如何缓解大模型幻觉?

来自： AiGC面试宝典

宁静致远 2023年09月29日 13:36

一、为什么 会 出现 大模型幻觉?

1.1 训练数据中存在 矛盾或者错误的表述 对LLMs训练影响

大模型幻觉出现的主要原因之一 是 **训练数据中存在 矛盾或者错误的表述问题。由于 用于 LLMs 训练的 标注数据 大多来自于互联网网上数据（新闻、文章、书籍、网站等）。虽然这些数据一定程度上提供了有价值的语言模式，但它也不可避免会包含一些不准确的信息（因为互联网上的信息并不是都经过审核的）。**

LLM的训练过程大多数是基于next token prediction的方式进行预训练，因此，它只能保证文本生成的流畅性，而无法辨别所遇到的信息是否真实或准确。

因此，**如果训练数据中包含一些矛盾或者错误的表述，就可能导致LLM也在学习这些错误的表达，从而一定程度导致了幻觉的产生。**

1.2 训练数据中存在 偏见表述 对LLMs训练影响

大模型幻觉出现的主要原因之一 是 **训练数据中存在 偏见表述 对LLMs训练影响。**

训练数据中可能存在社会偏见、文化信仰和个人观点之类相关的语料。这些偏见可能也会被LLM学会，从而导致LLM生成的文本会包含一些错误或有偏见的信息。

1.3 LLMs 学习到 知识 缺乏外部验证

大模型幻觉出现的主要原因之一 是 **LLMs 学习到 知识 缺乏外部验证。**

虽然大模型拥有在训练过程中学习并获得大量内部知识的能力，但是 由于 们缺乏实时访问最新信息或根据外部参考验证事实的能力，使得 他们 无法辨别 产生的信息是准确的还是虚构的。

此外，生成过程中缺乏事实核查机制使得LLM可以生成听起来合理但缺乏任何实质性证据或事实依据的文本。这也告诉我们不能太相信LLM。

LLM幻觉的原因在于训练数据、文本中存在的偏见以及LLM的固有局限性。通过解决这些基本因素，我们可以努力提高LLM的准确性和可靠性，尽量保证它们生成的信息不仅连贯，而且是准确的。

二、如何 缓解 大模型幻觉?

要克服LLM幻觉需要从多个方面一起考虑，比如从训练层面和用户层面，以下是一些常见的策略：

2.1 方法一：提高训练数据质量

提高训练数据的质量对于减小LLM幻觉至关重要。 比如可以考虑**剔除那些不准确、有偏见的数据，并纳入多样化和可靠（比如经过事实审查）的数据来源。**

用于LLM训练的数据量越大（广度的数量都要考虑），最终训练得到的LLM出现幻觉的可能性就可能越小。

2.2 方法二：使用合适的训练算法

LLM如何训练也关乎后续在推理时是否会产生幻觉，因此**可以考虑在训练阶段融入一些有助于生成与事实一致的文本的策略。**

2.3 方法三：事实核查和验证机制

训练好的LLM在推理阶段仍然可以进行一些事实审查或者验证，比如**通过在生成过程进行一些判断或者交叉引用从而保证准确性。**

2.4 方法四：外部知识集成

使LLM能够访问和利用外部知识来源可以显著提高它们生成准确和可靠信息的能力。**将结构化数据、知识图谱或垂直领域的知识库集成到训练/推理过程中可以增强模型对事实信息的理解，并提高其生成可靠文本的能力。**

2.5 方法五：Human-in-the-Loop

通过在训练和测试阶段融入一些人类的反馈，从而纠正和完善模型的输出。训练阶段比如RLHF就是很好的例子，推理阶段同样也可以考虑一些交互式的推理方式将人类反馈融入其中。

2.6 方法六：偏见缓解

一方面要**在数据上做一些预处理**，另一方面也要**定期评估和监控模型输出的偏见，同时还可以通过一些手段减少推理过程出现的偏见。**

2.7 方法七：用户教育和批判性思维

LLM的幻觉可以缓解，但是难以根除。因此用户也应该有批判性思维，不能依赖LLM的输出，可以通过交叉引用信息，多个来源信息综合考虑。

三、总结

总的来说，缓解LLM幻觉需要从多个维度进行努力。

知识星球