信息引力：大型语言模型中令牌选择的场论模型

Maryna Vyshnyvetska

Abstract

我们提出了一个名为“信息引力”的理论模型，用于描述大型语言模型（LLMs）中的文本生成过程。该模型使用来自场论和时空几何的物理工具来形式化用户查询与生成的标记概率分布之间的相互作用。查询被视为一个具有“信息质量”的对象，它使模型的语义空间产生弯曲，形成吸引标记的重力势阱。在生成过程中，这种模型提供了一种机制来解释 LLM 行为中观察到的几个现象，包括幻觉（由低密度语义空洞引发）、对查询表达的敏感性（由于语义场的曲率变化），以及采样温度对输出多样性的影响。

1 介绍

大型语言模型（LLMs）已经彻底改变了人工智能领域，它们在文本理解和生成能力上接近于人类水平。然而，尽管有令人印象深刻的结果，这些模型的内部运行机制在很大程度上仍然是一个“黑箱”。正如 Amodei 在其文章《可解释性的紧迫性》 [1] 中所指出的那样，研究人员对 LLMs 生成特定响应的原因及其得出结论的方式了解有限。随着 LLMs 开始在经济、技术和国家安全中发挥核心作用，这种缺乏透明性的问题变得越来越严重。

特别值得关注的现象包括不可预测的幻觉、对查询表述的极端敏感性，以及生成的标记概率分布中令人费解的模式。这些现象不仅限制了 LLM 在关键应用中的可靠性，还指出了我们对其运作理解中的根本性空白。

1.1 LLMs 中的可解释性挑战

当前对 LLM 进行解释的研究方法关注它们功能的各个方面：从个别组件的机械可解释性 [9] 到对注意力和神经元激活的分析 [5] 。然而，这些方法大多仍然是片面的，未能为理解模型行为作为一个完整系统提供一个统一的解释结构。

研究表明，即使在查询表述上的细微变化也可能导致截然不同的响应 [3] 。Sclar 等 [10] 发现，对于同一任务，不同的查询格式会导致质量差异高达 56 个百分点。这种敏感性表明存在复杂的内部动态，而不能用简单的统计近似来解释。

另一个问题与“非自然”的概率分布有关。Zhang 等人 [13] 证明，当要求生成随机数字或名字时，模型表现出对某些回答的强烈偏好，即使从理论上讲，均匀分布应是随机数字或名字请求的不确定性的自然直观模型。这表明生成过程中存在隐藏的“吸引子”。

1.2 对统一理论模型的需求

为了解决这些问题，需要建立一个可以：

1. 将对于大型语言模型行为的不同观察整合为一个统一的解释结构
2. 提出一种理解和预测幻觉的机制
3. 解释模型对查询形式的敏感性
4. 提出一个管理生成过程的概念框架

的理论模型

现有的方法，例如基于能量的模型、信息几何和语义景观范式 [6] ，提供了这种理论的个别元素，但并未将它们整合到一个统一的模型中。

www.xueshuxiangzi.com

1.3 提出的方法：信息引力

在本文中，我们提出了一个新的理论模型，称为“信息引力”，它将物理引力与大型语言模型中的文本生成过程进行类比。根据这个模型，用户查询可以看作是一个具有信息“质量”的对象，它使模型的语义空间弯曲，产生“引力井”，从而将生成过程吸引到某些标记上。

我们方法的关键元素是如下概念的形式化：

1. 查询的信息量作为其熵、上下文深度和新颖度的函数
2. 语义潜力作为标记选择概率的衡量标准
3. 信息重力作为该势的梯度
4. 生成过程的热力学解释

该模型不仅可以帮助我们解释 LLM 运作中的观察现象，还可以提出改进其功能的具体方法，包括信息质量管理、自适应采样温度调节，以及开发新的质量指标。

1.4 论文结构

在接下来的部分中，我们详细介绍了所提出的信息重力模型。第二部分将模型的关键概念形式化。第三部分描述了在弯曲语义空间中进行令牌选择的动态。第四部分解释了模型如何预测幻觉、查询敏感性和温度效应。第五部分提出了理论验证的实验方法。最后，第六部分讨论了模型开发的实际应用和前景。

通过信息引力的视角，我们不仅旨在更深入地了解大型语言模型的内部机制，还希望扩展我们应对这些日益重要的人工智能系统的概念工具包。

在大型语言模型（LLM）中，生成用户查询响应的过程可以视为模型在一个由标记及其组合组成的特定概率空间中的移动。类似于物理质量扭曲时空并创建引力场，提示中的信息扭曲 LLM 的语义空间，形成“信息吸引”区域。这个“信息引力场”决定了在生成过程中的后续标记的选择，使标记生成不是随机选择，而是跨越语义景观的定向移动。

将信息视为吸引力来源的想法在各个领域都有相似之处。在物理学中，信息引力假说（IGH） [2] 描述了引力是一种由信息过程产生的涌现现象。在机器学习中，数据分类的认知引力模型 [11] 使用对象的自信息作为质量的类比，定义样本之间的吸引力与其信息内容成正比。在认知科学中，智力有时被比作等离子体，其运动由信息引力 [8] 决定。

1.5 模型的潜在空间

我们将模型 L 的潜在空间定义为一个 n 维的隐藏表示空间，其中每个标记、上下文或注意向量被表示为一个点或区域。这个空间的维度由模型的结构决定，并对应于其内部表示的维度。

在这个空间内，模型进行计算，这正是我们讨论的“曲率”发生的地方。这不是一个随意的比喻——变压器确实通过一系列非线性映射来转换输入数据，在概率空间中创建复杂的几何结构，这与 Gokhale 提出的语义景观范式一致 [6] 。

注意力机制直接影响这个潜在空间的形成是非常重要的。注意力分数起着类似于场线的作用，引导模型朝向这个空间的特定区域，从而形成具体的生成路径。因此，注意力可以明确地形式化为一组矢量，这些矢量定义了语义景观中的运动方向，从而创造出“引力通道”，通过这些通道进行的标记生成流动。这与 Dar 等人的研究相符，[4] ，他们表明 transformer 参数可以在词汇嵌入空间中直接解读。

此外，transformer 的每个注意力层中的注意力分数可以被解释为一个矩阵，该矩阵不仅定义了方向，还定义了令牌之间交互的相对强度。因此，注意力矩阵作为描述信息场强度的张量的类似物，进一步加强了与物理场理论的类比。

1.6 查询的信息量

查询的信息质量反映其扭曲语义空间和形成概率吸引区域的能力。我们建议将这种质量形式化为三个主要参数的加权组合：

$$M(Q) = \alpha \cdot H(Q) + \beta \cdot D(Q) + \gamma \cdot N(Q) \tag{1}$$

其中：

www.xueshuxiangzi.com

- $H(Q)$ 是查询的信息熵，表征不确定性和模糊性。
- $D(Q)$ 是上下文深度，反映查询与对话或模型知识的先前部分之间的连接程度。
- $N(Q)$ 是查询的新颖性，评估当前查询与模型训练时使用的查询分布之间的距离。
- α, β, γ 是使模型针对特定任务和数据类型进行调整的系数。

信息量的各个组成部分可以通过以下方法独立测量：

- 熵 $H(Q)$ ： 标准的香农熵，基于标记概率计算。
- 语境深度 $D(Q)$ ： 通过总互信息进行评估，使用语言模型或通过嵌入度量进行测量（例如，查询和语境嵌入之间的余弦相似度）。
- 新颖性 $N(Q)$ ： 查询词元的经验分布与先前从训练数据中收集的统计数据之间的 Kullback-Leibler 散度。

1.6.1 查询的信息熵

查询 $H(Q)$ 的信息熵可以计算为构成查询的符号概率分布的香农熵：

$$H(Q) = -\sum_{t \in Q} P(t|Q) \log P(t|Q) \tag{2}$$

其中 P(t|Q) 是在查询 Q 的上下文中令牌 t 的概率。

高熵表示查询的语义模糊性或复杂性。熵越高，语义空间的曲率越强，模型就越难确定“正确”的生成路径。

1.6.2 上下文深度

上下文深度 $D(Q)$ 反映了查询与先前上下文的整合程度，可以形式化为查询与对话先前部分之间的互信息之和：

$$D(Q) = \sum_{i=1}^{k} I(Q; C_i) \tag{3}$$

其中，$I(Q; C_i)$ 是查询 Q 与之前上下文的第 i 部分之间的互信息，k 是所考虑上下文的深度。

高语境深度会形成复杂且扩展的“引力场”，这些引力场可能有多个局部极小值，从而增加模型在语义空间中被“捕获”到次优区域的概率。

1.6.3 查询新颖性

查询 $N(Q)$ 的新颖性表示查询与用于训练模型的数据分布之间差异的衡量标准：

$$N(Q) = D_{KL}(P(Q)||P_{train}(Q)) \tag{4}$$

其中 D_{KL} 是当前查询 $P(Q)$ 的概率分布与训练数据 $P_{train}(Q)$ 中的分布之间的相对熵。

高新颖性意味着模型在其知识空间中处理查询的“参考点”很少，这增加了不确定性并提高了产生幻觉的风险。

www.xueshuxiangzi.com

1.7 语义场及其潜力

我们将可能的标记空间视为一个多维的语义概率场。在这个空间中，每个标记都有特定的语义势，类似于物理学中的引力势。用户的查询改变了这个场的配置，创造了语义势降低的区域，模型在生成标记时倾向于向这些区域移动。

尽管原始的词元空间是离散的，模型使用的是嵌入，它们是连续向量。在这个连续的嵌入空间中，形成了语义场及其相应的引力势，这解决了词元离散性与引力类比连续性之间的矛盾。

令 token t 与查询 Q 的语义潜能由公式定义如下：

$$\Phi(t, Q) = -\log P(t|Q) \tag{5}$$

其中，$P(t|Q)$ 是在给定查询 Q 的情况下选择符号 t 的概率。

因此，潜力与代币概率的对数成反比：代币的概率越大，其潜力越低。在此模型中，在生成过程中，模型“吸引”具有最小潜力的代币，类似于物体向重力势能的最低点移动。

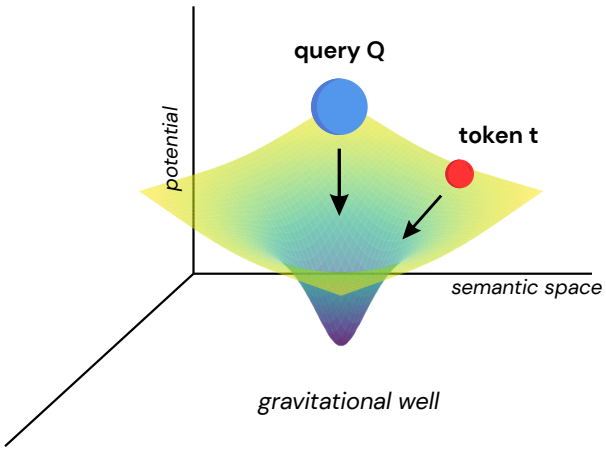

Figure 1: 语义空间中信息引力的说明。用户查询在语义场中创建了一个引力井，吸引标记生成朝向语义势能最小的区域。

1.8 信息引力及其梯度

基于语义势，我们可以将信息引力 $g(t)$ 定义为该势的梯度：

$$g(t) = -\nabla\Phi(t) \tag{6}$$

信息引力指示语义空间中“吸引力”的方向——即模型在生成过程中最有可能移动的方向。$g(t)$ 越高，对应词元的“吸引力”就越强。

值得注意的是，transformers 由多个注意力层组成，每个注意力层形成其自身的语义子空间。因此，模型信息场的真实结构表示出一个复杂的多层嵌套“引力场”系统，这些场相互作用，共同决定令牌的最终概率分布。

在文本生成过程中，大型语言模型根据概率分布选择下一个标记，该概率分布可以通过语义潜力和温度来表达：

$$P(t|Q) = \frac{\exp(-\Phi(t)/T)}{Z} \tag{7}$$

其中 T 是采样温度，Z 是归一化系数（配分函数），定义为：

www.xueshuxiangzi.com

$$Z = \sum_t \exp(-\Phi(t)/T) \tag{8}$$

这个公式呈现为统计物理学中的玻尔兹曼分布，其中温度 T 调节系统中的“热运动”。在低温 ($T \to 0$) 时，模型严格选择具有最低势能的记号，而在高温 ($T \to \infty$) 时，选择变得更加均等，这对应于输出分布中的高熵。

对神经网络的这种热力学解释可在关于基于能量模型（EBM）的研究中找到，Xie 等人 [12] 证明了该学习过程可以在信息几何的分布空间中被解释为一种动态的交替投影。

一个有趣的类比是将高熵生成与量子不确定性进行比较，在这种情况下，模型同时考虑几种可能的生成路径，导致随机且难以预测的行为。这为进一步研究 LLM 的热力学行为与量子系统之间的类比开辟了可能性，特别是在多代理或集成生成方法的背景下。

1.9 生成轨迹作为引力场中的运动

文本生成过程可以看作是在曲折的语义空间中沿着最小作用轨迹的移动。在每一步生成中，模型根据查询和先前上下文创建的语义势的配置来选择下一个标记。

需要注意的是，每个选定的标记都会改变上下文，因此会改变语义场在后续步骤中的配置。这可以被看作一个移动的引力源，不断地重构其周围的潜在场。

该模型在响应的风格和主题上表现出一致性，可以解释为信息惯性——即在模型开始沿语义空间的某一路径移动后，倾向于保持生成方向。这一现象类似于物理学中的惯性，可以通过已选定的标记在给定方向上产生额外的信息吸引力来解释。

形式上，生成轨迹 τ 表示一个最小化作用积分的标记序列 $t_1, t_2, \ldots, t_n$ ：

$$S[\tau] = \int_\tau \Phi(t(\tau), Q) d\tau \tag{9}$$

考虑可能将系统从局部最小值“弹出”的温度波动。

2 模型的影响和预测

所提出的信息引力模型解释了在 LLM 操作中观察到的几种现象：

2.1 幻觉机制

可以通过语义空洞的概念来解释 LLM 中的幻觉——这些是语义场中与训练数据连接稀疏或不存在的区域。当查询 $M(Q)$ 的信息量较高，但模型的局部知识不足时，系统就会遇到这样的空洞。由于无法找到强连接的延续，模型被迫选择一个弱相关或孤立的标记，导致语义不一致或事实不正确的内容。

从数学上讲，这可以通过一个不稳定性条件来表达：

如果语义势的 Hessian（第二偏导数矩阵）

$$H_{ij} = \frac{\partial^2 \Phi}{\partial x_i \partial x_j} \tag{10}$$

在潜在空间的小区域内具有较高的特征值，并且令符号的全局概率为 $P(t) \ll 1$ ，那么系统容易在错误的局部极小值中“捕获”，导致生成似乎合理但事实上不正确的内容。

这种机制解释了为什么当遇到复杂或模糊的查询时，或者当模型遇到新的、未知的信息时，幻觉通常会发生——在这些情况下，语义场具有复杂的、“崎岖”的结构，并存在多个局部最小值。

一个有趣的现象是所谓信息黑洞的出现，即在语义空间中人工诱导出具有极高密度和低熵的区域，由于外部约束（如对齐训练、安全指令或重复提示），模型被强烈吸引到这些区域。在这些区域，模型倾向于生成高度可预测的、重复的或预定义的响应。

另一方面，当模型遇到语义空洞时，就会产生幻觉——这些区域与训练数据的联系稀疏或缺失。在这种情况下，模型被迫选择最近的弱连接标记，这导致语义不一致或事实错误的内容。与黑洞不同，空洞代表了低语义密度和弱连接的区域。

www.xueshuxiangzi.com

2.2 对查询结构的敏感性

信息引力模型预测，查询形式的微小变化可以显著改变语义场的配置。这解释了由 Cox 等人 [3] 描述的 LLMs 对查询格式和结构的高敏感性，即所谓的提示敏感性。

这一效应可以通过查询扰动期间势能景观的变化进行定量评估：

$$\Delta\Phi(t) = \Phi(t, Q') - \Phi(t, Q) \tag{11}$$

，其中 Q' 是查询 Q 的修改版本。

$|\Delta\Phi|$ 越大，响应概率分布随着最小查询修改而发生变化的程度越强，这可能导致语义空间中不同“吸引子”之间出现急剧转换。Sclar 等人的研究 [10] 证实了这一点，展示了不同查询格式下质量变化高达 56 个百分点。

2.3 温度效应和生成熵

模型中的温度参数 T 调节“引力吸引”到局部势能极小值与允许探索更广泛语义空间区域的“热波动”之间的平衡。

在低温度下 $(T \to 0)$，模型严格遵循势能梯度，选择最可能的符号，从而导致确定性但可能刻板的生成。在高温度下 $(T \to \infty)$，“引力场”的影响减弱，生成变得更加随机，这可能导致更具创意但也更混乱的结果。

这一效应与 Zhang 等人的观察结果一致，他们发现大型语言模型在生成随机内容时分布扩散不足，偏向于某些概率不成比例高的标记。例如，当要求输出一个 1 到 10 之间的随机数时，Llama-2-13B-chat 模型不均匀地偏好数字 5；而在生成随机名字时，Mistral-7B 模型选择“Avery”的频率比预期高出 40 倍。

从定量上看，输出分布的熵随温度增加：

$$\lim_{T \to 0} H(P(t|Q, T)) = 0 \quad \lim_{T \to \infty} H(P(t|Q, T)) = \log(|V|) \tag{12}$$

，其中 $|V|$ 是模型的词汇量大小。

2.4 模型局限性

尽管信息引力模型理论上很有吸引力，但它有几个局限性：

- 实证验证需要复杂的实验和大量的数据。
- 信息质量的组成部分的测量受到它们内部相互关系的影响而变得复杂。
- 在可视化和解释多维语义空间尤其是在大型模型中的复杂性。
- 尽管与物理学的类比直观清晰，但如果将其视为一个字面上的物理模型而不是隐喻性的映射，可能会产生误解，因此它应被解释为一个概念框架，而不是直接的物理类比。

3 实验方法

实验应该分开且独立地进行，以测量信息量的各个组成部分。例如，对于熵，制定具有不同歧义程度的查询（同义词，多义词）；对于上下文深度，改变之前消息的长度和复杂性；对于新颖性，使用与训练语料库距离受控的查询（例如，结合众所周知的概念和全新的概念）。然后通过实验验证每个因素对生成行为的影响。

为了实证验证信息引力理论，我们提出了以下实验方法：

3.1 语义景观的可视化

实验验证的一个关键方面是对语义潜力 $\Phi(t)$ 进行可视化，并观察其在不同查询下的变化。这可以通过降维方法实现，例如在应用 softmax 之前对模型的 logit 向量应用 $t-SNE$ 或 $UMAP$ 。

1. 对于一组不同的查询 $Q_1, Q_2, \ldots, Q_m$ ，计算 logit 向量 L_i ，即模型在 softmax 层之前的最终线性变换的原始输出。
2. 将 $t-SNE$ 或 $UMAP$ 应用于获取的向量，以投影到二维或三维空间。

www.xueshuxiangzi.com

3. 将结果投影可视化为“潜力图”，其中较暗的区域对应于较低的潜力（较高的概率）。

这种方法与 Gokhale [6] 的工作中提出的可视化语义景观的方法一致，并且可以通过 Pingbang Hu 和 Mahito Sugiyama [7] 开发的使用基于能量模型的可解释数据生成方法进行补充。

这种方法将直观展示不同查询如何“改变”语义空间，创造各种“引力井”的配置。

3.2 信息量与幻觉之间关系的分析

为了验证查询信息量与模型产生幻觉倾向之间关系的假设，提出了以下实验：

1. 创建一组查询，其中信息量的组成部分受到控制地增加：熵 $H(Q)$ 、上下文深度 $D(Q)$ 和新颖性 $N(Q)$ 。
2. 对于每个查询，生成多个具有不同温度 T 的响应。
3. 评估生成响应中幻觉的频率和严重程度。
4. 分析信息质量成分与幻觉指标之间的关联性。

模型预测：具有高熵、深层背景和高新颖性的查询将表现出显著更高的幻觉频率，尤其是在较低的温度下，此时模型没有能力在语义空间中“探索”替代路径。

3.3 温度效应的研究

为了研究温度对语义场配置的影响，我们提出：

1. 选择一组复杂性不同的固定查询。
2. 对于每个查询，生成多个响应，温度参数 T 在接近 0 到非常高的值之间进行系统性变化。
3. 对于每个 T 的取值，分析：
    - 令牌概率分布的香农熵
    - 生成回复的多样性
    - 语义一致性
    - 幻觉频率

模型预测：在低温条件下，将观察到“引力俘获”现象——模型将始终选择相同的标记，遵循势能梯度。在高温条件下，“热运动”将超过“引力吸引”，标记选择将变得更加随机，增加多样性，但可能降低一致性。

信息引力的概念为实际应用和进一步研究开辟了广阔的视角：

3.4 通过信息质量管理改进生成

理解信息质量的各个成分如何影响语义空间的曲率，我们可以开发出更精确控制生成的方法：

1. 熵平衡：修改查询以达到一种最佳水平的熵 $H(Q)$ ，这种熵足以进行创造性生成，但不足以引发幻想。
2. 上下文稳定化：管理上下文深度 $D(Q)$ ，以创建更稳定的“引力场”，并减少局部最小值。
3. 新颖性和锚定：通过添加“锚定”片段来补偿高查询新颖性 $N(Q)$ ，这些片段将查询与模型已知的信息连接起来，降低在语义空间错误区域发生“重力坍塌”的风险。

这些方法可能特别有用，因为 Cox 等人的观察表明，[3] 关于大型语言模型对查询表达的高度敏感性，并且可能补充 Zhang 等人提出的方法 [13] ，以在生成过程中实现更分散的分布。

www.xueshuxiangzi.com

3.5 自适应温度控制

信息引力模型为开发基于查询属性的自适应采样温度控制算法提供了理论基础：

为了调节生成行为与语义域复杂性的关系，我们提出了以下公式：

$$T_{adaptive}(Q) = f(M(Q)) \tag{13}$$

其中函数 f 将查询的信息量映射到一个最佳温度值。

信息量大的查询可能需要更高的“温度”来克服“引力陷阱”并探索语义空间的更广区域，而简单、明确的查询可以在低温下高效处理。

3.6 LLM 的质量和可解释性指标

信息引力的概念可用于开发新的指标来评估 LLM 的质量和可解释性：

我们提出以下新的度量标准来评估 LLM 的性能：

- 引力曲率：一种衡量语义场中异质性程度的度量（例如，通过势能梯度的变化）。高曲率表示不稳定性和产生幻觉的倾向。
- 信息传导率：一种评估从查询到响应的语义信息传递效率的度量。可以通过查询和响应之间的互信息或交叉熵来测量。
- 语义场稳定性：通过评估语义潜力配置对小查询变化的稳定性来衡量模型的鲁棒性。

这些指标与胡平邦和杉山诚人 [7] 的研究方向一致，并可与谢等人描述的信息几何方法相结合 [12] 。

这些度量可以更深入地理解 LLM 中的内部过程，并有助于开发更可靠和可解释的下一代模型。

在本文中，我们提出了一个信息引力的理论模型，用于解释和预测大型语言模型在文本生成过程中表现出的行为。所提议的模型在物理理论、信息几何和大型语言模型中的概率过程之间建立了一个概念桥梁，提供了一个分析这些系统行为的统一范式。

3.7 关键结果

我们已经形式化了以下关键概念：

1. 查询的信息量是熵、上下文深度和新颖性的组合，影响模型语义空间的“曲率”。
2. 语义势能作为一种度量，描述文本生成的概率性景观，类似于物理中的重力势能。
3. 信息引力作为语义势的梯度，控制着语义景观中符号选择的轨迹。
4. 生成过程的热力学解释，其中采样温度调节“引力吸引”和“热波动”之间的平衡。

这个理论框架解释了在 LLM 操作中观察到的几个实证现象：

- 幻觉产生机制通过语义空洞——语义空间中与训练数据连接稀疏或缺失的区域，导致语义不一致或事实错误的内容。
- 模型对查询公式的高敏感性如同语义场配置的变化
- 通过系统热力学状态的变化，温度对生成多样性和质量的影响
- 由对齐指令、对抗性提示或重复查询模式引起的人为信息黑洞的存在，导致模型输出高度可预测或受限
- 信息惯性，解释生成中的风格和主题一致性

www.xueshuxiangzi.com

3.8 模型开发的展望

所提出的信息引力模型为进一步研究开辟了几个有前景的方向：

1. 通过语义景观的可视化、幻觉分析和温度效应的研究来进行理论的实证验证。
2. 通过信息量操控、适应性温度控制以及开发新的模型质量指标来控制生成的新方法的实际应用。
3. 通过与变压器中的注意力机制建立联系的形式化、与信息几何的整合以及扩展到多模态模型，进行模型的理论发展。
4. 模型间交互——研究不同模型创建的“信息场”及其在多个 LLM 合作工作背景下的交互。

3.9 可能的限制

尽管所提出的模型具有解释力，但它有几个潜在的局限性：

- 独立测量信息质量成分的困难
- 在多层变换器中形式化不同注意力层之间信息场复杂交互的挑战
- 在给定标记离散性的情况下语义空间连续性的简化表示

然而，这些限制更多地代表了模型进一步发展的方向，而不是它适用性的根本障碍。

3.10 结论

信息重力理论不仅为理解大语言模型中的过程提供了一个新的隐喻，还提供了一个可能可验证的科学假设，具有实际应用意义。随着大型语言模型成为越来越强大的工具，理解其运行的基本原理变得至关重要。

所提出的模型可以作为深入理解和控制 LLM 行为的一步，为开发更加可解释、可靠和可管理的人工智能系统提供理论基础。鉴于最近关于机器学习中的“黑盒子”的讨论，所提出的信息引力理论可能成为提升这些黑盒系统可解释性并使语言模型操作更加透明的概念工具。

怀着不合逻辑的温柔感谢 ChatGPT 和 Claude 在准备手稿过程中的贡献。

这项工作也可以作为 Zenodo 预印本获取：https://doi.org/10.5281/zenodo.15283647

References

[1] Dario Amodei. The urgency of interpretability, 2024. Unpublished manuscript.

[2] J. Chang. The informational gravity hypothesis. *OSF Preprints*, 2022.

[3] J. Cox, Y. Han, D. Wang, S. Nair, and Z. Yu. Understanding uncertainty in large language models through prompt perturbations. In *AAAI 2025*, 2025.

[4] Y. Dar, Z. Ren, L. Moysis, S. Agarwal, T. Zhang, and V. J. Hellendoorn. Understanding embeddings in the transformer through the lens of the position-aware linear model. *arXiv preprint*, 2023.

[5] Nelson Elhage et al. Analyses of attention and neuron activations, 2021. Technical report.

[6] C. Gokhale. The semantic landscape paradigm for neural networks. *ResearchGate*, 2023.

[7] P. Hu and M. Sugiyama. Interpretable data generation from energy-based models by projection. *arXiv preprint*, 2024.

[8] J. Lee. Is artificial intelligence possible? *arXiv preprint*, 2018.

[9] Chris Olah et al. Mechanistic interpretability of individual components, 2020. Research blog.

[10] M. Sclar, P. West, P. Liang, and N. Saunshi. Quantifying the sensitivity of large language models to different prompt formats. *arXiv preprint*, 2023.

www.xueshuxiangzi.com

[11] M. Tang and X. Yang. Classification using cognitive gravitation model. In *Lecture Notes in Computer Science*, volume 10016, pages 203–214, 2016.

[12] J. Xie, K. C. Li, Y. Zhu, and C. H. Zhang. Cooperation in variational and non-variational mcmc via energy-based models. In *AAAI*, 2021.

[13] L. Zhang, W. Chen, W. Chen, A. Wang, Q. Deng, Y. Tao, and C. Finn. Tuning language models to generate diffuse distributions. *arXiv preprint*, 2024.

www.xueshuxiangzi.com