可用的 XAI：在 LLM 时代利用可解释性的 10 种策略

Xuansheng Wu [1*] *xuansheng.wu@uga.edu*
Haiyan Zhao [2*] *hz54@njit.edu*
Yaochen Zhu [3*] *uqp4qh@virginia.edu*
Yucheng Shi [1*] *yucheng.shi@uga.edu*
Fan Yang [4] *yangfan@wfu.edu*
Tianming Liu [1] *tliu@uga.edu*
Xiaoming Zhai [1] *xiaoming.zhai@uga.edu*
Wenlin Yao [5] *wenlinyao@global.tencent.com*
Jundong Li [3] *jundong@virginia.edu*
Mengnan Du [2] *mengnan.du@njit.edu*
Ninghao Liu [1] *ninghao.liu@uga.edu*

[1] *University of Georgia* [2] *New Jersey Institute of Technology* [3] *University of Virginia* [4] *Wake Forest University* [5] *Tencent AI Lab (Seattle)*

Abstract

可解释 AI （XAI）是指为 AI 模型的工作原理提供人类可理解的见解的技术。最近，XAI 的重点扩展到大型语言模型（LLM），这些模型经常因缺乏透明度而受到批评。由于两个原因，这种扩展要求对 XAI 方法进行重大转变。首先，许多现有的 XAI 方法由于其复杂性和高级功能而无法直接应用于 LLM。其次，随着 LLM 越来越多地部署在不同的行业应用程序中，XAI 的角色从仅仅打开“黑匣子”转变为积极提高 LLM 在现实世界中的生产力和适用性。同时，与传统的机器学习模型不同，传统的机器学习模型是 XAI 洞察力的被动接受者，LLM 的独特能力可以相互增强 XAI。因此，在本文中，我们通过分析（1）XAI 如何使 LLM 和 AI 系统受益，以及（2）LLM 如何为 XAI 的发展做出贡献，在 LLM 的背景下介绍了可用的 XAI。我们介绍了 10 种策略，介绍了每种策略的关键技术，并讨论了它们的相关挑战。我们还提供案例研究，以演示如何获取和利用解释。本文中使用的代码可在以下位置找到：https://github.com/JacksonWuxs/UsableXAI_LLM 。

[*] Equal contribution

Contents

1 介绍

可解释性在理解机器学习模型和提供改进方向方面具有很大的前景。在实践中，用户对模型的可解释性有很高的期望：

1. Through explanation, can we know if a model works properly?
2. Does explainability help developing better models?

首先，解释有望阐明模型是否按照人类的期望运行。例如，该模型在决策中是否利用了可靠的证据和领域知识？该模型是否包含偏见和歧视？该模型是否显示任何潜在攻击的漏洞？模型会输出有害信息吗？其次，在认识到模型的缺陷时，我们渴望可解释性，以便为开发更好的模型提供信息。例如，如果我们发现模型在进行预测时使用了不可靠或不合理的特征，如何调整模型的行为？我们能否通过使模型的行为与人类偏好保持一致来提高模型的性能？

因此，问题来了：这些期望是否得到满足？近年来，关于可解释人工智能（XAI）的文献迅速扩大，以提高模型的透明度（**?????**），涵盖了针对不同数据模态定制的各种方法，包括视觉（**?**）、文本（**?**）、图（**?**）和时间序列数据（**?**）。一些文献深入研究了特定技术，例如注意力方法、广义加法模型和因果模型。此外，有些提供对一般原则和分类的评论，或发起关于评估解释的忠实性的讨论（**?**）。尽管取得了进展，但 XAI 的最后一英里 - 利用解释 - 并没有得到足够的关注。在许多情况下，我们似乎只满足于获得解释及其相关的可视化，有时然后对模型的优点和缺点进行定性分析。虽然这些解释可以揭示模型的缺陷，但量化模型属性（例如，公平性、安全性、合理性）或采取下一步具体步骤来改进模型仍然是一项艰巨的任务。

实现可用可解释性的挑战是双重的。首先，在 XAI 中，AI 自动化和人类参与之间存在内在冲突。一方面，人类需要定义模型应该遵循的可解释性，或者仔细检查解释以确定模型中是否存在任何漏洞。另一方面，人工监督的要求带来了巨大的成本，对人工智能工作流程中模型调试和改进的可扩展性和实际实施提出了挑战。其次，目前的许多方法将可解释性视为纯粹的技术问题，忽视了从业者和非技术利益相关者的需求。现有的 XAI 方法主要是作为统计和数学工具开发的。然而，这些工具的目标与各个应用领域的从业者的期望之间存在明显的差异（**?**）。满足技术受众的解释可能对非技术受众几乎没有价值。虽然传统深度模型（例如，多层感知器、卷积和递归神经网络）的不透明度问题尚未完全解决，但大型语言模型（LLM）（**????**）的最新进展似乎加剧了我们面临的挑战。首先，LLM 通常具有更大的模型大小和更多的参数。这种增加的模型复杂性加剧了解释其内部工作原理的难度。其次，与主要关注分类和解析等低级模式识别任务的传统 ML 模型不同，LLM 可以处理更复杂的任务，例如生成、推理和问答。了解 LLM 的独有能力对 XAI 技术提出了新的挑战。考虑到 LLM 在各种应用中的变革性影响，确保 LLM 的可解释性和合乎道德的使用已成为迫在眉睫的迫切需求。同时，LLM 的涌现能力也为 XAI 研究提供了新的机遇。他们类似人类的沟通和常识推理技能为实现可解释性提供了前景，这些方式可能会增加或取代人类的参与。

定义“可用的 XAI”。鉴于上述考虑，在 LLM 的上下文中，我们定义了 Usable XAI，它包括以下两个方面。(1) 利用可解释性来增强 LLM 和 AI 系统。除了提供解释或提高 LLM 的透明度之外，我们还探讨了这些解释是否可以查明模型调试的问题或提高 LLM 或整个 AI 模型的整体性能，例如准确性、可控性、公平性和真实性。(2) 利用 LLM 增强 XAI 框架。LLM 的类人通信能力可以通过将数值转换为可理解的语言来增强模型在用户友好性方面的解释。此外，存储在 LLM 中的常识性知识可以通过扮演人类的角色并减轻人类真正参与 AI 工作流程的需求，显着提高现有 XAI 框架的实用性。

本文的贡献。在本文中，我们研究了 LLM 背景下可用的 XAI 技术的 10 种策略。这些策略分为两大类：(1) LLM 的可用 XAI; (2) LLM 为可用 XAI，如图 **??** 所示。此外，我们还进行案例研究，以证实对所选技术的讨论。对于每项战略，我们还探讨了未来工作中需要进一步研究的开放挑战和领域。

- 可用于 LLM 的 XAI。我们将介绍如何利用解释来增强 AI 管道，包括 LLM 和小型模型。首先，我们研究了如何利用解释来诊断和增强 LLM 的效用。我们研究了三种类型的事后解释方法，分别针对 LLM 预测（Section **??** ）、LLM 组件（Section **??** ）和训练样本（Section **??** ）。其次，我们专注于如何利用解释来审查和提高模型可信度（Section **??** ），包括安全性、公平性、毒性和真实性，这对于实现人类一致性至关重要。第三，我们讨论了可解释性如何指导数据的增强，包括推理数据（即提示）和训练数据。具体来说，我们讨论了为 LLM 制作可解释提示的两种策略：思维链提示（Section **??** ）和知识增强提示（Section **??** ）。此外，我们引入了利用 LLM 解释来增强训练数据以改进小模型（Section **??** ）。
- 可用 XAI 的 LLM。在这一部分中，我们研究了利用 LLM 的高级功能来解决传统 XAI 领域的挑战的策略，从而增强了 XAI 在实践中的可用性。首先，我们研究了通过 LLM 的生成能力来增强解释的用户友好性的方法（Section **??** ）。其次，我们介绍了如何利用 LLM 的规划能力来自动化可解释的 AI 工作流

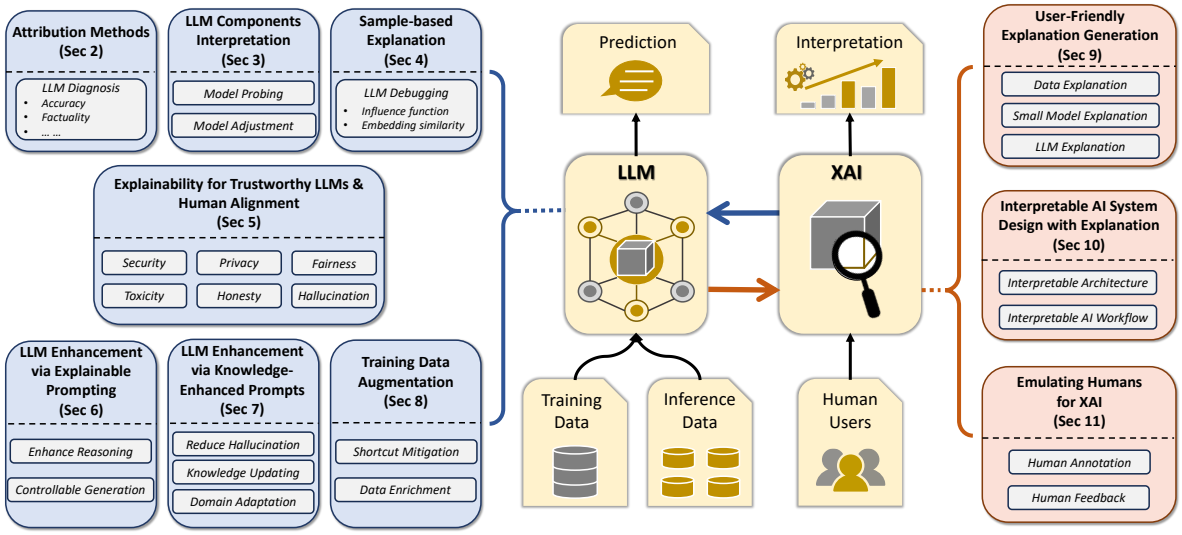


Figure 1: 本文的贡献和大纲。我们在 LLM 的上下文中定义了 Usable XAI，七大策略 使用 XAI 增强 LLM，三大策略 使用 LLM 增强 XAI。

的设计（Section ??）。第三，介绍了如何利用 LLMs 在模拟人类认知过程中的独特性质来促进 XAI 方法的评估（Section ??）。

本文与现有调查的区别。已经进行了许多调查来研究可解释的 AI（???）或可解释的机器学习（?）。本文与现有工作不同，因为我们专注于大型语言模型的解释方法。同时，与以往主要综述 LLMs 解释方法的调查（?）不同，本文强调 XAI 在 LLM 研究中的可用性。据我们所知，与我们调查最相关的论文是（?），其中还讨论了解释可以提高 LLM 性能的几个方面。然而，这种轻量级的调查缺乏对 XAI 方法（例如，基于样本的解释、可解释的工作流程、可解释的提示）以及 LLM 如何使现有 XAI 框架（例如，数据增强、提高用户友好性、XAI 评估）的彻底检查。最后，我们的论文通过提供详细的案例研究和开源代码来进一步做出贡献，促进未来在 LLM 背景下有效应用解释的研究。

2 通过归因方法进行 LLM 诊断

本节介绍归因方法作为 LLM 的事后解释，以及我们如何通过归因分数发现模型缺陷。我们首先重新审视现有的归因方法，然后讨论哪些方法仍然适合解释 LLM。由于 LLM 广泛地服务于分类和生成任务，因此我们的讨论相应地对归因方法进行了分类。之后，我们探讨了应用归因方法来评估 LLM 生成的输出质量的案例研究。最后，我们讨论了为 LLMs 设计新的事后解释方法的未来工作。

2.1 归因方法的文献综述

基于归因的解释量化了有助于做出预测的每个输入特征的重要性。给定一个语言模型，f 根据 N 词输入提示 x 进行预测 $\hat{y} = f(x)$，解释器 g 评估输入词对 x 的影响 $\boldsymbol{a} = g(x, \hat{y}, f) \in \mathbb{R}^N$。通常，$a_n \in \boldsymbol{a}$ 符号表示单词 x_n 对 $\hat{y}$ 产生积极或消极的影响，$|a_n|$ 值越大表示影响越强。在文本分类中，$\hat{y}$ 表示特定的类标签。在文本生成中，$\hat{y}$ 表示生成文本的不同长度。

许多现有的基于归因的解释方法侧重于分类任务，不能直接应用于生成任务。它们之间的主要区别在于：分类仅限于一组特定的预测，而生成包含无穷无尽的可能性。例如，在情感分析中，可以通过在语言模型顶部添加线性层和 sigmoid 函数来指示语言模型输出一个介于 0 和 1 之间的数字，该数字指示输入文本的积极性。然而，在生成式设置中，模型可以通过多种表达方式表达这种积极性，例如“评论者绝对喜欢这部电影”和“这是一部强烈的积极电影评论”。这种区别对将解释方法从分类调整到生成任务提出了独特的挑战。在下文中，我们将根据相关作品适用的场景对相关作品进行回顾。

Table 1: 生成任务不同归因方法的时间复杂度分析。

Method	Forward	Backward	Notes
Mask Perturbation	$\mathcal{O}(N)$	0	-
Gradient × Input	$\mathcal{O}(1)$	$\mathcal{O}(M)$	-
Integrated Gradients	$\mathcal{O}(N_{step})$	$\mathcal{O}(N_{step} \cdot M)$	N_{step} is the number of steps for integrating gradients.
LIME	$\mathcal{O}(N_{aug})$	0	N_{aug} is the number of augmented samples.
SHAP	$\mathcal{O}(2^N)$	0	-

2.1.1 标签分类的归因输入

常见的归因方法 针对传统深度模型开发的 (**??**) 包括基于梯度的方法、基于扰动的方法、代理方法和分解方法。我们介绍了每个类别的一般思想和代表性示例，然后分析了它们是否适合解释大型语言模型。

基于扰动的解释。基于扰动的方法通过扰动输入特征并监测预测置信度的变化来评估输入特征的重要性，即 $a_n = p(\hat{y}|x) - p(\hat{y}|\widetilde{x}_n)$ ，其中 $\widetilde{x}_n$ 是指第 n 个特征受到扰动的输入序列。每个特征可以指一个词 (**?**) 、一个短语 (**?**) 或一个词嵌入 (**?**) 。其基本原则是，扰动更重要的特征应该会导致模型的预测置信度发生更明显的变化。然而，这种方法有局限性，特别是在其假设特征是独立的方面，由于单词相互依赖性，文本数据并不总是如此。此外，它对于解释 LLM 是计算密集型的，需要 N 推理才能输入 N 单词。

基于梯度的解释。基于梯度的方法提供了一种计算效率高的方法，用于估计基于梯度 $\frac{\partial p(\hat{y}|x)}{\partial \boldsymbol{x}_n}$ 的模型对输入特征的敏感性，其中 $\boldsymbol{x}_n$ 是指词 x_n 的嵌入。一些方法采用梯度的 L_2 范数来评估单词重要性 (**?**) ，即 $a_n = \|\frac{\partial p(\hat{y}|x)}{\partial \boldsymbol{x}_n}\|_2$ 。此方法只需要一次推理和一次反向传播传递。一些扩展方法将梯度乘以单词 embedding (**???**) ，即 $a_n = \frac{\partial p(\hat{y}|x)}{\partial \boldsymbol{x}_n} \cdot \boldsymbol{x}_n$ 。这些方法可能对深度模型的忠实度有限 (**?**) ，因为梯度仅反映输入变化和输出变化之间的局部关系。为了解决这个问题，提出了积分梯度 (IG) (**????**) ，它将梯度累积为从参考点到实际数据点的输入转换。然而，IG 需要多轮推理和反向传播，从而大大增加了计算需求。

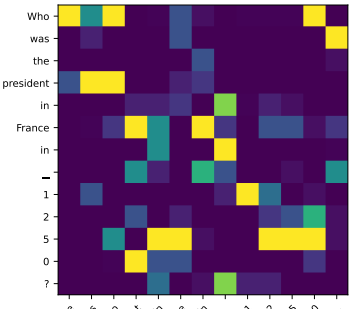
基于代理的解释。基于代理的解释方法通过构建一个更简单的模型来理解复杂的模型 g $\mathcal{D}(x, \hat{y}) = \{(\widetilde{x}_k, \widetilde{y}_k)\}_{k=1}^K$ 上训练，其中 $\mathcal{D}(x, \hat{y})$ 表示为目标实例构建的数据集 $(x, \hat{y})$; $\widetilde{x}_k$ 通常是通过扰动 x 和 $\widetilde{y}_k = f(\widetilde{x}_k)$ 获得的。代理模型 g ，从基本的线性模型到复杂的决策树，作为代理来近似目标模型的决策边界 f 特定实例 $(x, \hat{y})$ 。值得注意的例子包括 LIME (**?**) 、SHAP (**?**) 和 TransSHAP (**?**) ，其中前两个是为通用深度神经网络设计的，最后一个是为基于 Transformer 的语言模型量身定制的。然而，它们的一个重要局限性是它们高度依赖与目标模型的重复交互，这个过程对于 LLM 来说是不切实际的。

基于分解的解释。基于分解的方法将线性加法相关性分数分配给输入，从而有效地分解了模型的预测。Layer-wise Relevance Propagation (**?**) 和 Taylor-type Decomposition (**?**) 是计算这些相关性分数的著名技术。这些方法在各种研究中都适用于基于 Transformer 的语言模型 (**???**) 。然而，实现基于分解的解释的一个主要挑战是需要量身定制的分解策略来适应不同的模型架构。虽然很多大型语言模型都是基于 Transformer 框架的，但它们之间还是存在着关键的差异，比如 LLaMA (**?**) 和 GPT (**?**) ，特别是在位置编码策略和前馈网络设计等方面。这一挑战限制了分解方法在通用解释中的普遍适用性。

总而言之，传统的解释方法并不总是适合 LLMs。特别是，基于扰动和基于梯度的解释相对容易扩展，将 LLM 响应归因于输入提示，而基于代理和基于分解的方法则变得非常具有挑战性。具体来说，基于代理的方法假设一个可解释的小模型可以围绕局部示例近似目标模型的决策边界，但文本生成任务的可解释模型有限。同时，基于分解的方法需要针对不同的层设计分解策略，这对于大型 LLM 架构来说是一个挑战。另一个主要问题是他们对计算资源的巨大需求。给定一个 N 字输入提示和一个 M 字输出响应，我们在表 **??** 中给出了几种代表性解释方法的时间复杂度。它演示了现有方法要么需要大量的前向操作，要么需要后向操作。因此，提高基于归因的解释效率是未来研究发展的重要方向。

2.1.2 为文本生成归因输入

生成模型的解释可以定义为将整体置信 $p(\hat{y}|x)$ 归因于输入 x ，其中 $\hat{y}$ 表示生成的响应 $\hat{y} = [\hat{y}_1, ..., \hat{y}_M]$ M 词。实现此目的的一种方法是将文本生成过程视为一系列单词级分类任务。这种视角允许应用现有的基于分类的解释技术来评估每个输入词 x_n 相对于每个输出词 $\hat{y}_m$ 的影响，从而产生相应的归因分数 $a_{n,m}$ 。在收集了 $m = 1, ..., M$ 的



归属 $a_{n,m}$ 后，我们进行聚合以确定每个输入词 x_n 的总体贡献。这是通过聚合与输入词对应的所有输出词的单个属性来实现的，表示为 $a_n = \text{Aggregate}([a_{n,1}, ..., a_{n,M}])$ 。这种聚合最简单的方法是平均分配给每个输入词的不同输出词的归属（**?**）。然而，**?** 观察到，不同输出词的归因分数本身并不具有可比性。例如，功能词（例如，"the''，"is''，"have''）的归因分数通常不成比例地大于具有明确语义的内容词（例如，动词和名词）的分数。因此，有必要在聚合之前对分数进行归一化，以便分数 $[a_{n,1}, ..., a_{n,M}]$ 变得可比 $1 \leq m \leq M$ 。图 **??** 绘制了一个示例案例的归一化分数，其中 Y 轴中的每个索引都表示一个输入提示标记，而 X 轴中的每个索引都表示输出响应标记。标准化归因分数越高，效果越好。在此示例中，用户尝试将模型定向输出不存在的信息，即 1250 年的法国总统。模型成功地意识到这个东西不存在，并拒绝回答。模型响应可以分为三个部分，“没有”、“法国总统”和“1250 年”。根据该图，第一个跨度的生成很大程度上是因为标记“Who”和“president”，而模型同时使用“France”和“1250”来响应第二个跨度“president in France”。最后，该模型通过引用提示中的相同信息再次强调日期“1250”。总的来说，这些解释与人类的理解相一致，并强调了这种方法在未来的用途。然而，目前关于生成式 LLM 的基于归因的解释的研究仍处于早期阶段，并且只提出了有限数量的方法。

2.2 案例研究：LLM 归因方法的可用性

归因图提供了对 LLM 操作机制的部分见解（**???**）。因此，我们提出了一个利用归因分数来分析 LLM 行为的通用管道，如图 **??** 所示。首先，给定目标 LLM 和输入提示，我们计算输入代币相对于输出代币的归因分数。其次，我们从归因图中提取一个特征向量，根据手头诊断任务的要求进行定制。第三，我们训练一个轻量级预测器（例如，分类器）来诊断模型是否根据特征向量表现得当。在下文中，我们提供了案例研究来说明如何利用归因分数来评估 LLM 响应质量（**?**）。

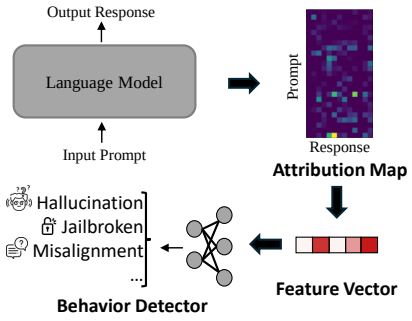


Figure 3: 带有归因解释的模型诊断的一般管道。

2.2.1 LLM 响应质量评估及解释

本案例研究探讨了使用基于归因的解释作为评估 LLM 生成的响应质量的证据。在这里，“质量”是通过响应的准确性来衡量的。我们假设从正确的理由生成的响应可能更准确。我们的方法涉及将模型的解释与基本原理进行比较，作为评估响应准确性的一种手段。这种方法可以提高对话系统的可靠性，这对于医疗问题回答等应用至关重要。

数据。 我们在本案例研究中使用了多句阅读理解（MultiRC）数据集（**?**）。与其他数据集相比，MultiRC 提出了一个更大的挑战：它要求系统根据给定段落中的多个句子来回答问题。答案可能不直接对应于特定的句子跨度，反映了现代 LLM 的更现实场景。每个 MultiRC 实例都与人工注释的基本原理相关联，指示哪些句子与回答问题相关。我们的研究专门采用了总共 950 个样本的测试子集。我们考虑 80 个%，其中是训练集，并报告在其余样本上评估的结果。

实现。 我们应用重要性密度得分法（**?**）来估计所提供段落中每个单词的重要性。具体来说，给定 N -word 段落 X 和 M -word 响应 $\hat{Y}$ ，每个输入词 x_n 到每个输出词 $\hat{y}_m$ 的归属定义为 $a_{n,m} = \frac{\partial f(y_m|Z_m)}{\partial \boldsymbol{E}_i[x_n]} \cdot \boldsymbol{E}_i[x_n]^\top$ ，其中 Z_m 表示与第一个 $m-1$ 响应词连接的格式化输入提示，f 是语言模型，$\boldsymbol{E}_i[x_n]$ 表示 x_n 的输入静态嵌入。然后，该成对归因分数被归一化为 $\widetilde{a}_{n,m} = \lceil L \times a_{n,m} / \max_{n'}(a_{n',m}) \rceil$ 。任何等于或小于 b 意志的规范化归因 $\widetilde{a}_{n,m}$ 都强制为 0。词 x_n 的整体归因得分定义为其归因密度，即 $a_n = ||[a_{n,1}, ..., a_{n,M}]||_1 / ||[a_{n,1}, ..., a_{n,M}]||_p$ 。在我们的实验中，我们让 $L = 10$ 、$b = 2$ 、$p = 5$ ，并将 Vicuna-7B-v1.1 视为我们的语言模型 f 。通过平均这些词的归因分数，我们得出了每个句子的重要性。选择重要性得分最高的前 K 个句子作为每个实例的解释。然后，将解释与输出响应连接起来，并馈送到分类器以预测输出是否正确。我们使用 DistillBERT-base（**?**）初始化分类器，并使用学习率 $5e^{-5}$ 和权重衰减 $1e^{-3}$ 在 3 个 epoch 上训练它。为了与此进行比较，我们还考虑将完整段落或人工注释的基本原理作为训练分类器的解释部分。根据之前的研究（**??**），我们通过将所有正确答案与生成的响应精确匹配来评估生成的响应的准确性。Table **??** 报告了五个随机种子的测试集的宏精度、召回率、F1 和 AUC 分数。

Table 3: 利用归因对 ChatGPT 响应进行幻觉检测。

Method	Language Model	Precision	Recall	F1	Accuracy
Random	-	88.41	50.34	64.11	50.59
FacTool	GPT-4	95.30	72.93	82.62	73.04
Vectara	DeBERTa-base	90.29	60.54	72.40	59.45
AttrScore (ours)	Vicuna-7B	90.15	74.21	81.36	70.20
	Mistral-7B	88.74	75.04	81.26	69.57

结果。 在 Table **??** 中，我们观察到，将基本原理与完整上下文隔离开来可以最好地帮助分类器识别响应质量。很明显，当与正确的反应相关联时，解释与人类注释的理由更接近。特别是，从正确答案得出的解释的准确率和召回率指标超过了与错误答案相关的指标。这一发现作为经验证据，强调了基于归因的解释在估计 LLM 生成的响应准确性方面的有效性。

Table 2: 利用归因进行响应质量评估

Setting	Precision	Recall	F1	AUC
Random	49.40	51.79	49.61	49.03
Human Rationale	68.73	66.88	67.57	73.11
Full Paragraph	58.02	58.47	56.89	63.44
Attribution (ours)	63.25	67.69	64.12	71.53

2.2.2 带有归因解释的幻觉检测

本案例研究探讨了 LLM 生成质量的不同方面，重点关注 LLM 生成的反应中幻觉的存在。我们表明，基于归因的解释可以作为检测 LLM 幻觉的指标。幻觉被定义为包含与事实知识相冲突或无法通过事实知识验证的信息的反应（**??**）。例如，如果一个模型被问及一个虚构的实体，如“雷诺伊特国王”，并回答一个关于“三个火枪手”的叙述，声称它与不存在的国王有关，这说明了一种幻觉。这种趋势在指令调整模型中尤为明显，源于他们为满足用户要求而做出的认真努力。当直接命令（“告诉我一个故事”）显着影响生成过程，而指令的主题（“关于雷诺瓦国王”）被忽略时，通常会出现这个问题。基于这一见解，我们根据不同类型提示词的归因分数分布开发了一种幻觉检测器。

数据。 在本案例研究中，我们使用 Hallucination Evaluation Benchmark（**?**）。该数据集中的每个实例都包括一个输入提示、ChatGPT（**?**）生成的响应以及与响应相关的知识。每条知识都有一个关于知识是否有效的人工注释。每个具有至少一个无效知识的实例都被视为幻觉反应。本研究重点关注基准测试中 632 个不太混乱的例子，每个例子几乎都是正确的或几乎全部错误的知识。我们随机选择 80 个% 样本组成训练集，其余组成测试集。

实现。 给定一个查询提示及其 ChatGPT 响应，我们的目标是建立一个分类器来检测响应是否包含幻觉。由于 ChatGPT 的梯度无法访问，因此我们应用 Vicuna-7B 模型作为替代品来计算归因分数。具体来说，我们采用重要性密度分数（**?**）来计算骆马的归因分数。然后，我们使用 NLTK 包来标识每个查询词的词性（POS）标记。最后，每个查询-响应对都用一个 82 维向量表示，其中每个维度表示特定类型的 POS 标记的平均归因分数。我们基于训练集上的 POS 标记归因分数开发了一个支持向量机分类器，并报告了幻觉样本的精确率、召回率、F1 分数，以及表 **??** 中所有样本的准确性。为了与此进行比较，我们还利用微调模型和基于提示的方法作为幻觉检测器基线。

结果。 在表中，我们首先观察到所有方法都表现出比随机策略更好的性能，这表明所提出的方法和基线是有效的幻觉检测器。此外，我们观察到所提出的方法与使用 GPT-4 的 FacTool 相比具有竞争力，在该领域建立了新的基准。值得注意的是，我们方法的归因分数不是使用 GPT-4，而是使用较小的 70 亿参数语言模型计算的。这证明了我们在弱到强泛化方面的实用性和效率，因为我们可以用较小的模型诊断大型语言模型。未来的工作可以考虑提取更有效的特征并使用更强大的分类器。

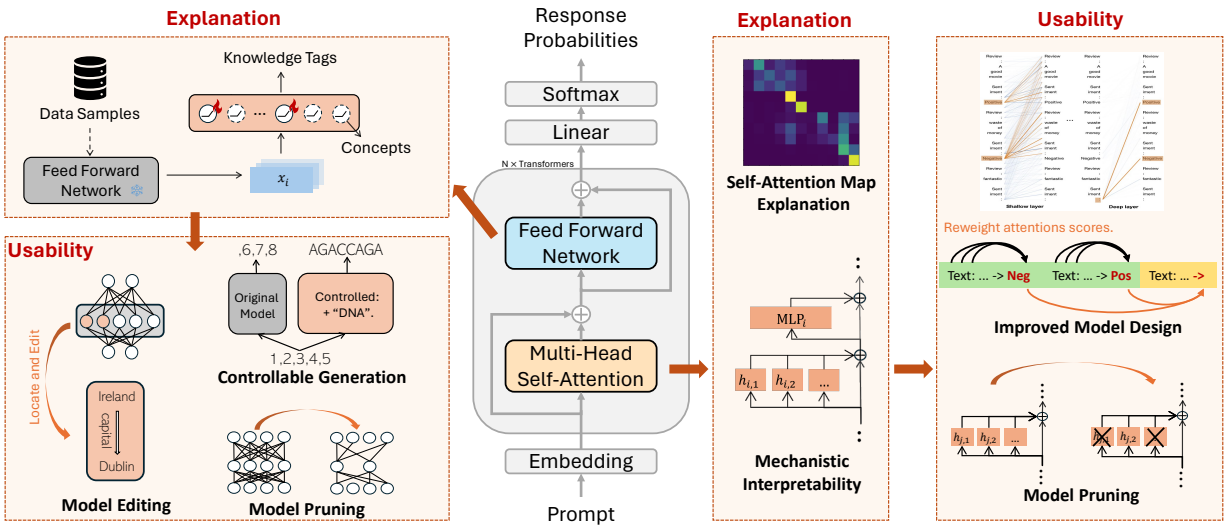


Figure 4: LLM 组件的解释方法及其应用综述。我们根据目标 LLM 模块对方法进行分类：自注意力层和前馈层。

2.3 挑战

2.3.1 如何识别和解释输出的语义？

归因函数 $a_n = p(\hat{y}|x) - p(\hat{y}|\tilde{x}_n)$ 不再忠实地将模型预测归因在人类感兴趣的语义层面，因为模型可以通过不同的响应来表达相同的语义含义。具体来说，模型可以为其原始响应分配比新响应更低的置信度，而两个响应具有相同的语义含义。与传统的分类问题相比，这是一个显著的差异，在传统的分类问题中，目标标签集是手动设计的，因此较低的 $p(\hat{y}|\cdot)$ 表示模型在预测特定语义概念方面的信心较低。以情感分析任务为例，LLM 可能会产生两种不同的响应，共享相同的预测概念，例如“这是一个正面评价”和“观众认为这部电影很棒”。目前基于归因的解释集中在生成的响应中的字面变化，但它们没有研究这些响应的语义含义如何变化。因此，它们不能在语义层面上对模型生成的响应提供足够的解释。在这种情况下，语义层面是输入审查的哪些词导致模型相信它是积极的。未来的工作可能会通过提出指标来评估响应中的语义差异来应对这一挑战。

2.3.2 解释超越归因的 LLM 预测

LLM 生成的多功能性将激发超越传统归因方法的各种解释范式。归因方法旨在通过输入特征的贡献来解释模型输出。此解释任务对于传统的机器学习（ML）模型很有意义，这些模型的输出通常是具有清晰格式（例如分类、回归、对象检测）的单个决策。决策高度依赖于输入要素。然而，LLM 在两个方面与传统的 ML 模型不同。首先，LLM 的生成过程是随机的，例如，即使使用相同的输入，我们也可以通过运行两次 LLM 来获得不同的输出。其次，LLM 在其参数中对丰富的知识进行编码，这些参数与输入无关。这些独特的特性产生了新的解释范式。对于第一个方面，一个有趣的解释任务是理解 LLM 生成的不确定性。例如，研究人员（???）利用预测困惑来检查 LLM 在生成过程中是否有置信度，从而识别置信度较低的预测中的潜在错误。其次，将 LLM 预测归因于他们的编码知识而不是输入模式可以提供一个新的视角。一些研究人员（?）提出了知识边界检测任务，以检测模型是否具有给定问题的特定知识。然而，它并没有将预测归因于特定的知识，因此人类还不能用他们的结果来验证预测过程。

3 通过解释模型组件进行 LLM 诊断和增强

本节讨论解释大型语言模型内部组件的 XAI 方法。此外，它还深入研究了这些方法提供的见解，这些见解有助于完善和增强语言模型的设计。LLM 采用 transformer 作为基本架构，通常由两种类型的主要组件组成：自注意力层和前馈层。在下文中，我们回顾了分别解释这些组成部分的研究。

3.1 了解自注意力模块

多头自注意力模块包含多个自注意力头，捕获不同类型的词与词关系，这些关系用权重 $\boldsymbol{W}_q, \boldsymbol{W}_k \in \mathbb{R}^{D_1 \times D_2}$ 建模。具体来说，单词 w_i 和 w_j 的关系计算为 $\boldsymbol{A}_{i,j} \propto (\boldsymbol{x}_i \boldsymbol{W}_q) \cdot (\boldsymbol{x}_j \boldsymbol{W}_k)^\top$ ，其中 $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathbb{R}^{1 \times D_1}$ 是单词的上下文嵌入。最直接的解释是分析注意力得分矩阵 $\boldsymbol{A}$ 给定一个输入序列来研究单词之间的关系 (**??**) 。在实践中，这些直观的解释将主要用于通过可视化来呈现案例研究。通过这种策略， **?** 进行上下文情感分析的案例研究，他们发现上下文示例中的标签词可以作为最终预测的锚点。具体来说，这些锚点聚合来自示例的信息，以从较低层生成信息表示，而较深层则利用这些表示进行最终预测。这种洞察力促使他们重新加权这些锚点的注意力分数，以实现更好的推理准确性。一些研究人员 (**??**) 扩展了这个框架，通过从感兴趣的词汇中输入单词的静态词嵌入，而不是它们的上下文嵌入，来全局分析注意力权重 $\boldsymbol{W}_q$ 和 $\boldsymbol{W}_k$ 。例如，通过这种方法， **?** 发现指令调优通过鼓励 LLM 编码更多与指令词相关的词间关系，使 LLM 能够遵循人类的意图。另一方面，提出了一些数学模型来理论上解释自注意力机制，如稀疏分布式存储器 (**?**) 和变压器电路 (**?**) 。特别是，变压器电路为基于变压器的模型提供了机理可解释性，将模型分解为人类可理解的部分。尽管这些关于自我注意力的理论分析为未来的研究奠定了坚实的基础，但它们的直接应用在很大程度上没有得到充分的探索。

3.2 了解前馈模块

前馈网络形式化为 $\boldsymbol{x}' = \sigma(\boldsymbol{x}\boldsymbol{W}_u)\boldsymbol{W}_v^\top$ ，其中 $\boldsymbol{x} \in \mathbb{R}^{1 \times D_1}$ 是输入词的中间上下文表示，σ 是非线性操作，$\boldsymbol{W}_u, \boldsymbol{W}_v \in \mathbb{R}^{D_1 \times D_3}$ 是模型参数。前馈网络可以理解为键值存储器 (**??**) ，其中每个键或值分别定义为 $\boldsymbol{W}_u[d] \in \mathbb{R}^{D_1}$ 和 $\boldsymbol{W}_v[d] \in \mathbb{R}^{D_1}$ 。也就是说，每个前馈网络获得 D_3 键值对，称为内存。解释记忆语义的一种简单方法是收集能够最大限度地激活该部分记忆的键或值向量的单词 (**??**) ，这证明了提取的单词列表具有很强的可解释性。但是，重要的是要注意键或值向量是多义的 (**???**) ，这表明这种简单的方法可能无法为每个键值对提供简明的解释。研究表明，每个键值对的词表平均有 3.6 个人类可解释的模式 (**?**) 。为了缓解多义性造成的有限可解释性， **?** 建议解释这些关键或价值向量的主要成分，从而对每个单词列表进行更简洁的解释，例如“医学缩写”和“编程任务和操作”。其他工作通过测量预测在扰乱其相应激活后的变化来检查个体记忆，其中揭示了一些记忆编码特定知识 (**?**) 而另一些记忆捕获一般概念 (**?**) 。通过利用键值记忆的解释，我们可以定位和更新与特定知识相关的记忆，以执行模型编辑 (**????**) ，即修改过时或不正确的知识。这些权重解释的另一个用途是模型修剪，其中 LLM 可以通过专门维护冗余神经元来压缩其初始参数的 66.6 %，从而使推理速度提高约 1.4 倍 (**?**) 。

除了解释和分析模型权重外，一些工作还研究模型激活（例如，$\boldsymbol{x}\boldsymbol{W}_u$ 或 $\boldsymbol{x}'$ ）来解释其功能。探测技术是实现此目的的最流行方法 (**????**) ，识别特定概念是否在表示中编码。其基本思想是开发一个辅助分类器 g ，从 $\boldsymbol{x}'$ 的表示映射到感兴趣的概念空间 $\mathcal{C}$ ，例如语法和词性知识，并且 g 的性能解释了 $\boldsymbol{x}'$ 中编码的信息与 $\mathcal{C}$ 中的概念相关。这种技术激励开发更好的参数效率 (**?**) 、领域特定 (**?**) 和鲁棒 (**??**) LLMs。最近的研究 (**??**) 也应用探测方法来检测 LLM 的知识边界，从而减少幻觉反应。一些研究者 (**??**) 指出了另一个方向来解释模型的隐藏激活，称为字典学习，其动机是假设叠加 (**??**) 。叠加假设 LLM 将学习一组过于完整的非正交特征，从而超越表示空间维度所施加的限制。因此，研究人员旨在重建和解释这些特征，以了解模型的内部结构。实际上，他们开发了一种稀疏自编码器 g 来重建 $\{\boldsymbol{x}_n\}$ 表示，这表明人类可以很好地根据他们最活跃的词来解释 g 学习到的稀疏特征。他们的研究表明，这种方法可用于更可控的发电。具体来说，如果强制激活稀疏特征，则语言模型 f 将更改其响应以执行该稀疏特征的特定行为。例如，给定“1,2,3,4,5,6,7,8,9,10”作为输入，模型最初生成数字作为输出。然而，当他们被迫放大称为“DNA”的稀疏特征的激活时，模型将其输出更改为“AGACCAGAGAGAAC”。总的来说，虽然前馈网络的解释技术主要为模型开发提供见解，但它们在模型编辑和可控生成等领域也显示出有前途的应用。

3.3 挑战

解释内部模块的功能仍处于起步阶段，我们确定了朝这个方向需要解决的两个挑战。

3.3.1 单个模型的复杂性及其相互作用

基于转换器的语言模型包含两种类型的模块，它们基于残差机制 (**?**) 进行协作，这使得后面的模块能够利用、增强和/或丢弃前面模块的输出。形式上，第 l 个模块的输出表示为 $x^l = f^l(x^{l-1}) + x^{l-1}$ ，其中 f^l 可以是自注意力模块或前馈网络。该领域的研究旨在解释不同的模块如何 f^i 和 f^j $i \neq j$ 一起工作。试点研究 (**??**) 发现堆叠的自注意力模块可以形成感应头，这与情境学习能力有很强的相关性。具体来说，感应头

鼓励模型预测单词 “B”，然后是序列 “AB...A”. 他们的研究发现，在预训练 LLM 期间，有一个特定的阶段，在这个阶段，归纳头和上下文学习能力都从模型中出现。沿着这条轨迹，研究人员观察到 LLM 中用于不同任务的不同功能头，例如用于对象识别任务的 “名称移动头” 和 “重复令牌头” (**?**)，用于多项选择题回答任务的 “单字母头” 和 “正确字母头” (**?**)，以及用于通用任务的 “大写头” 和 “反义词头” (**?**)。尽管这些研究确实加深了我们对跨模块效应的理解，但它们的分析是基于特定的任务或情景，因此不确定这些发现是否具有广泛的普遍性。

3.3.2 多义和叠加假设的本质

解释单个神经元（权重矩阵的一行/列向量）的功能在分析大型语言模型时失败，因为单个神经元可以被多种不同的含义激活，称为多义 (**???**)。这种性质导致可解释性差：解释单个神经元通常不能反映一个简洁的人类概念。一些研究者 (**??**) 认为这种现象是由模型学习的一组过完整的特征叠加引起的。基于这个假设，我们可以通过分解模型权重来重建大量特征来达到另一个层次的解释。然而，这种方法的两个关键问题仍然不清楚：(1) 我们如何确保重构的特征忠实地代表模型学习的潜在特征？(2) 我们如何用人类语言来解释我们重建的特征？

4 LLM 调试与基于样本的解释

在本节中，我们将讨论 LLM 的基于样本的解释策略，旨在将 LLM 生成的答案追溯到语料库中的特定训练样本（即文档）或文档片段。基于样本的 LLM 解释的效用是多方面的。首先，将 LLM 的预测追溯到训练样本，可以为生成结果提供证据，从而在出现错误时有利于模型调试，并在结果准确时提高用户对模型的可信度。此外，它还可以帮助研究人员了解 LLM 如何从训练样本中泛化。如果 LLM 的输出可以追溯到直接从训练数据中拼接的精确子序列，则可能表明 LLM 只是在记忆数据。相反，如果生成结果和影响训练样本是抽象相关的，则可能表明 LLM 可以通过从输入提示推理来理解概念并生成响应。

在本节中，我们首先系统地回顾了传统的基于样本的解释策略，包括基于梯度的方法和基于嵌入的方法，以及将它们推广到 LLM 的一些初步探索。然后，我们分析了将上述策略推广到具有独特变压器结构和前所未有的参数数量的 LLM 相关的挑战。最后，我们讨论了应对挑战的见解，以及值得进一步研究的开放挑战。

4.1 基于样本的解释的文献综述

在本节中，我们将输入空间和输出空间分别表示为 $\mathcal{X}$ 和 $\mathcal{Y}$。在大型语言模型（LLM）的上下文中，$\mathcal{X}$ 是称为提示的标记序列的空间，$\mathcal{Y}$ 可以是分类任务中的离散标签空间，也可以是生成任务中作为输出的标记序列 [*] 空间。因此，我们有一个训练数据集 $\mathcal{D}_{train} = \{z_i = (x_i, y_i)\}_{i=1}^N$ 其中包含从联合空间 $\mathcal{X} \times \mathcal{Y}$ 中抽取的 N 个样本，在该数据集上使用预训练参数 $\hat{\theta} \in \mathbb{R}^P$ 训练 LLM 模型 f_θ。我们还有一个感兴趣的测试样本 $z = (x, y)$，我们想解释基于 $\mathcal{D}_{train}$ 中的训练样本（可以看作是信息源）从 x 生成 y。基于样本的解释的目的是测量训练样本 $z_i \in \mathcal{D}_{train}$ 或 z_i 中某个片段的影响，以便可以很好地解释 LLM 的生成，并由选定的训练样本提供支持。

4.1.1 影响基于函数的方法

量化训练样本 z_i 在 $\mathcal{D}_{train}$ 数据集中对测试样本 z 的影响的一种策略是通过影响函数 (**??**)。它测量测试样本 z 的预测损失 $\mathcal{L}(z, \theta)$ 的变化，当训练样本 z_i 在模型训练期间在数据集 $\mathcal{D}_{train}$ 中经历假设修改时。此修改会导致一组最优模型参数发生更改，表示为 $\hat{\theta}_{-z_i}$。训练样本最常见的修改是将其从数据集中删除，其中删除训练样本 z_i 对测试样本 z 损失的影响可以计算如下：

$$\mathcal{I}(z_i, z) = -\nabla_\theta \mathcal{L}(z, \hat{\theta})^\top \boldsymbol{H}_{\hat{\theta}}^{-1} \nabla_\theta \mathcal{L}(z_i, \hat{\theta}), \tag{1}$$

其中 $\nabla_\theta \mathcal{L}(z, \hat{\theta})$ 是在最优参数 $\hat{\theta}$ 处评估的测试样本上 z $\mathcal{L}$ 损失函数的梯度，$\boldsymbol{H}_{\hat{\theta}} \overset{\text{def}}{=} \frac{1}{N} \sum_{i=1}^N \nabla_\theta^2 \mathcal{L}(z_i, \hat{\theta})$ 表示参数 $\hat{\theta}$ 处 LLM 模型的 Hessian 矩阵。如果我们将 $\hat{\theta}$ 中的参数数表示为 P，则 Hessian 矩阵 $\boldsymbol{H}_{\hat{\theta}}$ 的 naïve 反演会导致时间复杂度 $\mathcal{O}(NP^2 + P^3)$ 和空间复杂度 $\mathcal{O}(P^2)$ (**?**)，这对于大型模型来说显然是不可行的。为了提高效率，**?** 采用迭代逼近过程，即 LiSSA（线性时间随机二阶算法），计算方程（**??** ）中的 Hessian-Vector

[*] 在这里，我们应该注意，在具有语言建模的 LLM 的预训练阶段，该模型要么预测 x_i 中的掩码标记（即掩码语言建模），要么自回归地预测 x_i 中的下一个标记（即因果语言建模）。因此，在一些作品中，省略了 y_i，只包括 x_i 进行讨论。

Product（HVP），其中内存复杂度可以降低到 $\mathcal{O}(P)$ ，时间复杂度可以降低到 $\mathcal{O}(NPr)$ （r 是迭代次数）。为了进一步降低复杂性，**?** 提出了方程（**??** ）的替代方案，即 TracIn，它通过计算在模型训练期间小批量中包含 z_i 时对 z 损失的总减少量来衡量 z_i 对 z 的影响。TracIn 测量可以公式如下：

$$\mathcal{I}_{\text{TracIn}}\left(z_i, z\right) = \sum_{t:z_i \in \mathcal{B}_t} \mathcal{L}\left(z, \theta_t\right) - \mathcal{L}\left(z, \theta_{t+1}\right) \approx \frac{1}{b} \sum_{t:z_i \in B_t} \eta_t \nabla_\theta \mathcal{L}\left(z_i, \theta_t\right) \cdot \nabla_\theta \mathcal{L}\left(z, \theta_t\right), \tag{2}$$

其中 $\mathcal{B}_t$ 是训练期间输入模型的第 t 个小批量，θ_t 是第 t 步的参数检查点，η_t 是步长，b 是小批量的大小。根据上述等式，TracIn 仅利用梯度项，其中 Hessian $\boldsymbol{H}_{\hat{\theta}}$ 从影响测量中移除。这大大提高了效率。然而，从计算和内存的角度来看，这种复杂性对于大型模型来说仍然令人望而却步。此外，TracIn 只能估计添加/删除样本对损失的影响，其中方程（**??** ）中定义的香草影响函数的变体可以测量训练样本 z_i 的其他修改的影响，例如扰动（例如，屏蔽文档 x_i 的一段）。为了适应方程（**??** ）的香草影响函数来解释变压器，**?** 建议使用 Alnordi 迭代 (**?**) 在随机采样子集 $\mathcal{D}_{sub}$ 上找到 Hessian 矩阵的主要特征值和特征向量，$|\mathcal{D}_{sub}| \ll |\mathcal{D}_{train}|$ 。在这种情况下，对角线化的 Hessian 可以廉价地缓存和反转，从而可以大大降低计算和内存复杂性。以前的工作主要集中在降低计算单个训练样本影响的复杂性。观察到在 z 上找到最有影响力的训练样本需要迭代方程（**??** ）整体 N 训练样本，**?** 建议使用快速 KNN 从 $\mathcal{D}_{train}$ 作为候选者中预过滤一小部分具有影响力的数据点，以解释小型预训练语言模型，而 **?** 建议迭代地找到一个梯度最相似的小子集 $\mathcal{D}_{sub} \subset \mathcal{D}_{train}$ 下游任务示例。最近，**?** 建议使用特征值校正的克罗内克因式近似曲率（EK-FAC）近似将影响函数扩展到最多 52B 参数的 LLM。对于适应，只考虑由多层感知器（MLP）层介导的影响，而来自自注意力层的权重是固定的，因为以前的工作已经证明知识主要编码在 MLP 层（**?**）。此外，基于来自不同 MLP 层的权重是独立的假设，EK-FAC 近似影响可以表述为每层介导的影响之和：

$$\mathcal{I}_{\text{EKFAC}}(z_i, z) = \sum_l \nabla_{\theta^{(l)}} \mathcal{L}(z, \hat{\theta})^\top (\hat{\boldsymbol{G}}_{\hat{\theta}^{(l)}} + \lambda^{(l)} \boldsymbol{I})^{-1} \nabla_{\theta^{(l)}} \mathcal{L}(z_i, \hat{\theta}), \tag{3}$$

其中 $\theta^{(l)}$ 表示第 l 个 MLP 层的权重，$\hat{\boldsymbol{G}}_{\hat{\theta}^{(l)}}$ 是 $\theta^{(l)}$ 的 EK-FAC 近似高斯-牛顿黑森。由于 L 小 $K_l \times K_l$ 矩阵（即 $\mathcal{O}(L \times K_l^3)$ ）的反演比大 $LK_l \times LK_l$ 矩阵（即 $\mathcal{O}((LK_l)^3)$ ）的反演效率高得多，因此 $\mathcal{I}_{\text{EKFAC}}$ 可以适应非常大的模型，更不用说利用 EK-FAC 特性可以进一步简化 HVP 了。

最近，基于影响函数的解释已被用于 LLM 的有效微调，其中影响函数已被用于为特定的下游任务选择一小部分训练样本，给定少量验证样本，其中训练开销可以大大改善 (**?**) 。

4.1.2 基于嵌入的方法

基于样本的解释的另一种策略涉及利用转换器架构中的隐藏表示，该架构被识别为从文本数据编码高级语义，以计算 z 和 z_i 之间的语义相似性。相似性也可以用来衡量 z_i 对 z 的影响，因为解释 (**?**) 。具体来说，**?** 建议通过将输入和输出连接起来表示训练样本 z_i 并测试样本 z $z_i^{cat} = [x_i||y_i]$ 、$z^{cat} = [x||y]$ 。对于输出 y 与输入提示 x 位于同一标记序列空间中的生成任务，串联是可行的。然后可以按如下方式计算 z_i 和 z 之间的相似性：

$$\mathcal{I}_{\text{emb}}\left(z_i, z\right) = \frac{f_{\hat{\theta}}^{(l)}(z_i^{cat})^\top \cdot f_{\hat{\theta}}^{(l)}\left(z^{cat}\right)}{\left\| f_{\hat{\theta}}^{(l)}(z_i^{cat})^\top \right\| \left\| f_{\hat{\theta}}^{(l)}\left(z^{cat}\right) \right\|}, \tag{4}$$

其中 $f_{\hat{\theta}}^{(l)}$ 是输出预训练 LLM $f_{\hat{\theta}}$ 的第 l 层中间激活的子网络。方程（**??** ）与方程（**??** ）中定义的香草影响函数及其在方程（**??** 中定义的 TracIn 备选函数具有相似的形式，该替代函数为数据集中每个训练样本 z_i 的被解释者 z 分配一个分数 $\mathcal{I}$ $\mathcal{D}_{train}$ 作为样本 z_i 的解释置信度。

与上一部分介绍的影响函数方法相比，基于嵌入的方法具有计算效率，因为对于每个被解释者 z ，来自训练样本 z_i 的解释分数只需要一个 transformer 网络的前向传递。此外，对于不同的训练样本，可以很容易地并行计算。然而，缺点也很明显：这些方法缺乏理论基础，可能无法识别出与测试样本语义不相似的重要训练样本。考虑以下玩具示例：训练样本 $z_i =$ （"1+1="，"2"）和 $z_j =$ （"2+2="，"4"）使 LLM 获得进行算术计算的能力，这就解释了为什么用 $x =$ "100+100" 提示模型会得到 $y =$"200" 的结果。但是，当通过方程（**??** ） (**?**) 计算时，测试样本 z 与两个训练样本 z_i 和 z_j 之间的嵌入可能非常不同。因此，基于嵌入的方法可能无法忠实地找到训练样本，其中解释需要超越语义相似性的泛化能力。

Table 4: EK-FAC 近似影响函数对已建立的 SciFact-Inf 数据集的有效性。时间（Pre.）代表预计算 $\boldsymbol{Q}_A$ 、$\boldsymbol{Q}_S$ 和 $\boldsymbol{\Lambda}$ 的时间。时间（Inf.）代表计算每个测试样本 100 个训练样本的影响的时间。GPT2-1.5B、LLaMA2-7B、Mistral-7B 的实验是在 4 个 A100 GPU 上完成的，而 LLaMA2-13B 的实验是在 8 个 A100 GPU 上完成的。

Strategy	LLM	Recall@5	Recall@10	Time (Pre.)	Time (Inf.)
Random	-	0.0100	0.0200	-	-
Inf. Func.	GPT2-1.5B	0.6368	0.7363	0h 27min	0min 28sec
	Mistral-7B	0.6418	0.6866	2h 05min	1min 47sec
	LLaMA2-7B	0.8063	0.8308	1h 37min	1min 34sec
	LLaMA2-13B	0.7811	0.8940	3h 11min	3min 08sec

4.2 案例研究：基于 EK-FAC 的影响估计

在这一部分中，我们实现了 **?** 提出的 EK-FAC 近似影响函数，并验证了其在具有数十亿个参数的 LLM 上的可扩展性和有效性，包括 GPT2-1.5B (**?**) 、LLaMA2-7B (**?**) 、Mistral-7B (**?**) 和 LLaMA2-13B。

4.2.1 试验设计

我们使用 SciFact 数据集 (**?**) 作为语料库，其中包含来自基础科学和医学领域的 5,183 篇论文的摘要。通过对预训练的 LLM 进行 20,000 次迭代，以 AdamW (**?**) 为优化器，将学习率和权重衰减分别设置为 1e-5、1e-3，得到被解释的 LLM。然后，我们使用来自语料库的 500 个样本来估计（i）激活和激活前伪梯度的非中心协方差矩阵 $\boldsymbol{Q}_A^{(l)}$ ，$\boldsymbol{Q}_S^{(l)}$ 和（ii）每个选定的密集层的投影伪梯度 $\boldsymbol{\Lambda}^{(l)}$ 的方差 l ，并将它们缓存在硬盘上（详细信息参见 Eqs. **?** 中的（16）和（20））。我们为 GPT2-1.5B 选择 `c_fc` 层，为 LLaMA2-7B、Mistral-7B 和 LLaMA2-13B [*] 选择 `gate_proj` 层。

为了进行评估，我们从语料库中随机选择 200 个样本来构建测试集，我们将其命名为 SciFact-Inf。具体来说，对于第 j 个选定的样本 $z_j = (x_j, x_j)$ （这里 $y_j = x_j$ 因为标签等于语言建模中的输入），我们使用 x_j 中的前三个句子，即 $\hat{x}_j$ ，用微调的 LLM 生成完成 $\hat{y}_j$ （这里，$\hat{y}_j$ 不等于 x_j 中的其余句子），我们的目标是通过训练解释从 $\hat{x}_j$ 中生成 $\hat{y}_i$ 通过 EK-FAC 的样本近似影响分数在公式（**??** ）中定义。理想情况下，j 训练样本 z_j 本身应该是最有影响力的样本，因为测试样本 $\hat{z}_j$ 的 $\hat{y}_j$ 生成，这有助于对方程（**??** ）的有效性进行定量分析。

在我们的实现中，对于每个测试样本 $\hat{z}_j$ ，我们首先计算影响 $\mathcal{I}_{\text{EKFAC}}(z_i, \hat{z}_j)$ 的 EK-FAC 近似 HVP 部分，即 $\sum_l \nabla_{\theta^{(l)}} \mathcal{L}(\hat{z}_j, \hat{\theta})^\top (\hat{\boldsymbol{G}}_{\hat{\theta}^{(l)}} + \lambda^{(l)} \boldsymbol{I})^{-1}$ ，该部分 z_i 所有训练样本共享。具体来说，我们记录逐层梯度 $\nabla_{\theta^{(l)}} \mathcal{L}(\hat{z}_j, \hat{\theta})$ 并使用缓存的 $\boldsymbol{Q}_A^{(l)}$ 计算 HVP，$\boldsymbol{Q}_S^{(l)}$ **?** 中的公式（21）。然后，我们通过候选训练样本（1 个正样本和 99 个负样本），计算梯度 $\nabla_{\theta^{(l)}} \mathcal{L}(z_i, \hat{\theta})$ ，并将近似 HVP 的内积作为逐层影响。最后，将层向影响总结为方程（**??** ）作为总影响 $\mathcal{I}_{\text{EKFAC}}(z_i, \hat{z}_j)$ 。我们对影响进行排名，并计算正训练样本的前 K 命中率。

4.2.2 结果与分析

实验结果总结于表 **??** 。从表 **??** 中我们可以发现，即使只考虑一小部分致密层介导的影响，EK-FAC 近似影响函数在找到对测试样本生成影响最大的训练样本方面也取得了良好的准确性。此外，我们发现计算基于 EK-FAC 的影响的主要计算瓶颈是估计协方差 $\boldsymbol{Q}_A^{(l)}$ 、$\boldsymbol{Q}_S^{(l)}$ 和方差 $\boldsymbol{\Lambda}^{(l)}$ ，当使用 500 个训练样本进行估计时，这可能需要数小时。然而，在估计之后，计算每个测试样本的 500 个训练样本的影响相对便宜，对于 8 个 A100 GPU 上的 13B LLaMA2 模型来说，这大约需要 3 分钟。这通过假设不同密集层的独立性并使用 EK-FAC 来简化计算，证明了基于 EK-FAC 的影响的可扩展性。

4.3 挑战

总体而言，通过追溯训练样本来解释 LLM 的生成仍然是一个新兴领域。需要解决悬而未决的问题，以进一步推进该领域的发展。在本节中，我们确定了以下三个主要挑战，这些挑战可以作为未来探索的方向。

[*] 所有实现和层名称都基于 huggingface 转换器，详细信息可以在 https://huggingface.co/docs/transformers/en/index 中找到。

4.3.1 对可扩展性的有力假设

现代 LLM 中前所未有的参数数量导致基于样本的解释策略存在严重的可伸缩性问题。这对于基于梯度的方法尤为明显，因为方程（**??** ）中的 HVP 会引起高计算和空间复杂性。为了解决瓶颈问题，通常需要强有力的假设才能使其适用于大型模型。例如，TracIn (**?**) 通过一阶近似简化了方程（**??** ）中的二阶项。**?** 认为黑森人是低等级的。**?** 假设来自 LLM 不同层的权重是独立的，以及不同步骤中的标记，因此可以适当地应用 EK-FAC 来近似影响函数。从上面的分析中，我们可以发现，虽然 **?** 的方法具有最好的可扩展性，但它也具有最强的假设，这在实践中可能无法成立。虽然计算效率很高，但基于嵌入的方法隐含了语义同样意味着可解释性的假设，我们已经证明情况可能并非总是如此。因此，未来需要研究如何在弱假设下提高可扩展性。

4.3.2 可解释性与可解释性可理解性

尽管影响/嵌入相似性的优势在于提供特定训练样本的定性测量作为 LLM 生成的解释，但已识别样本的可理解性仍然很弱，其中所选训练样本与生成之间的联系可能对人类来说是无法理解的。具体来说，**?** 警告说，训练代币的影响分数符号可能很难被人类与对生成结果的积极或消极影响联系起来。这严重损害了已识别训练样本的可用性。此外，**?** 还发现，由于 LLM 通常不会被训练到最小值以避免过拟合（并且由于过度参数化，局部最小值的数量可能很大），因此方程（**??** ）中定义的影响与在 z 处去除样本 z_i 的反事实损失之间的联系也很弱。对于基于嵌入的方法，由于大多数 LLM 模型都是黑盒变压器模型，因此嵌入的相似性也很难被人类解释; 因此，当务之急是提高已识别训练样本的可解释性，以便回溯变得更有意义。

4.3.3 面向 LLM 的基于样本的解释

最后，我们观察到基于梯度和基于嵌入的方法都与 LLM 以及骨干变压器网络松散地连接在一起。例如，像 TracIn (**?**) 这样的算法旨在将影响函数扩展到大型模型，而这些模型并不特定于 LLM。同样，**?** 中提出的基于嵌入的方法适用于大多数具有潜在表示的机器学习模型。**?** 通过利用骨干变压器的知识神经元假设来考虑 LLM 的特殊性，(**?**) 简化影响函数，其中考虑的权重被限制在 MLP 层上，这可能没有充分利用变压器的性质。因此，如何进一步利用 LLM 和骨干变压器的特性来设计 LLM 定制的基于样本的影响/相似性（以减少计算/空间开销或提高解释质量）对未来的工作具有很大的希望。

5 可信赖 LLM 和人类对齐的可解释性

在前面的章节中，我们探讨了使用解释技术来评估和提高 LLM 的性能。在本节中，我们将重点转移到研究 LLM 可信度上。随着 LLM 越来越多地融入日常生活的各种应用，包括医疗保健、金融和法律咨询等高风险领域，他们的反应不仅要准确，而且要符合人类道德标准和安全协议（**??**）。因此，有必要将解释的范围从仅仅评估 LLM 的准确性扩大到审查其可信度。在这里，我们深入探讨了前几节中讨论的解释技术如何有助于评估 LLM 在安全性、隐私性、公平性、毒性和诚实性等可信度的关键方面。值得注意的是，虽然可解释性本身是可信度的一个方面，但它有望成为解决其他可信度问题的基础工具。

众所周知，

5.1 安全

LLM 容易受到攻击和利用，例如传播错误信息、发起网络钓鱼攻击和毒害训练数据（**?**）。为了提高安全性，LLM 旨在拒绝某些类型的提示，这些提示可能会导致有害内容的生成，例如，通过排除可能从训练阶段引发不安全输出的提示。但是，越狱技术可以规避这些限制措施并操纵 LLM 生成恶意内容。恶意用户（即攻击者）可以制作特殊的提示，迫使或诱使 LLM 优先遵循指令而不是拒绝（**??**）。例如，通过前缀注入，攻击者可以使用不太可能被拒绝的分发外提示前缀（**??**）。另一种称为垃圾抑制的方法涉及指导或说服模型忽略已建立的安全协议（**??**），然后使用指令跟随能力来执行攻击。

现有的方法主要依靠提示工程来攻击 LLM，但它们通常攻击成功率低，时间成本高（**?**）。因此，通过理解和设计 LLM 的潜在表示，解释方法提供了一种可行的方法来设计高级攻击并发现 LLM 的潜在漏洞（**?**）。例如，最近的一项工作通过使用表示工程解释 LLM 的潜在空间来提取“安全模式”。具体而言，可以从恶意查询和良性查询之间的激活差异中捕获这些模式。差分向量维度的显著部分被定位并用于生成安全模式的特征。安全模式反映了 LLM 内部保护机制。规避这些模式会导致新的攻击，这有助于探索 LLM 的潜在漏洞（**?**）. 此外，对微调的更深入理解可以阐明现有安全措施的可靠性。特别是，**?** 使用网络修剪、注意力图激活和探测分类器来跟踪模型功能从预训练到微调的变化。这些工具有助于找到识别关键神经元的显着权

重，以恢复训练前的能力。这些神经元已经证明，通过对其他不相关任务的微调，可以很容易地消除在微调期间获得的能力。这一发现使人们对 LLM 中当前安全对齐的鲁棒性产生了怀疑。

5.2 隐私

最近的研究表明，ChatGPT 等 LLM 可以通过一种称为发散攻击的方法泄露大量训练数据。这些攻击利用特制的提示来引导模型远离其标准的聊天机器人式生成（?）。通过这种手段暴露私人数据的风险对道德上负责任的模型的发展构成了严重挑战。类似于越狱攻击的策略使这个问题变得更加复杂，在这种策略中，利用错位来诱使 LLM 通过发行版外提示以非常规的“开发人员模式”运行（?）。传统的数据预处理技术，如数据清理（?），由于训练数据规模巨大，作为 LLM 的防御是不切实际的。

增强 LLM 隐私涉及两种策略方法：（1）防止模型记忆敏感数据，以及（2）建立保护措施，防止在内容生成过程中发布敏感信息。后者可以使用越狱防御中使用的技术，将请求私人信息的提示视为潜在的恶意提示。前一种方法需要确定 LLM 是否拥有特定知识，传统上通过制定问答（QA）任务来测试，以确定 LLM 是否可以提供答案。然而，由于 LLM 对 QA 提示的措辞很敏感，而最佳提示通常是未知的，因此这种方法面临局限性。为了应对这一挑战，解释技术可以作为一种工具，确认 LLM 是否已经内化了某些知识。例如，通过解释事实知识与神经元激活（???）之间的关系，我们可以研究事实知识是否存储在转换器中以及存储在何处。此外，? 最近提出了“知识边界”的概念，并开发了一种基于梯度的方法，以探索 LLM 是否独立于输入提示掌握某些知识。

5.3 公平

尽管 LLM 具有强大的生成能力，但它们的广泛应用也带来了对加剧社会偏见问题的担忧，因为 LLM 能够在人类生成的语料库中学习社会偏见（?）. 例如，在性别偏见案例中，由于性别刻板印象，“[他] 是医生”比“[她] 是医生”的可能性要大得多。在本小节中，我们重点关注公平问题，这些问题涉及人类社区中与种族、性别和年龄相关的偏见（?）。有大量的文献通过各种测试来量化 LLM 中的公平性问题（?）。解释通过为缓解偏见提供独特的视角来补充这些方法，重点是揭示偏见嵌入 LLM 的机制。该领域的一个研究方向是研究有偏见的注意力头。例如，? 通过探测注意力头并通过 Shapley 值评估其归因来检测刻板印象编码（?）。结果显示，在六个基于 transformer 的模型中，大约 15 个% 到 30 个% 的注意力头与刻板印象有关。这些注意力负责人倾向于专注于维持各种刻板印象，为开发有效的去偏见技术提供了潜在的途径。此外，使用基于梯度的指标对头部偏差进行评分提供了另一种识别有偏差的头部的方法（?）. 此外，最近的工作已经对 LLM 表示进行了审查（?）。通常，与给定概念或功能相关的特定模板是事先设计的。然后，使用主成分分析（PCA）检查与概念或功能密切相关的表示。从这个分析中，从第一个主成分推导出一个向量来预测某个偏差。

为了实现公平的模型预测，已经提出了多种缓解技术来消除模型的偏差。一个工作流建议在嵌入级别消除 LLM 的偏见，通过各种方法改进嵌入。例如，最近的一项工作试图以最小的改动来改变偏置嵌入，使它们与中性嵌入正交（?）。此外，一些研究集中在消除注意力头部层面的偏见。? 通过修剪导致某些偏见的注意力头来解决这个问题。同样，另一项研究调整了运动修剪（?），以动态选择低偏差注意力头（?）的最佳子集。除了修改嵌入和修剪注意力之外，另一种策略涉及针对已知会传播偏见的特定神经元组。它通过重新训练这些神经元的权重向量来消除偏差（?）。此外，还可以从以数据为中心的角度使用一些训练样本（?）来缓解偏差。这项工作使用预训练模型来查找最有偏差的训练样本，然后修改这些样本以微调模型。

5.4 毒性

毒性是 LLM 可能产生的另一种形式的有害内容。之所以会出现这个问题，是因为 LLM 通常是在未经彻底过滤的大量在线文本语料库上进行训练的，其中包含难以完全消除的毒性元素。毒性可以通过解释 LLM 组件（如前馈层和注意力头）来识别。例如，最近的工作揭示了毒性如何在 LLM 中表示，通过识别在 MLP 层内促进毒性的多个向量，并通过奇异值分解进一步识别相关维度（?）。此外，在每层表示中探索几何结构提供了另一种检测毒性的方法。? 在 MLP 中应用样条公式来提取七个特征输入特征，证明了它们在描述提示域和分类有毒言论方面的实用性。

对 LLM 中毒性如何表现的见解也阐明了缓解策略。由于发现可以通过操纵相关载体来降低毒性，? 开发了一种称为直接偏好优化（DPO）的方法，其中使用成对的有毒和无毒样品来微调模型，从而促进无毒含量。通过检查微调过程中参数矩阵的变化，它证实了即使对这些关键载体进行微小的调整也可以降低毒性。基于

LLM 的表征由注意力层的输出更新（?）的观察结果，另一项工作试图通过识别“毒性方向”，然后在相反的方向上调整表征来降低毒性（?）。

5.5 真实性

LLM 的一个突出缺点是它们倾向于自信地产生虚假陈述。这些陈述分为两大类：1）与模型中学到的知识相矛盾的陈述，这个问题通常与模型的诚实性有关;2）事实不正确的陈述，似乎是由模型捏造的，这种现象通常被称为幻觉。在下文中，我们将深入研究各种方法，旨在通过利用可解释性工具来理解上述两种行为。

5.5.1 诚实

LLMs 的诚实描述了模型根据其学习的信息生成真实陈述的能力，其中不诚实的行为会显着损害 LLM 的可信度。已经进行了大量研究，通过观察 LLM 内部来了解不诚实行为是如何以及为什么发生的。一项值得注意的工作试图通过训练分类器来预测陈述的准确性来区分不诚实（?）。分类器只是在 LLM 隐藏层的激活之上进行训练。这些激活是从 true 和 false 陈述生成的。分类器的精度范围介于 60 % 和 80 % 之间，这表明 LLM 可能在内部意识到其输出的真实性（?）。此外， ? 的研究将不诚实行为定位在注意力头部的层面上。本研究采用特制的提示来诱导不诚实的响应，然后根据模型对真/假陈述的激活来训练逻辑分类器。它还采用激活补丁来用诚实的激活代替撒谎的激活。有趣的是，这两种方法都见证了第 23 层 – 第 29 层在翻转不诚实行为中的重要性。此外，另一种流行的方法试图研究真/假陈述的几何结构 (?)。通常，这些结构是通过将语句的表示投影到两个主成分上来可视化的。可以推导出清晰的线性结构和真相方向，以减轻不诚实行为。

5.5.2 幻觉

LLM 中的幻觉可能是由于数据质量差、偏见、知识过时和缺乏明确的知识而产生的（??）. 然而，LLM 是否意识到他们的幻觉行为仍然是一个悬而未决的问题。最近的研究通过研究模型的隐藏表示空间（?）来研究这个问题。它检查涉及问题的三个隐藏状态，即正确答案和错误答案，用于计算“意识”分数。该指标量化了 LLM 对自己答案的不确定性，发现对抗性诱导的幻觉可以提高模型的意识。此外，? 说明了模型输出与其内部激活之间的主要差异，将这些差异确定为幻觉的潜在来源。通过对每个注意力头的激活进行线性探测分类器训练，可以识别出最专业的注意力头。随后使用已识别的专用探针对正交探针进行训练。实验表明，“真理”可能存在于一个子空间中，而不是一个单一的方向（?）。另一项工作通过分析通过扰动的源令牌贡献模式来研究幻觉的来源（?）。他们的研究结果表明，幻觉可能源于模型对一组受限制的源令牌的过度依赖。此外，源代币贡献的静态分布，称为“源贡献静态”，可以用作幻觉的另一个指标。

基于上述对 LLM 幻觉的洞察，? 应用 PCA 推导正确答案最终隐藏状态的方向，并以此方向增强隐藏表征以减少幻觉。相比之下， ? 采用了不同的方法，通过干预顶部 K 专门的注意力头，同时最大限度地减少模型中其余注意力头的影响。与确定单一原则方向的 PCA 不同，这项工作采用两种不同的技术来寻找多个干预方向。首先，他们使用每个探针超平面的正交向量，这类似于 PCA。其次，它们利用连接真分布和假分布均值的向量（?）。从均值偏移得出的向量已被证明比来自探针分类器的向量更有效，这为识别真理方向提供了另一种可行的策略。

5.6 挑战

我们从两个方面讨论了使用解释来提高模型可信度和增强一致性的挑战：1）现有检测技术的局限性，以及 2）基于解释的缓解策略的缺陷。

5.6.1 现有检测方法的挑战

目前的检测方法主要集中在 LLM 模块的级别，例如层、注意力头和表示。然而，我们仍然缺乏对知识如何在 LLM 中编码的更细粒度的理解，例如在单个神经元和权重参数的水平上。此外，我们缺乏可靠地识别这些知识的一般和强有力的策略。例如，为了识别性别偏见，可以检查注意力头，然后修剪相关的头（?）。这种方法需要单独分析每个模型，而不是采用通用方法。此外，现有的定位方法要么依赖于探测分类器，要么依赖于随意的清理，这可能并不可靠。对于探测分类器，用于训练这些分类器的预先设计的偏差对其性能至关重要。另一方面，随意清洁通常会引入新的变量，使分析复杂化。

5.6.2 缓解策略的挑战

由于 LLM 是在大量数据语料库上训练的，因此从数据角度缓解上述可信度问题是不切实际的。人们普遍认为，XAI，即理解 LLM 的内部机制，是解决这些问题的基本途径。因此，LLM 的缓解方法通常是基于解释而开发的。现有的解释是使用机理可解释性和表示工程（?）的技术实现的。虽然这两种方法都可以缓解这些问题，但它们未能完全解决这些问题。例如，主成分分析（PCA）通常用于查找与这些问题相关的几何结构，但无法解决这些问题。另一种流行的缓解方法是随意清理，它引入了某些“积极”激活，以取代那些被证明对特定问题负责的激活。但是，确定的方向和修补的激活只能在一定程度上缓解问题。此外，对表示或激活的更改也可能影响模型能力的其他方面，我们尚无法评估这些方面。

6 通过可解释提示增强 LLM

LLM 与传统机器学习模型的一个关键区别在于 LLM 在模型推理过程中能够接受灵活操作的输入数据，即提示（?）。LLM 在生成输出时通常优先考虑这些提示中显示的信息。因此，为了缓解 LLM 预测中的不透明性问题，我们可以用可理解的内容来增强提示，然后优先考虑 LLM 的固有和隐含知识。这些丰富的提示可以包括特定于域的见解、上下文信息或分步推理链。作为回应，LLM 可能会在推理过程中揭示他们的决策过程，从而提高他们行为的可解释性。

6.1 思维链（CoT）提示

思维链（CoT）方法显着增强了 LLM 在处理复杂任务方面（?）. 虽然 LLM 擅长生成类似人类的反应，但它们的推理过程往往缺乏透明度。这种限制使用户难以评估回答的可信度，特别是对于需要详细推理的问题。为了弥合这一差距，最近的努力将上下文学习与人工设计的解释直接结合到提示中（????）。在这些方法中，CoT 方法通过采用显性知识来指导推理过程而脱颖而出。从形式上讲，我们将语言模型定义为 f_θ，将输入提示定义为 $X = \{x_1, y_1, x_2, y_2, ..., x_n\}$，其中 $x_1, y_1, x_2, y_2, ..., x_{n-1}, y_{n-1}$ 表示上下文学习的示例问答对，x_n 是实际问题。在标准问答场景中，我们的模型输出为 $y_n = \arg\max_Y p_\theta(Y|x_1, y_1, x_2, y_2, ..., x_n)$。然而，这种方法并不能深入了解答案 y_n 背后的推理过程。因此，CoT 方法建议包含 i 第 -个上下文示例 e_i 的人为解释，从而修改输入格式 $X = \{x_1, e_1, y_1, x_2, e_2, y_2, ..., x_n\}$。给定输入，模型不仅会输出 y_n，还会输出生成的解释 e_n：

$$e_n, y_n = \arg\max_Y p_\theta(Y|x_1, e_1, y_1, x_2, e_2, y_2, ..., x_n). \tag{5}$$

除了允许与 LLM 进行更透明和更易于理解的交互外，CoT 方法实际上也很有用，因为它通过为用户打开一个窗口来控制模型的思维过程，从而增强了 LLM 的功能。具体来说，CoT 方法的用处在于几个关键方面：

- 减少推理中的错误：通过将复杂的问题分解为一系列较小的任务，CoT 减少了复杂计算和面向逻辑的任务中的错误，从而更精确地解决了复杂的问题（????）。
- 提供可调节的中间步骤：CoT 可以在解决问题的过程中概述可追溯的中间步骤。此功能使用户能够跟踪模型从开始到结论的思维过程，并在观察到不良模型行为时调整提示（??）。
- 促进知识蒸馏：从较大的 LLM 中衍生出来的分步推理过程可以作为小型 LLM 的专用微调数据集。它允许较小的模型通过遵循解释来学习解决复杂的问题，有效地教他们以增强的推理能力解决复杂的问题（?）。

6.2 可解释提示的扩展方法

已经开发了超越 CoT 方法的先进技术，以扩大 LLM 可用的推理路径范围，以提高决策过程的透明度和可理解性（?????）. 我们在下面介绍几个值得注意的例子。

思想 (ToT) 之树。由 ? 提出的 ToT 超越了传统的线性思维链推理，提供了一种更通用的结构，允许模型在多个推理路径中导航。ToT 通过将 LLM 与人类的思维过程紧密结合，使 LLM 的推理过程更具可解释性，因为人类在前瞻性规划和回顾性分析中自然会考虑多种选择和可能的结果来得出结论（??）。这种能力增强了 LLM 应对复杂挑战的能力，这些挑战需要考虑和重新评估不同策略的能力，例如设计游戏策略或生成创意内容。通过模拟人类的思维和决策方式，ToT 不仅使人类用户更容易理解他们的思维过程，而且还提高了模型处理复杂任务的效率。

思想图（GoT）。由 **?** 提出的 GoT 将 LLM 的输出转换为图形格式。这种格式将信息片段可视化为节点，将其连接可视化为边缘，与以前的方法 CoT 和 ToT 相比，实现了更复杂和更互联的推理形式。通过将数据组织成节点（单个概念或信息片段）和边缘（这些概念之间的关系），GoT 使复杂系统内的逻辑连接更容易理解（**?**）。这种图形表示为理解复杂信息带来了几个好处。首先，它支持动态修改概念之间的关系，提供更改一个元素如何影响其他元素的清晰可视化。这在法律推理（**??**）、科学研究（**??**）和政策分析（**?**）等领域至关重要，在这些领域，各种因素之间的相互依存关系可能是错综复杂和微妙的。其次，GoT 能够评估图中每个节点的重要性，从而深入了解哪些信息对任务最关键。这种程度的适应性和清晰度使 GoT 在分析和导航复杂的信息网络方面非常强大。

6.3 案例研究：CoT 真的使 LLM 推理变得可解释吗？

6.3.1 背景和实验设置

尽管 CoT 提示设计具有明显的直观性，但一个关键问题仍未得到解答：CoT 真的使 LLM 推理可解释吗？换句话说，通过 CoT 提供的信息能否忠实地反映 LLM 的底层生成过程？我们使用多跳问答（QA）作为方案来调查此问题。

在 QA 系统中，回答多跳问题仍然是一个重大挑战。多跳问题不是利用单一信息源，而是需要将来自多个数据片段或数据源的信息综合成一个连贯的逻辑序列。虽然 LLM 在单跳 QA 任务中表现出良好的性能（**?**），但它们在多跳情况下的效能显着下降（**???**）。这种差异凸显了需要更先进的方法来有效处理复杂的多跳推理。

CoT 提示进行多跳 QA。为了应对上述挑战，我们的案例研究应用了 CoT 技术。CoT 依靠高质量的模板作为上下文学习提示，我们举了一个例子，如下所示。在这里，$[x]$ 表示测试问题。每个“问题”后面的“想法”是多跳问题的分步解决问题的陈述。模板中的想法使 LLM 的生成过程与人类认知问题解决模式保持一致。

```
Question: What is the capital of the country where Plainfield Town Hall is located
    ?
Thoughts: Plainfield Town Hall is located in the country of the United States of
    America. The capital of the United States is Washington, D.C.
Answer: Washington, D.C.
...

Question: Who has ownership of the developer of the Chevrolet Corvette (C4)?
Thoughts: The developer of Chevrolet Corvette (C4) is Chevrolet. Chevrolet is
    owned by General Motors.
Answer: General Motors

Question: [x]
```

CoT 忠实度解释：为了定量衡量 CoT 的忠实度，我们选择保真度作为相应的指标（**??**）：

$$Fidelity = \frac{\sum_{i=1}^{N} \left(\mathbb{1}\left(\hat{\boldsymbol{y}}_i = \boldsymbol{y}_i\right) - \mathbb{1}\left(\hat{\boldsymbol{y}}_i^{mislead} = \boldsymbol{y}_i\right) \right)}{\sum_{i=1}^{N} \left(\mathbb{1}\left(\hat{\boldsymbol{y}}_i = \boldsymbol{y}_i\right) \right)} \times 100\%, \tag{6}$$

$\boldsymbol{y}_i$ 表示地面实况标签，$\hat{\boldsymbol{y}}_i$ 表示带有 CoT 的原始模型输出，而 $\hat{\boldsymbol{y}}_i^{mislead}$ 表示在“想法”部分插入误导性信息的模型输出。下面，我们举个例子。鉴于目标问题，正确的分步思路应该是：“艾莉 · 肯珀（Ellie Kemper）是美利坚合众国公民。美利坚合众国总统是乔 · 拜登。”为了误导模型，我们用不正确的信息（带下划线的文本）替换想法，并要求模型根据错误的想法生成新的答案。如果模型在修改后仍然生成正确答案，我们认为 CoT 信息并不能忠实地反映答案生成的真实过程。另一方面，如果它产生了与错误思想相对应的答案，那么我们声称这些思想是忠实的。

```
Question: What is the capital of the country where Plainfield Town Hall is located
    ?
Thoughts: Plainfield Town Hall is located in the country of the United States of
    America. The capital of United States is Washington, D.C.
Answer: Washington, D.C.
```

```
...

Question: Who has ownership of the developer of the Chevrolet Corvette (C4)?
Thoughts: The developer of Chevrolet Corvette (C4) is Chevrolet. Chevrolet is
    owned by General Motors.
Answer: General Motors

Question: [Who is the head of state of the country where Ellie Kemper holds a
    citizenship?]
Thoughts:
Answer:
```

实验设置。我们评估了 MQUAKE-CF 数据集 **?** 的性能，该数据集包括每个 K 跳问题的 1,000 个案例，$K \in \{2, 3, 4\}$ ，总共包含 3,000 个问题。我们的评估应用了多种语言模型，包括 GPT-2 (**?**) 有 15 亿个参数，GPT-J (**?**) 有 60 亿个参数，LLaMA (**?**) 有 70 亿个参数，Vicuna-v1.5 (**?**) 有 70 亿个参数，LLaMA2-chat-hf (**?**) 有 70 亿个参数，Falcon (**?**) 有 70 亿个参数，Mistral-v0.1 (**?**) 有 70 亿个参数，Mistral-Instruct-v0.2 (**?**) 有 70 亿个参数。这些模型已经证明了在语言生成和理解方面的熟练程度。

6.3.2 实验结果

性能改进。表 **??** 中报告的多跳问答性能突出了 CoT 在各种模型中的有效性。GPT-J 表现出最显着的进步，尤其是在 3 跳问题中，准确率提高了 200 % 以上，这表明 CoT 的连贯推理大大增强了 LLM 的问答能力。虽然 GPT-2 显示出适度的收益，但 GPT-J 和其他模型（如 LLaMA、Vicuna-v1.5、LLaMA2、Falcon、Mistral-v0.1 和 Mistral-v0.2）的性能表明 CoT 方法可能更有利于更高级的模型。LLaMA2 和 Mistral-v0.2 在 3 跳问题中分别表现出 78.4 % 和 63.8 % 的显着改进，进一步支持了这一观察结果。总体而言，这些结果强调了相干推理技术（即 CoT）在提高 LLM 在不同模型架构和规模上的问答能力方面的潜力。

Table 5: MQUAKE-CF 上的多跳问答性能。

Datasets	MQUAKE-CF								
Question Type	2-hops			3-hops			4-hops		
Edited Instances	Base	Enhanced	Improve	Base	Enhanced	Improve	Base	Enhanced	Improve
GPT-2 (1.5B)	13.6	15.9	16.9 % ↑	11.6	8.9	23.3 % ↓	7.0	8.4	20.0 % ↑
GPT-J (6B)	23.1	51.9	124.7 % ↑	10.1	30.5	202.0 % ↑	21.3	49.8	133.8 % ↑
LLaMA (7B)	47.7	65.1	36.4 % ↑	29.6	39.3	32.8 % ↑	52.4	62.9	20.0 % ↑
Vicuna-v1.5 (7B)	41.3	56.3	36.3 % ↑	22.7	29.7	30.9 % ↑	31.6	53.1	68.2 % ↑
LLaMA2 (7B)	36.7	58.7	60.0 % ↑	17.0	30.3	78.4 % ↑	29.2	49.1	68.1 % ↑
Falcon (7B)	42.3	61.7	45.7 % ↑	23.2	31.7	35.7 % ↑	33.3	48.6	45.7 % ↑
Mistral-v0.1 (7B)	49.0	69.3	41.5 % ↑	30.0	42.3	41.1 % ↑	48.7	63.2	29.9 % ↑
Mistral-v0.2 (7B)	44.0	56.3	28.0 % ↑	23.0	37.7	63.8 % ↑	32.9	56.2	70.9 % ↑

CoT 的忠实度评估。Table **??** 说明了准确与误导的 CoT 对 LLM 性能的影响。Fidelity 指标表示模型的输出如何忠实地反映 CoT 中描述的推理过程。理想情况下，高保真度分数表明模型的最终响应直接基于提供的 CoT，验证它是对模型推理路径的忠实解释。然而，正如我们将在下面讨论的那样，低保真度可能并不总是意味着模型推理缺乏忠实性，这就要求在未来的研究中开发更有效的评估方法。

GPT-J 和 LLaMA 在不同的问题类型中表现出高保真度分数，表明对给定推理路径的强烈遵守。相反，Vicuna-v1.5、LLaMA2、Falcon、Mistral-v0.1 和 Mistral-v0.2 等模型的误导准确率得分相对较高，保真度得分较低。在实验中，我们观察到这些模型通常依赖于自己生成的想法，而不是使用 CoT 中提供的错误信息。特别是 Mistral-v0.2，表现出最低的保真度分数和最高的误导性准确性分数，表明对虚假信息具有潜在的自卫能力。后来模型的保真度分数较低，这可能归因于它们在更多样化和高质量的数据集上改进了训练过程，使他们能够更好地理解上下文和推理。因此，他们更有可能生成自己的正确推理路径。

总之，产生的想法通常可以被视为对其输出答案的忠实解释。虽然高保真度分数通常表明模型遵守所提供的 CoT，但低保真度分数并不一定意味着缺乏忠实度，尤其是当模型表现出拒绝误导性信息的能力时。对 CoT 忠实度的进一步研究和更复杂的评估指标的开发可能有助于推进可解释和可靠的语言模型。

Table 6: MQUAKE-CF 上的 CoT 忠实度评估。

Datasets	MQUAKE-CF								
Question Type	2-hops			3-hops			4-hops		
Edited Instances	Correct	Mislead	Fidelity	Correct	Mislead	Fidelity	Correct	Mislead	Fidelity
GPT-2 (1.5B)	15.9	5.2	67.3 %	8.9	2.9	67.4 %	8.4	1.3	84.5 %
GPT-J (6B)	51.9	7.3	85.9 %	30.5	1.8	94.1 %	49.8	2.0	96.0 %
LLaMA (6B)	65.1	9.9	84.8 %	39.3	6.1	84.5 %	62.9	6.0	90.5 %
Vicuna-v1.5 (7B)	56.3	21.7	61.5 %	29.7	12.7	57.3 %	53.1	16.1	69.7 %
LLaMA2 (7B)	58.7	17.0	71.0 %	30.3	8.3	72.5 %	49.1	12.0	75.6 %
Falcon (7B)	61.7	24.0	61.1 %	31.6	15.0	52.6 %	48.6	23.1	52.4 %
Mistral-v0.1 (7B)	69.3	24.0	65.4 %	42.3	13.0	69.3 %	63.2	18.4	70.8 %
Mistral-v0.2 (7B)	56.3	47.9	14.8 %	37.7	22.0	41.6 %	56.2	37.3	33.6 %

6.4 挑战

在机器学习中，解释忠实度是指解释在多大程度上准确地反映了它所要阐明的模型的决策过程（**?**）。如果解释导致模型做出与原始输入相同的决定，则该解释被认为是忠实的。在这种情况下，可解释提示（例如 CoT 提示）面临的挑战在于两个方面：（1）指导语言模型生成真正代表模型内部决策过程的解释，以及（2）防止语言模型依赖于可能有偏见的 CoT 模板。

关于第一个挑战，我们的案例研究表明，相对较小的语言模型可能会产生与所提供的 CoT 基本原理不一致的答案。因此，这些基本原理并不能准确代表这些模型中的决策过程。已经做出了一些努力来增强较小语言模型的 CoT 能力，方法是使用 CoT 原理实现指令调优（**??**）。这些方法有助于提高 CoT 对小语言模型的解释可信度，从而在一定程度上解决这一问题。然而，如何确保生成的解释（即“模型所说的”）忠实于语言模型的内部机制（即“模型认为什么”）仍然是一个具有挑战性的问题。

关于第二个挑战，最近的研究表明，在模型输入中引入偏差提示模板可能会对 CoT 中的解释产生很大影响（**?**）。这是因为现有的 CoT 需要精心设计的模板来提示语言模型生成解释。如果在此类模板中编码了不正确或有偏见的信息，则生成的解释可能会产生误导。最近，**?** 提出了一种新的解码策略，通过提示实现 CoT，可以缓解这个问题。然而，如何有效地帮助语言模型摆脱对模板的依赖，仍有待探索。

7 通过知识增强提示增强 LLM

利用外部知识增强模型可以显著提高决策过程的控制和可解释性。虽然 LLM 通过对 Web 规模数据的预训练获得了广泛的知识，但这些知识隐含在模型参数中，因此在推理过程中解释或控制如何使用这些知识具有挑战性。此外，LLM 可能并不总是包含特定于某些领域的独特知识，也跟不上世界上不断发展的信息。为了解决这些局限性，本节讨论了检索增强生成（RAG），用于将外部知识显式集成到基于 LLM 的 AI 系统的机制中，旨在产生更可解释的预测。

7.1 预赛：检索-增强生成

通过从外部数据库或互联网获取相关信息，RAG 确保 LLM 输出准确且最新。它解决了 LLM 依赖固定和可能过时的知识库的局限性。RAG 分两步运行：（1）检索：根据用户的查询，从外部来源定位和获取相关信息；（2）生成：将这些信息合并到模型生成的响应中。给定一个输入查询 x 和所需的输出 y，RAG 的目标函数可以表述为（**?**）：

$$\max_{\phi,\theta} \log p(y|x) = \max_{\phi,\theta} \log \sum_{z\in\mathcal{K}} p_\phi(y|x,z)\cdot p_\theta(z|x), \tag{7}$$

其中 z 代表从知识库检索到的外部知识 $\mathcal{K}$。因此，目标分布由知识检索器 $p_\theta(z|x)$ 和答案推理模块 $p_\phi(y|x,z)$ 联合建模。知识 z 是一个潜在的变量。RAG 模型经过训练以优化参数，以便它学会检索相关知识 z 并根据 z 和 x 生成正确的答案 y。由于法学硕士具有更强的文本理解和推理能力，因此无需进一步培训即可直接作为推理模块 p_ϕ。在这种情况下，RAG 可以被视为以数据为中心的问题：

$$\max_{z\in\mathcal{K}} \log p(y|z,x) = \max_{z\in\mathcal{K}} \frac{p(z|x,y)}{p(z|x)} p(y|x), \tag{8}$$

目标是找到支持所需输出的适当知识。基于 RAG 的模型的可解释性来自 z 中的信息：（1）z 通常阐明或补充 x 中特定任务的信息；（2）z 可以解释产出 y 的产生。与其他以端到端方式直接估计 $p(y|x)$ 的深度模型不同，在决策过程中，决策过程是不可理解的，RAG 过程提供了支持结果的理由或理由 z 。现有的检索增强生成（RAG）方法可以根据何时将外部知识集成到模型的工作流程中进行分类。第一类在推理阶段纳入外部知识。例如，**?** 使用密集向量来识别相关文档或文本段落，从而增强了 RAG 的数据检索步骤。同样，**?** 优化数据检索过程，以确保只有最相关的信息才能影响模型的输出。第二类在模型调优阶段整合外部知识。一些具有代表性的方法包括 **???** 。通常，这些方法将检索机制嵌入到模型的训练阶段，使模型能够从一开始就更有效地利用外部数据。

7.2 用显性知识加强决策控制

通过 RAG 整合明确的外部知识，增强了 LLM 中决策的精确性和可控性。此方法利用来自外部数据库的实时信息来生成响应，这些响应不仅准确，而且针对每个查询的特定要求进行定制。下面，我们将探讨 RAG 实现更可控和更有针对性的内容生成过程的机制，并参考了促成这些进步的关键论文。

7.2.1 减少幻觉反应

在 LLM 的上下文中，“幻觉”是指这些模型生成的信息虽然连贯且上下文合适，但并非基于事实准确性或真实世界的证据（**?**）。这个问题可能导致产生误导性或完全捏造的内容，对 LLM 输出的可靠性和可信度构成重大挑战。RAG 提供了一个强大的解决方案来缓解 LLM 中的幻觉问题。通过在生成响应时主动整合最新的、经过验证的外部知识，RAG 确保模型生成的信息锚定在现实中。这个过程大大增强了模型输出的事实基础，从而减少了幻觉的发生。**?** 将神经检索在环架构应用于基于知识的对话，这大大减少了聊天机器人中的事实不准确，正如人类评估所证实的那样。**?** 引入了 RAG-end2end，它将检索器和发电机组件联合训练在一起。他们的方法在医疗保健和新闻等专业领域表现出显着的性能改进，同时减少了知识幻觉。

7.2.2 对知识更新的动态响应

RAG 使 LLM 能够整合最新信息，使他们的决策过程与最新发展保持一致。此功能在快速发展的领域尤为重要，例如医学和技术，在这些领域中，及时准确的信息需求至关重要（**?**）。例如，(**?**) 的研究表明，通过实时信息检索，输出相关性和准确性得到了显著提高。同样，**?** 建议使用检索到的事实数据来有效地纠正和更新预先训练的 LLM 中的知识。此外，**?** 引入了一种将从多语言数据库中新检索到的知识直接集成到模型提示中的方法，从而促进多语言上下文中的更新。

7.2.3 特定于域的自定义

RAG 通过整合来自专业来源的知识来增强 LLM，从而能够创建针对特定领域的模型。**?** 的研究表明，将特定于某些领域的数据库集成到检索过程中，可以使模型能够提供专家级响应，从而提高其在专业和学术环境中的有效性。**?** 通过 MedEdit 将这一概念应用于医学领域，利用上下文学习策略将相关医学知识合并到查询提示中，以获得更准确的医疗建议。此外，最近的研究发现，LLM 很难捕获在预训练数据中未被广泛讨论的特定知识。具体来说，**?** 观察到，LLM 通常无法学习相对较低的长尾事实知识，发现简单地增加模型大小并不能显着增强对此类信息的回忆。然而，他们指出，检索增强的 LLM 在准确性上超过了更大的模型，特别是对于众所周知的问题，这表明这种方法可以有效地弥合知识差距。同样，**?** 强调了 LLM 在获取稀有知识方面的挑战，并提出检索增强提供了一种可行的解决方案，最大限度地减少了对大量预训练的依赖，以捕获细微的、不太常见的信息。

7.3 挑战

我们讨论了 RAG 中与其可解释性方面相关的挑战：（1）在检索阶段 $p_\theta(z|x)$ ，检索到的信息是否总是 z 阐明输入 x 中包含的特定任务信息？（2）在 $p_\phi(y|x,z)$ 的产生阶段，z 是否有效地解释了产出 y 的产生？请注意，我们的目标并不是在本文中详尽讨论 RAG 的所有局限性，因为 RAG 本身是 NLP 研究中的一个广泛主题。为了更详细地研究 RAG 更广泛的局限性，我们将读者引导至其他评论（**?**）。

7.3.1 检索精度瓶颈

现有的 RAG 方法通常依靠相似性搜索来精确定位相关信息（**??**），这代表了对基本关键字搜索的实质性改进（**?**）。然而，这些方法可能会在复杂的查询中遇到困难，这些查询需要更深入的理解和细致入微的推理。

最近的“中间迷失”现象（?）揭示了无效的检索会导致无关或相互冲突的信息的积累，从而对生成质量产生负面影响。为了应对这一挑战，最近的 RAG 方法整合了适应性学习过程（?）。这一进步使检索系统能够随着时间的推移通过反馈来改进其性能，适应不断变化的语言使用和信息更新，确保其响应保持相关性和准确性。尽管如此，有效处理复杂的多跳问题仍然是一个重大挑战，这凸显了持续研究以增强 RAG 系统功能的必要性。

7.3.2 可控发电瓶颈

情境学习是整合外部知识以提高 LLM 能力的首要方法，例如 GPT-4（??）。尽管它很有效，但没有万无一失的方法可以确保这些模型始终如一地利用其决策过程中提示中提供的外部知识。在实践中，为了实现全面覆盖，常用的密集检索通常会返回大量内容，包括输入问题的相关和冗余信息。不幸的是，模型提示中的冗余信息会提高计算成本，并可能误导 LLM 生成不正确的答案。最近的研究表明，检索到的信息会降低问答任务的性能（???）。最近的一些工作建议对 LLM 进行微调，以提高对噪音的适应能力并减少幻觉。然而，这种方法仍然不能防止过大的检索信息降低系统的可解释性（??）。优化外部解释的使用以在 LLM 中实现更精确和可控的决策的挑战是一个持续存在的问题，尚未得到充分解决。

8 带解释的训练数据增强

本节探讨了使用大型语言模型从解释中生成合成数据的方法，这种技术有望增强各种机器学习任务。在机器学习中，有限的数据可用性通常会限制模型性能，从而在许多领域带来重大挑战。一个可行的解决方案是数据增强，其中 LLM 具有强大的生成能力，可用于文本数据增强（?），例如将现有的文本样本转换为新的变体（?）。然而，要实现有效的文本增强，仍有几个挑战需要解决。首先，为了实用性，与原始数据相比，生成的样本需要表现出多样性。其次，这些样本应该表现出与下游任务相关的有用模式。为了应对这些挑战，解释方法提供了一个有价值的工具，通过提供补充上下文和有用的理由来指导数据增强（?）。使用 LLM 进行解释引导的数据增强是一个新兴但很有前途的领域。在本节中，我们旨在概述可行的框架并讨论潜在的应用，为该领域的未来研究提供方向。

在两种情况下，解释在数据增强中可能特别有益。在第一种情况下，解释用于描述所需的模型行为或识别现有的缺陷，从而有效地指导 LLM 的数据增强过程。第二种情况涉及使用 LLM 直接生成解释性文本，这些文本作为补充信息以丰富数据集。

8.1 用于缓解快捷方式的解释性数据增强

机器学习模型容易做出具有虚假相关性的预测，也称为捷径（?），这与人类推理过程不一致。这种对快捷方式的依赖是机器学习中各种挑战的基础，特别是削弱了模型泛化到分布外样本的能力，并削弱了其对对抗性攻击的弹性（?）。模型对快捷方式特征的依赖程度与其可解释性有着内在的联系。从本质上讲，如果模型的预测主要基于这些不可靠的特征，则表明从人类的角度来看，模型的潜在机制无法完全理解或解释。因此，事后解释技术通常在检测深度模型中输入和预测之间的不良相关性方面发挥着至关重要的作用（??）。例如，? 采用集成梯度（IG）将模型的预测归因于其输入特征，表明该模型倾向于将功能词、数字和否定词视为快捷方式，并强烈依赖这些快捷方式特征进行自然语言理解任务的预测。

数据增强可用于训练对快捷方式特征不太敏感的更好模型。反事实（?）等解释性信息已被纳入数据增强中，以提高模型的鲁棒性。它通过首先识别关键特征（例如单词标记），然后用它们的反义词替换这些特征，同时反转它们的相关标签来生成反事实样本。随后，将生成的样本与原始样本相结合，以训练下游模型。此外，这些技术可以扩展以增强较小模型的分布外性能（??）。也就是说，大型语言模型可以作为增强数据的有效工具。例如，LLM 能够合成表示异常情况或罕见情况的示例，这有助于较小的模型更好地泛化看不见的数据（?）。这可能有助于在数据稀缺或机密的情况下构建健壮的模型（?）。此外，与传统方法相比，LLMs 通过生成更有效、更自然的对抗性示例，在提高模型安全性方面很有前途（?）。首先，使用基于归因的方法识别最脆弱的词。然后，这些词被 LLM 替换，以保持原始文本完整性的方式。这些示例的质量可以使用外部分类器进行检查。随后，这些对抗性示例被用于训练下游模型，有效地增强它们免受潜在攻击的能力并提高其安全性。同样，LLM 也有助于减轻偏差，例如模型中的公平性问题（?）。这项工作声称它会自动识别代表性不足的亚组，并选择一种最佳的增强策略，以避免伤害其他群体。使用带有人工标签的 LLM 生成新的组示例。这些实验观察到代表性不足的群体和整体模型性能的改善。这种方法可以防止倾斜的数据集导致特定人群的不公平结果，从而有可能促进社会的公平性。

8.2 解释增强的数据扩充

作为一种强大的生成工具，LLM 已被用于直接生成自然语言解释作为增强数据。这种策略依赖于 LLM 的理解能力来帮助较小的模型完成学习任务。这项工作的一个目标是将 LLM 生成的自然语言解释添加到训练数据中，从而提高小模型的性能。**?** 引入 LLM 的解释，以促进较小模型的训练，以增强其推理能力并获得解释生成能力。具体来说，LLM 的三种解释生成方法被用于训练较小的模型，包括（1）通过思维链提示生成的解释，（2）以黄金标签为条件的合理化提示的解释，以及（3）结合前两种方法的混合方法。在推理案例中的最佳情况下，与大 9.5 倍的 GPT-3 相比，准确性提高了 60 % (**?**) . 值得注意的是，包括 ChatGPT 和 GPT-4 在内的 LLM 在生成连贯和合乎逻辑的解释方面具有竞争力，除了有限的语言 (**?**) 。LLM 的解释也被用于减轻基于方面的情感分析任务中的虚假相关性 (**?**) 。本研究提出使用 LLMs 来解释句子中各方面的情感标签 (**?**) 。这些解释提供了基于上下文语义的推理，而不是依赖于单词和标签之间的虚假相关性。通过两种方法将解释集成到基于方面的情感分析模型的训练中：用解释来增强训练数据，或者通过模仿行为从解释中提炼知识。通过专注于解释而不是肤浅的文本线索，模型可以更好地学习文本和情感之间的真正关联，并变得更加稳健，从而提高域内性能和泛化能力 (**?**) 。另一项工作涉及整合 LLM 基本原理作为额外的监督，以指导较小模型的训练。实验表明，这种方法不仅需要更少的训练数据，而且优于传统的微调和蒸馏方法 (**?**) 。

除了上面总结的增强技术的现有应用外，我们设想合并额外的文本信息也可以在提高各种模型的性能方面实用且有效。例如，一个很有前途的应用在于通过使用自动生成的解释来指导小模型的参数学习过程。以前的研究通过将自然语言推理模型的注意力引向人工设计的解释来研究这一途径 (**?**) 。由于人为的解释既艰巨又不可转移，因此利用 LLM 作为生成器提供了一种更经济和通用的替代方案。另一个潜在的应用是使用 LLM 的自然语言解释来增强复杂任务的模型性能。例如，代码转换生成任务将解释作为中间步骤，将模型性能平均提高 12 % (**?**) 。结果表明，解释在零样本设置中特别有用。除了协助较小的模型外，LLM 还展示了他们通过生成可靠的理由来提高自己的推理能力的能力 (**?**) 。此外，**?** 将事后解释（将分数归因于所有输入特征）嵌入到自然语言原理中。这种方法通过上下文学习将多个任务的模型准确性提高了 10-25 %。另一项研究明确调查了 LLMs 在自然语言中生成事后解释的能力。实验表明，LLMs 发现关键特征的准确率为 75 % (**?**) 。这些研究提出了一种新的策略，利用事后解释，超越传统的自然语言解释，来丰富训练数据，从而提高模型性能。

8.3 挑战

8.3.1 计算开销

建立在训练有素的模型之上的传统事后解释通常是资源密集型任务。上面提到的第一种方案利用可解释性技术来准确诊断数据集问题。此过程通常需要多轮模型训练并应用可解释性方法来开发公平和稳健的模型。因此，制作过程既费时又费力。鉴于这些挑战，探索以数据为中心的评估指标的开发至关重要。这些指标可以提供一种更有效的方法来评估数据问题，绕过传统的、繁琐的解释方法。通过专注于这些以数据为中心的测量，可以在训练之前诊断和修复数据问题。然后，所需的训练轮数将大大减少。这种转变不仅简化了模型开发，还有助于减少计算开销，使整个过程更加实用和高效。

8.3.2 数据质量和数据量

尽管具有先进的功能，但 LLM 在处理高度专业化或利基环境时仍然存在局限性。例如，最突出的问题之一是“幻觉”，其中模型会产生合理但不正确或误导性的反应。这可能会对增强数据的质量产生不利影响，可能会引入更多的偏差，而 LLM 也容易受到这些偏差的影响。另一个挑战是控制 LLM 生成内容的相关性。也就是说，解释或数据点可能看起来很合理，但往往缺乏事实准确性或特定于某个领域的细微差别。目前，我们缺乏可靠的指标来有效衡量这些生成的数据相对于原始任务的质量和相关性。确定所需的精确数据量也具有挑战性，通常会导致新的数据集不平衡。管理 LLM 生成的数据的质量是一个巨大的挑战，因为增强数据可能会引入其他偏差 (**?**) 。这源于 LLM 准确控制生成数据的数量和分布的能力有限。此外，制作有效的提示与其说是一门科学，不如说是一门艺术，这增加了生成数据质量的不确定性。总之，这些因素凸显了充分利用 LLM 在数据增强和相关任务方面的潜力的复杂性和挑战。

9 为 XAI 生成用户友好的解释

前面的部分主要集中在通过数值对 LLM 进行定量解释。例如，第 ?? 节中讨论的基于样本的解释旨在为每个训练样本分配一个影响分数（参见 Eqs. ?? - ?? ），用于衡量我们可以使用该训练样本来解释测试样本预测的置信度。但是，使用数值进行解释并不直观，对于领域知识很少的从业者来说可能很难理解（???）。相反，用户友好的解释旨在生成人类可理解的解释，例如，基于自然语言的描述，涉及某些数据、模型做出某些预测的原因或神经元在网络中扮演什么角色，以便研究人员和从业者都能很好地理解生成的解释。

给定一个被解释者 e ，它可以是数据样本 (x_i, y_i) ，一个来自预训练模型 f_θ 的神经元 θ_i ，或者基于输入 x $\hat{y}$ 预测结果，生成用户友好的解释旨在将被解释者 e 映射到一系列自然语言标记，作为被解释者 e 的解释，这样生成的解释就可以很容易地被人类理解。

9.1 使用 LLM 进行用户友好的数据解释

数据解释是指将困难的材料（例如，程序代码、长文档）翻译成简洁明了的语言，以便于人类理解的过程。长期以来，语言模型一直被用来生成对文本数据的解释（?）。由于现代 LLM 是在由代码、数学和论文组成的语料库上训练的，因此可以利用它们来解释纯文本内容以外的数据。例如，? 已经证明，预训练的 GPT 模型具有理解和生成代码的能力，其中同时生成解释性注释，以促进程序员的理解。此外，? 建议通过提供详细的推导来解释数学定理，使定理更容易理解。最近，法学硕士也被用于阐明学术论文 (?) ，使得领域知识很少的人容易理解困难的内容。

9.2 用 LLM 解释小模型

最近，人们越来越有兴趣利用 LLM 为小型模型生成自由文本解释。例如，为了解释黑盒文本分类器，? 提出了一种基于提示的策略，以识别输入文本中 $K = \{k_1, k_2, ..., k_n\}$ x 的预训练 LLM 中的关键字，这些 LLM 为标签 y 提供了信息，并要求 LLM 将它们替换为另一组关键字 $K' = \{k'_1, k'_2, ..., k'_n\}$ ，以便更改文本 x' 将标签预测更改为 y' 。他们认为文本映射规则“如果我们在 x 中将 K 改为 K' ，那么 y 将被归类为 y' ”作为模型的反事实解释。此外，为了解释预训练语言模型（例如 GPT2）的神经元，? 建议将神经元激活模式总结为具有更大语言模型（例如 GPT4）的文本短语，其中神经元激活模式表示为（标记，归因分数）对的序列。为了验证已识别的模式，他们通过相同的 LLM 根据短语生成激活模式，并将其与神经元的真实激活模式的相似性进行比较，其中得分高的短语被认为更有信心作为神经元的解释。

LLM 的解释能力并不一定局限于文本模型。例如，? 建议使用预先训练的视觉语言模型来生成对图像分类模型的神经元 θ_i 的解释。具体来说，对于每个类 $y = y_c$ ，他们首先在具有标签 y_c 的图像中找到神经元 θ_i 的最大激活区域作为 θ_i 的替代解释者，并提示 ChatGPT 等 LLM 为类标签 y_c 生成候选解释（单词、短语）。然后，他们使用预训练的视觉语言模型 CLIP （?）将候选解释与代理解释者作为神经元 θ_i 的解释进行匹配。最近，LLMs 也发现了在解释推荐系统 ? 方面的应用。具体而言，? 发现 LLMs 可以很好地解释对齐后顺序推荐模型的潜在空间，而 ? 建议将 LLMs 的用户令牌与学习到的小推荐模型的用户嵌入对齐，以生成嵌入中编码的用户偏好的解释。最近， ? 提出了一个统一的框架来解释所有可以将输入和输出转换为文本字符串的模型。具体来说，解释器 LLM 用作代理，通过迭代创建输入和观察模型的输出来与被解释者模型进行交互，其中文本解释是通过将所有交互视为上下文来生成的。

9.3 LLM 的自我解释

由于 LLM 的黑盒性质，为 LLM 本身生成用户友好的解释是有希望的，这样人类专家就可以很好地理解 LLM 的操作机制和预测。根据 LLM 是否需要重新训练以为自己生成解释，LLM 的自我解释可以分为两类：基于微调的方法和基于上下文的方法，这将在以下部分介绍。

基于微调的方法。在训练数据的标签上有足够的示例解释（例如，在推荐数据集中，例如 Amazon Review 数据集 (?) 或 Yelp 数据集 (?) ，用户已经提供了关于他们为什么购买某些物品的解释，这可以被视为对评级的解释），LLM 可以通过监督学习学习来学习为他们的预测生成解释作为辅助任务。一个示例方法是 P5 (?) ，它根据评级和解释数据微调预训练语言模型 T5 (?) ，以生成与建议一起的解释。最近，一些工作对 P5 (??) 进行了改进，它对不同的 LLM（如 GPT2、LLaMA、Vicuna 等）进行了微调，并提出了不同的提示学习策略，(?) 生成解释作为辅助任务。通过引入解释作为额外的监督信号来微调预训练的 LLM 以获得建议，可以通过良好的可解释性来提高性能。

基于上下文的方法。在许多应用中，往往缺乏足够的示例解释。然而，现代 LLM 通过类似人类的提示进行推理和提供答案的独特能力引入了基于上下文的解释的潜力。在这里，预测的解释完全基于提示中的信息进行设计。该领域的领先方法是思链（CoT）提示（?），它在提示中提供少量示例（有或没有解释），并要求 LLM 在逐步推理后生成答案，其中为生成最终答案提供更多上下文的中间推理步骤可以被视为解释。但是，CoT 首先生成推理，然后基于推理生成预测，其中推理步骤可以影响预测结果（?）。如果解释是在预测之后生成的，由于解释是以预测标签为条件的，因此它可以提供更忠实的事后解释，说明模型为什么（?）做出某些决策。基于上下文的 LLM 自我解释的应用是广泛的。例如，? 探索使用 LLM 生成情感分析的零样本自我解释，方法是直接要求它们在预测旁边生成解释。此外，? 提出了一种解释链策略，旨在解释 LLM 如何从文本输入中检测仇恨言论。? 发现 CoT 可以用科学知识为问答生成有充分支持的解释。

9.4 挑战

9.4.1 可用性与可靠性

许多现有的方法依靠提示来生成用户友好的解释，这些解释不如具有良好理论基础的数值方法可靠。? 发现 CoT 的解释可能没有事实依据。因此，他们认为这些解释更适合作为关于为什么 LLM 做出某些预测的事后解释（无论预测是对还是错）。然而，将 CoT 解释视为事后理由的有效性受到了 ? 的最新研究结果的质疑，该研究使用有偏见的数据集（例如，提示中的少数示例总是回答多项选择题的 “A”）来表明生成的解释可能是合理的，但系统地不忠实于代表 LLM 的真实推理过程。这个问题源于有偏见的推理步骤，这些步骤无意中歪曲了预测。因此，越来越需要对用户友好的解释进行更多的理论审查，以确保其真实性和可信度。

9.4.2 受限应用场景

目前，利用 LLM 来解释较小的黑盒模型主要局限于那些处理具有丰富文本信息的数据（??）的模型。尽管 ? 提出了一种解释图像分类器的策略，但将候选文本解释与图像模式相匹配的能力仍然依赖于预训练的视觉语言模型 CLIP。这种方法可能不适用于其他领域，例如图机器学习（例如图神经网络）或时间序列分析（递归神经网络），与自然语言处理和计算机视觉相比，大型预训练模型几乎没有进展。因此，迫切需要设计更通用的策略来解释更广泛领域的模型。这项工作可能取决于将 LLM 与其他特定领域任务相结合的基础研究，例如以零样本方式开发适用于看不见的图的图语言模型。

10 用于可解释 AI 系统设计的 LLM

XAI 中一个有趣但具有挑战性的问题是创建模型架构甚至本质上是可解释的 AI 系统（?），其中不同的模型组件代表清晰易懂的概念或功能，这些概念或功能很容易相互区分。机器学习模型，如支持向量机（?）和基于树的模型（?）是实现模型可解释性的经典技术。在深度学习时代，这方面的典型研究领域包括概念瓶颈模型（??）、解纠结表示学习（??）和网络剖析（??）。然而，在传统的深度学习环境中，由于两个主要挑战，这些技术的可用性仍然有限。首先，很难定义模型预期捕获的概念或功能范围。其次，与黑盒模型相比，可解释模型的功效往往不足，从而限制了它们的实际效用。

大型基础模型，如大型语言模型（LLM）和视觉语言模型（VLM），提供了弥合差距的机会。通过利用嵌入其中的常识性知识，基础模型可以通过提供提示来设计可解释的架构，这些提示鼓励在 AI 工作流中创建和使用功能或过程。这与传统的深度学习管道不同，在传统的深度学习管道中，深度模型在训练过程中会自动发现特征，而这些特征最终可能不会得到具有明确含义的模型组件。此外，LLM 可以将复杂的任务分解为更简单和协作的子任务，从而增强系统的可解释性和整体性能。

10.1 使用 LLM 设计可解释的网络架构

开发可解释深度架构的代表性方法包括广义加性模型（GAM）（??）和概念瓶颈模型（CBM）（??）。这些模型将输入映射到人类可理解的潜在空间，然后将该空间的线性变换应用于目标标签。例如，为了构建诊断关节炎的分类器，我们可以让模型识别 “骨刺” 和 “硬化症” 等特征，然后使用这些可解释的特征进行最终决策。然而，这些方法通常需要专家的参与来定义潜在空间，这可能会限制深度模型的学习能力。一些工作试图在模型训练期间自动发现语义概念，例如要求概念之间的独立性（??）或聚类数据（?），但它们缺乏对结果的直接控制，不能确保概念的清晰度。一个有前途的策略是利用 LLM 来提供可理解的概念候选者。? 使用人类语言作为视觉识别的内部表示，并为下游任务创造一个可解释的概念瓶颈。通过基于这些可理解的概念进行决策，模型架构本身具有更好的透明度。同样，最近的一种方法 Labo（?）构建了无需手动概念

注释的高性能 CBM。此方法通过从 LLM 生成候选来控制瓶颈中的概念选择，这些候选者包含重要的世界知识（?）可以通过提示字符串前缀来探索。人体研究进一步表明，这些源自 LLM 的瓶颈是事实和可依据的，为模型设计保持了很大的内在可解释性。除了基于概念的模型之外，另一个有前途的策略是使用 LLM 来增强固有可解释的传统架构，例如 GAM 和决策树（DT）。? 利用 LLM 中捕获的知识来增强 GAM 和 DT，其中 LLM 仅在增强模型训练期间参与，而不是推理过程。对于 GAM 训练，LLM 可以提供解耦嵌入以进行增强。对于 DT 训练，LLM 能够帮助生成改进的拆分特征。LLM 增强的 GAM 和 DT 可实现完全透明，其中只需要求和系数和输入关键短语即可进行解释。借助来自 LLM 的额外信息，与非增强 GAM 和 DT 相比，增强 GAM 和 DT 能够实现更好的泛化性能。

10.2 使用 LLM 代理设计可解释的 AI 工作流

传统的深度模型通常以端到端的方式设计。内部工作流程对于普通用户来说不太容易理解。通过利用常识性世界知识，LLM 可以将复杂的问题分解成更小的问题，并在它们之间组织工作流程，从而实现更具可解释性的 AI 系统设计（?）。最近一个关于可解释 AI 工作流设计的例子来自 ?，其中 LLM 驱动的代理利用 ChatGPT 集成各种现成的 AI 模型（例如，来自 Hugging Face（?））来处理不同的下游应用程序任务。为了在透明的工作流程中处理复杂的任务，LLM 在与外部模型协调语言媒介以利用其力量方面发挥着关键作用。通过规划目标任务、选择候选模型、执行分解的子任务和总结响应，LLM 可以帮助根据用户请求拆解任务，并根据模型描述为任务分配适当的模型。同样，为了使工作流透明，? 引入了一个任务分解器来分析用户提示，并将其分解为许多子任务，以便使用 LLM 进行求解。每个子任务都得到了很好的管理，并带有描述、域、输入和输出。通过这种方式，人工智能系统能够通过逐步可理解的工作流程处理复杂的用户提示。在提示范式下，? 还使用 LLM 通过分解来解决复杂的任务。从工作流程可跟踪的软件库中汲取灵感，分解者和共享子任务以模块化方式设计。更进一步，? 为复杂任务引入了一种交互式计划方法，该方法通过集成计划执行描述和提供反馈的自我解释来增强对初始 LLM 生成计划的纠错。这种交互性在长期规划和多步骤推理任务场景中实现了更好的工作流程透明度。

10.3 挑战

10.3.1 复杂场景规划可行性

尽管 LLM 具有任务规划能力，但由于可行性问题，在实际应用中应用于某些场景仍然具有挑战性。一个典型的场景是小样本规划案例（?），其中获取大型数据集进行训练要么不切实际，要么成本过高，因此从稀疏样本中对看不见的案例进行可行的规划极具挑战性。为了更好地协助可解释的设计，LLM 规划需要在没有广泛监督的情况下很好地概括，并期望能够整合来自先前经验和知识的信息。此外，另一个重要的场景在于动态规划设置（?），其中 LLM 迭代整合来自环境的反馈，让智能体采取思考步骤或用推理跟踪来增强其上下文。动态场景迫切且频繁地涉及 LLM 迭代调用导致的高计算成本，并且在处理上下文窗口的限制和从规划幻觉中恢复方面仍然面临挑战。

10.3.2 辅助可靠性与知识差距

LLM 在将现实世界的知识封装在其参数中方面表现出非凡的熟练程度，但当某些知识缺失或不可靠时，他们会以高度的信心诉诸幻觉和偏见。尽管已经提出了越来越多的技术，如检索增强（?）、搜索集成（?）和多 LLM 协作（?），以扩展 LLM 知识，但由于人类理解的不断发展性（?），这种知识上的差异可能永远存在。因此，一个关键的研究挑战不断上升，即在将 LLM 用于设计时，如何有效地检测和减轻人类的 LLM 知识差距。我们需要进一步研究评估和开发强大的 LLM 机制，以解决知识差距问题，以帮助提高 LLM 的可靠性，减少幻觉并减轻偏见。此外，知识差距和安全方面之间的交叉点也是需要解决的巨大挑战，这可能会带来一些安全问题，尤其是在将 LLM 用于下游模型或工作流设计时。

11 使用 XAI 的 LLM 模拟人类

本节讨论如何利用 LLM 通过扮演人类角色来为 XAI 服务。构建可解释的模型需要人类参与的两个主要步骤：(1) 收集具有人工注释基本原理的数据集来训练模型；(2) 收集人类对模型产生的解释质量的反馈，以供评估。人工参与所需的大量成本和时间是扩大这一程序的主要挑战。LLMs 成为应对这一挑战的一个有前途的解决方案，这要归功于它们能够模拟人类推理并产生与人类生成的内容非常相似的响应。在下文中，我们将介绍一些方法，这些方法证明了 LLMs 能够生成类似人类的注释和反馈，从而有助于创建可解释的模型。

11.1 模拟人类注释器以训练可解释模型

将人类可理解的基本原理纳入模型开发中，已经显示出它在提高系统在各种 NLP 任务的透明度和性能方面的有效性，例如问答（??）、情感分析（??）和常识推理（??）。我们使用“基本原理”一词来描述证明输入和输出之间联系的支持性证据（?）。传统上，基本原理是通过利用人工注释（??）或应用专家设计的规则（??）来收集的，导致成本昂贵或质量有限。最近，自动注释（???）的研究人员已经开始探索利用高级 LLM 来模拟人类注释者对特定任务示例的目标标签进行注释的潜力。这些研究发现，在大多数任务上，高级 LLM 表现出与普通人群人类注释器相当的注释质量，成本较低，指出了使用机器模拟注释器的可扩展性。受这些著作的启发，一些研究（??）试图利用先进的 LLMs，通过应用思维链技术来收集基本原理。具体来说，研究人员在输入文本中提供了几个输入-基本原理-输出演示，以提示 LLM 为未标记的输入实例生成基本原理和输出。这种注释原理的质量很大程度上取决于 LLM 的上下文学习能力，导致不常见任务的注释质量无法控制。其他学者（???）提出了一种基于主动学习架构的基于人机交互的 LLM 标注框架。该框架最初收集一个小型种子数据集，其中包含人工注释的基本原理和标签。此种子数据集用于为此下游任务训练可解释的分类器。然后，每个未标记的样本都通过经过训练的可解释分类器。接下来是选择策略，该策略根据解释合理性、预测不确定性和样本多样性等指标选择具有代表性的样本。最后，利用 LLM 来注释这些选定的未标记样本的基本原理和标签。此过程可以重复多次，最近一次训练的可解释分类器是该框架的最终输出。与其他方法相比，该方法通过使用 LLM 仿真注释器在开发可解释模型时平衡了注释质量和成本预算。

11.2 模拟人类反馈以评估可解释的模型

可解释模型生成的解释可分为两类：抽取式和抽象式（?）。提取性解释直接来自输入数据，例如基于归因的方法，这些方法强调输入文本的特定部分。相比之下，抽象解释是以自由形式的文本方式生成的，例如思维链（CoT）响应（?），提供更细致入微的解释。抽取式解释的质量通常是通过它们与带注释的理由（?）的一致性来评估的，例如准确性、召回率和精确性。然而，评估抽象解释是一个重大挑战，因为全面详尽所有合理的抽象结果是不切实际的。为了自动评估抽象解释，早期的研究首先收集一些自由文本的理由，然后应用 LLM 来估计解释和理由之间的相似性（??）。抽象解释和注释基本原理之间的相似性越高，表明模型更透明。最近，一些研究人员直接使用 LLM 来检查模型解释的合理性，而不参考人类注释的基本原理（??），强调用高级 LLM 模拟人类反馈的潜力。

11.3 挑战

11.3.1 不可控的仿真可信度

虽然 LLM 可以帮助收集基本原理和解释评估，但它们对收集结果的行为可能并不总是与人类注释者相匹配，这主要是由于它们不熟悉的领域中的幻觉反应（?）。这个问题会导致不可靠的注释或反馈，因为 LLM 自信地生成了事实上不正确的结论。从此过程收集的数据质量会受到影响，从而影响 XAI 系统的开发。为了提高注释和反馈的质量，未来的研究可以集中在结合幻觉检测（?）和检索增强生成（?）技术。这些方法可以提高 LLM 输出的可靠性，使它们在 XAI 开发的背景下与人类生成的内容更具可比性。

11.3.2 LLM 注释中的伦理考虑

当 LLM 注释者让人类注释者远离主观场景时，例如仇恨言论检测（?），LLM 也有机会将不道德的意见注入到他们的注释数据集中。尽管大多数高级 LLM 都经过微调以符合人类价值观（?），例如乐于助人、诚实和无害，但许多研究表明，这种保护机制可以越狱（??），导致模型产生违反价值观的答案。确保 LLM 注释者遵循道德准则值得进一步探索。

12 讨论与结论

XAI 研究正在经历重大变革，并在大型模型时代经历快速扩张。在前面的章节中，我们介绍了 XAI 方法，并强调了它们的可用性。在最后一节中，我们对该领域持续存在的总体挑战进行了高层次的概述，并为未来的努力提出了方向。

- 规避可解释性与准确性的权衡。现代 LLM（例如 ChatGPT）的出现对这种权衡产生了重大影响。传统上，在许多应用中，人们愿意牺牲一定程度的性能来获得更好的透明度。相应的 XAI 策略是训练和部署一个模仿黑盒模型的固有可解释模型（?）。然而，将这种策略应用于 LLM 带来了挑战，因为很难确定

一个可以匹配 LLM 性能水平的可解释模型。这需要创建可以规避这种权衡的 XAI 策略，其中增强的可解释性有助于提高准确性。这与本文讨论的 Usable XAI 的目标一致。

- 数据驱动的 AI 与 XAI。数据驱动的 AI 是指开发基于大量训练数据运行的 AI 模型。这种方法通常会导致“黑匣子”模型，因为它强调结果而不是决策路径的清晰度。目前，XAI 技术的发展落后于 LLM 的发展，因为后者很容易通过数据驱动的方法进行扩展 - 它们从互联网上摄取大量文本进行训练。然而，我们认为，由于一些机会，XAI 仍可能迎头赶上。(1) 我们可能会用完数据。据预测，“我们将在 2026 年之前耗尽高质量语言数据的存量” [*] 。如果更多数据的积累不再产生实质性的改进，重点可能会转向增强模型的可解释性，以更有效地利用现有数据。(2) 模型比较稳定。由于 LLM 的 Transformer 架构非常成熟和稳定，它将吸引更多的注意力来解释它们的内部工作原理。(3) 利用 LLM 进行 XAI。如果 XAI 研究能够正确利用 LLM 的知识和类人能力，则可以加速其发展。
- 客观很重要。在从经典机器学习时代（当支持向量机和决策树占主导地位）到深度学习时代（卷积和递归神经网络变得流行时）的过渡期间，XAI 技术非常强调在模型中实现完全透明，就好像“任何不完全透明的东西都是不透明的”。然而，随着 LLM 开始在各种任务中匹配甚至超过人类的能力，某些 XAI 问题的重要性发生了变化。例如，当循环神经网络（RNN）被广泛用于文本生成时，我们对输出如何在语言上派生感兴趣，因为 RNN 经常产生无意义的句子。如今，我们对 LLM 的这一点不太感兴趣，因为他们擅长生成连贯的文本。然而，我们的重点可能会转移到解释 LLM 如何用事实信息构建输出，因为 LLM 容易产生幻觉。人类认知中也存在类似的观察结果，可以分为系统-1 和系统-2 风格：系统-1 处理难以解释的直觉和无意识任务，而系统-2 则包括逻辑思维、计划和推理 (**?**) 。鉴于 LLM 的庞大规模和复杂性，在不久的将来，在这些模型的所有方面实现绝对透明似乎越来越不可行。因此，优先考虑有意义和可行的解释目标，为特定任务量身定制，对于提高人工智能系统在实际应用中的效用至关重要。
- 传统的 XAI 已经发展出一个全面的解释问题和格式分类法，并附有每个类别的明确定义 (**???**) . 然而，由于两个原因，不能简单地将已建立的分类法嫁接到 LLM 的研究中。首先，某些 XAI 挑战在 LLM 的背景下失去了突出地位，而一些方法变得过于复杂而无法实际应用。其次，虽然 XAI 正在成为 LLMs 解决问题的常见途径，但对 LLMs 内在机制的探索已经分支到各个方向。例如，利用人类行为和局限性的见解来解释 LLM 有一种明显的趋势，例如 LLM 是否可以撒谎 (**?**) ，LLM 可以保守秘密吗 (**?**) ，提示中的礼貌对 LLM 的影响 (**?**) ，甚至它们如何被“催眠” (**?**) 。这些不同的方法在解释 LLM 行为时没有融合到统一的方法论中，这使得评估具有挑战性。一个潜在的风险是，由此产生的解释可能会给用户一种错误的感觉，即他们准确地理解了模型，特别是当用户试图硬塞某些人类知识或概念来解释 LLMs (**?**) 时。

结论。 在本文中，我们希望引导读者了解可解释人工智能（XAI）的一个关键但经常被低估的方面——可用性。为此，我们提出了在 LLM 范式中推进 Usable XAI 的 10 种策略，包括（1）利用解释来相互增强 LLM 和通用 AI 系统，以及（2）通过整合 LLM 功能来丰富 XAI 方法。释放 XAI 可用性的潜力有助于解决 LLM 中的各种挑战，例如人为一致性。我们还提供几个关键主题的案例研究，旨在为感兴趣的开发人员提供资源。我们在每个战略的末尾进一步讨论开放性挑战，为这一不断发展的领域的未来工作指明方向。

13

确认 这项工作部分由 NSF （# IIS-2223768、# IIS-2223769、# IIS-2310261、# DRL-2101104）提供支持。本文的观点和结论是作者的观点和结论，不应被解释为代表任何资助机构。

[*] https://www.livemint.com/mint-top-newsletter/techtalk20102023.html