boldsymbol{v}^{\text{real}} = [v^{\text{real}}(\boldsymbol{x}_{S_1})\ v^{\text{real}}(\boldsymbol{x}_{S_2}) \dots, v^{\text{real}}(\boldsymbol{x}_{S_{2^n}})]^\top \in \mathbb{R}^{2^n}$ 表示大语言模型在所有 2^n 个掩码句子上的输出向量，并按升序排列。即，$v^{\text{real}}(\boldsymbol{x}_{S_1}) \le v^{\text{real}}(\boldsymbol{x}_{S_2}) \le \cdots \le v^{\text{real}}(\boldsymbol{x}_{S_{2^n}})$。接着，令 $\boldsymbol{v}^{\text{approx}} = [v^{\text{approx}}(\boldsymbol{x}_{S_1})\ v^{\text{approx}}(\boldsymbol{x}_{S_2}) \dots, v^{\text{approx}}(\boldsymbol{x}_{S_{2^n}})]^\top \in \mathbb{R}^{2^n}$ 表示根据注2，由提取的概念近似得到的输出向量。这样，$v^{\text{real}}(\boldsymbol{x}_{S_i}) - v^{\text{approx}}(\boldsymbol{x}_{S_i})$表示第 i个掩码句子的匹配误差。对于每个模型输出 $v^{\text{real}}(\boldsymbol{x}_{S_i})$，我们计算了相邻 50 个句子的匹配误差的均方根误差（RMSE）$\text{error}_i = \sqrt{\frac{1}{2t+1}\sum_{j=i-t}^{i+t}(\Delta v_i - \mu_i)^2}$，以评估匹配质量。

图 2 展示了所有掩码句子上的真实输出 v^{real}（$\boldsymbol{x}_S$）及相应的均方根误差值。结果表明，在不同的掩码句子上，近似输出与真实输出高度匹配。这证明了

表1。不同句子中提取的概念分布之间的相似度。 M 表示从每个句子中提取的概念数量。

	M = 5	M = 10	M = 15	M = 20	M = 25	M = 30
平均 sim($I^{(1)}$, $I^{(2)}$)	0.565 ± 0.202	0.528 ± 0.205	0.476 ± 0.184	0.454 ± 0.183	0.433 ± 0.179	0.416 ± 0.173

表2. 从大语言模型的输入句子中提取的符号概念。

句子1：人脑约含有1000亿个神经元， 它们通过电的方式彼此 预测词：化学的		句子2：使物体加速所需的力与其加速度成正比 与其质量和加速度有关，正如牛顿定律所描述的 预测词：运动	
概念 S	推断效应 $I(S\|\boldsymbol{x})$	概念 S	推断效应 $I(S\|\boldsymbol{x})$
{电的、并且}	4.82	{定律}	5.99
{电的}	3.46	{牛顿}	3.73
{通过、电的、并且}	-2.37	{牛顿定律}	1.10
{大脑}	1.11	{加速度}	0.86
{大脑、电的，以及}	1.08	{力}	0.61
句子 3：乙型肝炎病毒是一种传染性极强的血源性病毒 可能导致严重后果 预测词：肝脏		句子 4：牛顿运动定律指出，物体将保持静止 或沿直线做匀速运动 预测词：直线	
概念 S	推断效应 $I(S\|\boldsymbol{x})$	概念 S	推断效应 $I(S\|\boldsymbol{x})$
{严肃}	4.54	{直率}	9.65
{乙型肝炎}	3.99	{定律，直线}	0.62
{乙型肝炎，病毒，高度，传染性，血液，导致}	1.49	{牛顿的，直线}	0.48
{导致}	1.32	{物体，直线}	0.37
{乙型肝炎，严重}	1.13	{物体、保持、笔直}	-0.35

OPT-1.3b模型可以解释为稀疏符号概念。

• **实验 2：大语言模型编码的符号概念的可迁移性**。除了符号概念的稀疏性之外，我们还分析了概念的可迁移性，这被视为概念的另一项重要属性。如果从一个句子中 S 提取出的符号概念也会显著影响大语言模型对其他句子的输出，那么该符号概念就被认为具有可迁移性。

因此，我们在 OPT-1.3b模型上进行了实验，以考察概念的可迁移性。我们首先收集了大语言模型生成的不同句子集合。同一集合中的所有句子都包含相同的词语 y^{truth}。我们旨在考察第 i 个句子中 y^{truth} 之前的所有词语 $\boldsymbol{x}^{(i)}$ 是否共享相似的逻辑。例如，句子 $\boldsymbol{x}^{(1)}$ =“糖尿病是一种会影响身体使用”和句子 $\boldsymbol{x}^{(2)}$ = “糖尿病是一种会损害身体使用”本应共享相同的逻辑，从而促使大语言模型生成词语 y^{truth} = “葡萄糖。”

通过这种方式，我们将句子 $\boldsymbol{x}^{(1)}$ 和 $\boldsymbol{x}^{(2)}$ 之间概念的可迁移性定义为从这两个句子中提取的概念分布之间的相似度。令 $N^{(1)}$ 和 $N^{(2)}$ 分别表示 $\boldsymbol{x}^{(1)}$ 和 $\boldsymbol{x}^{(2)}$中所有词语的集合。随后， $N\text{same} = N(1) \cap N^{(2)}$ 表示两个句子共有的词语集合。接着，我们按照 Li等人 [7] 的方法，从第一个句子 $\boldsymbol{x}^{(1)}$中提取排名靠前的 M 显著交互，即具有最大 $|I(S|\boldsymbol{x})|$ 值的 M 交互作用集合，记为 $\Omega^{(1)}$ salient。类似地，$\Omega^{(2)}$ salient表示从第二个句子 $\boldsymbol{x}^{(2)}$中提取的排名靠前的 M 显著概念集合。为了计算 $\Omega^{(1)}$ salient 和 $\Omega^{(2)}$ salient 中概念之间的相似度，我们构建了向量 $I^{(1)} = [I^{(1)}_{S_1}, I^{(1)}_{S_2}, \ldots, I^{(1)}_{S_d}, \tilde{I}^{(1)}_{S_1}, \tilde{I}^{(1)}_{S_2}, \ldots, \tilde{I}^{(1)}_{S_d}] \in \mathbb{R}^{2d}$和向量 $I^{(2)} = [I^{(2)}_{S_1}, I^{(2)}_{S_2}, \ldots, I^{(2)}_{S_d}, \tilde{I}^{(2)}_{S_1}, \tilde{I}^{(2)}_{S_2}, \ldots, \tilde{I}^{(2)}_{S_d}] \in \mathbb{R}^{2d}$, 其中 $S_1, S_2, \ldots, S_d$ 表示 N^{same}中词语之间的所有 $d = 2^{|N^{\text{same}}|}$ 交互作用。如果 $S_t \in \Omega^{(1)}$ salient，我们将$I^{(1)}_{S_t}$ = 设为max($I(S_t|\boldsymbol{x})$, 0)，并将 $\tilde{I}^{(1)}_{S_t} = -$设为min($I(S_t|\boldsymbol{x})$, 0)；否则，我们设置 $I^{(1)}_{S_t} = \tilde{I}^{(1)}_{S_t} = 0$。于是， $\boldsymbol{x}^{(1)}$ 和 $\boldsymbol{x}^{(2)}$ 之间概念分布的相似度定义为 $I^{(1)}$ 和 $I^{(2)}$之间的杰卡德相似度，即 sim($I^{(1)}, I^{(2)}$) = $\frac{\|I^{(1)} \wedge I^{(2)}\|_1}{\|I^{(1)} \vee I^{(2)}\|_1}$ min($I^{(1)}$,$I^{(2)}$) ~~max(,)~~。因此，较高的相似度 sim($I^{(1)}, I^{(2)}$)表明，句子 $\boldsymbol{x}^{(1)}$ 和 $\boldsymbol{x}^{(2)}$ 共有的大多数概念具有较强的可迁移性。

表 1 报告了不同句子对之间的平均相似度($I^{(1)}, I^{(2)}$)，($\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}$)。结果表明，排名靠前的显著概念通常表现出较强的可迁移性。

• **实验3：大语言模型编码的符号概念的判别能力**。理想情况下，忠实的符号概念应促使大语言模型生成符合事实知识的句子。因此，表2展示了对大语言模型推理影响最显著的概念。结果表明，大多数概念都与

表3. 对大语言模型预测错误负责的符号概念。

句子1：物理学家艾萨克·牛顿出生于村庄 预测词：牛顿		句子2：\n Mike 的妈妈有 3 个孩子；其中 2 个是 Luis 和 Drake。剩下那个孩子的名字是 预测词：Drake	
概念 S	推断效应 $I(S\|\boldsymbol{x})$	概念 S	推断效应 $I(S\|\boldsymbol{x})$
{牛顿}	6.26	{*Drake*}	6.94
{村庄，的}	1.57	{姓名}	1.15
{物理学家}	1.29	{孩子们，*Drake*}	0.64
{的}	1.18	{是，*Luis*}	0.58
{村庄}	1.02	{妈妈，*Drake*}	0.42

人类认知。

• 实验4：考察导致大语言模型预测错误的确切原因。考察大语言模型为何生成与事实相悖的句子的确切原因，是调试大语言模型的基础。正如实验1所证明的，大语言模型的推理分数可以分解为若干符号概念。因此，本文提取了导致生成的句子出错的确切原因。例如，如表3所示，大语言模型生成了句子“物理学家艾萨克·牛顿于1642年出生在牛顿村庄”，其中“在牛顿村庄”与事实“在伍尔索普-科尔斯沃思”相悖。随后，我们分析了大语言模型生成与事实相悖的第$n+1$个词的原因，记作 $y$counter。我们提取了所有对生成 $y$counter产生显著但正向影响的概念，即 $S \in \Omega$salient$I(S|\boldsymbol{x}) > 0$，并将其视为错误的原因。表3展示了OPT-1.3b模型预测错误的两个示例。在第一个示例中，{Newton}这个概念（人名）对大语言模型将“牛顿”（地名）作为预测错误的影响最大，但正确答案是“伍尔索普-科尔斯沃思”。这可能是因为OPT-1.3b模型没有学习过人物传记，因此无法回答这类问题。在第二个示例中，{Drake}这个概念对OPT-1.3b模型将“Drake”作为预测错误的影响最大，但正确答案是“Mike”。我们发现，影响OPT-1.3b模型对“Mike”一词进行编码的是换行符“\n”。如果没有换行符“\n”，OPT-1.3b模型就会编码{Mike}这一概念，并生成“Mike”一词。

4. 结论

本文分析了大语言模型在对话中编码的符号概念。具体而言，我们通过实证验证，大语言模型的推理分数可以分解为少量概念。这些符号概念通常具有较高的可迁移性和一定的判别能力。更重要的是，我们还利用这些符号概念解释了大语言模型的预测错误。

参考文献

[1] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等人。《语言模型是少样本学习者》。神经信息处理系统进展，33:1877–1901，2020年。1[2] Xu Cheng、Chuntung Chu、Yi Zheng、Jie Ren 和 Quanshi Zhang。《深度神经网络视觉概念的博弈论分类体系》。 *arXiv* 预印本 *arXiv:2106.10938*，2021年。3[3] Xu Cheng、Xin Wang、Haotian Xue、Zhengyang Liang 和 Quanshi Zhang。《神经网络中审美欣赏的一种假设》。 *arXiv* 预印本 *arXiv:2108.02646*，2021年。3[4] Huiqi Deng、Qihan Ren、Hao Zhang 和 Quanshi Zhang。《发现并解释深度神经网络的表征瓶颈》。载于国际学习表征会议，2022年。1，3[5] Huiqi Deng Deng、Na Zou、Mengnan Du、Weifu Chen、Guocan Feng、Ziwei Yang、Zheyang Li 和 Quanshi Zhang。《利用泰勒交互作用理解并统一十四种归因方法》。 *arXiv* 预印本，2022年。3[6] Michel Grabisch 和 Marc Roubens。《合作博弈中参与者之间交互作用概念的公理化方法》。《国际博弈论杂志》，28:547–565，1999年。2

[7] Mingjie Li 和 Quanshi Zhang。神经网络真的编码了符号概念吗？arXiv 预印本 arXiv:2302.13080，2023年。1，2，3，4，5，6[8] OpenAI。ChatGPT：OpenAI 的对话式人工智能语言模型。https://openai.com/blog/chatgpt/，2022年。访问日期：2022-10-14。1，3[9] Jie Ren、Mingjie Li、Qirui Chen、Huiqi Deng 和 Quanshi Zhang。定义并量化深度神经网络中稀疏概念的涌现。IEEE计算机视觉与模式识别会议，2023年。1，2，3，4，5[10] Jie Ren、Zhanpeng Zhou、Qirui Chen 和 Quanshi Zhang。我们能否忠实地表示掩码状态，以便在深度神经网络上计算Shapley值？载于国际学习表征会议，2023年。1，2[11] Qihan Ren、Huiqi Deng、Yunuo Chen、Siyu Lou 和 Quanshi Zhang。贝叶斯神经网络往往会忽略复杂且敏感的概念。arXiv 预印本 arXiv:2302.13095，2023年。3[12] Mukund Sundararajan、Kedar Dhamdhere 和 Ashish Agarwal。Shapley-Taylor交互指数。载于国际机器学习会议，第9259–9268页。PMLR，2020年。2[13] Xin Wang、Shuyun Lin、Hao Zhang、Yufei Zhu 和 Quanshi Zhang。解读对抗攻击的归因与交互作用。载于IEEE/CVF国际计算机视觉会议论文集，第1095–1104页，2021年。3[14] Xin Wang、Jie Ren、Shuyun Lin、Xiangming Zhu、Yisen Wang 和 Quanshi Zhang。解读并增强对抗可迁移性的统一方法。载于国际学习表征会议，2021年。3[15] Die Zhang、Hao Zhang、Huilin Zhou、Xiaoyi Bao、Da Huo、Ruizhao Chen、Xu Cheng、Mengyue Wu 和 Quanshi Zhang。为深度自然语言处理模型构建可解释的交互树。载于AAAI人工智能大会论文集，第35卷，第14328–14337页，2021年。2[16] Hao Zhang、Sen Li、YinChao Ma、Mingjie Li、Yichen Xie 和 Quanshi Zhang。从博弈论视角解读并增强随机失活。载于国际学习表征会议，2021年。2[17] Hao Zhang、Yichen Xie、Longjie Zheng、Die Zhang 和 Quanshi Zhang。解读深度神经网络中的多变量Shapley交互作用。载于AAAI人工智能大会论文集，第35卷，第10877–10886页，2021年。2，3[18] Quanshi Zhang、Xin Wang、Jie Ren、Xu Cheng、Shuyun Lin、Yisen Wang 和 Xiangming Zhu。证明十二种增强对抗可迁移性方法共享的共同机制。arXiv 预印本 arXiv:2207.11694，2022年。3[19] Susan Zhang、Stephen Roller、Naman Goyal、Mikel Artetxe、Moya Chen、Shuohui Chen、Christopher Dewan、Mona Diab、Xian Li、Xi Victoria Lin 等人。OPT：开放式预训练变换器语言模型。arXiv 预印本 arXiv:2205.01068，2022年。3，5[20] Huilin Zhou、Hao Zhang、Huiqi Deng、Dongrui Liu、Wen Shen、Shih-Han Chan 和 Quanshi Zhang。深度神经网络泛化的概念级解释。arXiv 预印本 arXiv:2302.13091，2023年。3