大模型面经——关于大模型幻觉问题的深化理解

原创 喜欢卷卷的瓦力 瓦力算法学研所 2024年06月18日 10:00 上海

◈ 技术总结专栏 ◈

作者：喜欢卷卷的瓦力

本篇是基于大模型幻觉问题的进一步讨论，围绕具体度量大模型幻觉问题的方法与缓解方案，以及实际经验中幻觉问题最容易出现的情况，继续深入探讨大模型幻觉问题；值得一看~

大模型幻觉之前的介绍与缓解方案有讲过一些，也推荐大家参考~

大模型的幻觉问题
控制大模型幻觉也太难了吧

本篇来继续讲具体的幻觉问题的度量方法与缓解方案，来具体解决某些应用场景下的问题，例如：

- 应用于医疗垂直领域时如何判断大模型生成的内容是否存在幻觉?
- 应用于文档生成领域时如何判断生成内容与参考材料之间是否一致?

下面是一个本篇的快捷目录。

1. 幻觉问题如何量化
2. 如何缓解幻觉问题
3. 大模型在哪些问题上最容易出现幻觉
4. 幻觉一定有害吗?

一、幻觉问题如何量化

1. 应用于垂直领域时

应用于垂直领域时，由于存在一些领域内比较经典的命名实体词汇以及经典的实体关系三元组，因此量化幻觉问题可以基于这一点。

举个例子，如果是医疗领域大模型，一般各大研究院都会积累有一个相关知识库或知识图谱。当询问大模型糖尿病治疗的药物有哪些，问答过程之间可能关联到的实体与三元组类型有如下可能：

- 实体

症状——糖尿病

症状——高血压

药物——胰岛素

- <头实体，关系，尾实体>三元组

<糖尿病，推荐药物，胰岛素>

<糖尿病，关联症状，高血压>

那么由此就可以对大模型生成结果的幻觉进行量化了，这里推荐两种方法。

1）命名实体误差

命名实体（NEs）是"事实"描述的关键组成部分，那么可以利用NE匹配来计算生成 文本与参考资料之间的一致性。这种方法直接评估生成答案中实体词是否属于知识图谱。

直观上，如果一个模型生成了不在知识图谱中的实体，比如模型输出"糖尿病推荐的药物是健胃消食片"，那么它可以被视为产生了幻觉（或者说，有事实上的错误）。

2）利用信息提取系统

此方法使用信息提取模型将大模型生成内容的知识简化为关系元组<头实体，关系，尾实体>；并与从知识图谱中提取的元组进行比较。

2. 应用于文档生成时

这种方法跟提供的参考材料直接相关，理论上来说，提供的参考材料质量越高越详尽，那么这种量化方法就越准确。

1）蕴含率

定义为被参考文本所蕴含的句子数量与生成输出中的总句子数量的比例。为了实现 这一点，可以采用成熟的蕴含/NLI模型。

2）基于外置的问答系统

此方法的思路是，如果生成的文本在事实上与参考材料一致，那么对同一个问题， 其答案应该与参考材料相似。也就是说，我们可以通过问大模型一些参考材料中隐含的事实，然后基于大模型生成的答案来验证大模型的幻觉问题。

具体而言，对于给定的生成文本，问题生成模型会创建一组问题-答案对。接下来，问答模型将使用原始的参考文本来回答这些问题，并计算所得答案的相似性。

上面的话有点绕，我们来举个例子，也就是我们会训练一个输入是参考材料，输出是问题-答案对的问题-答案对生成模型，那么接下来：

- 假设我们要问大模型的prompt：

```
prompt：已知糖尿病一般使用胰岛素降血糖，请问是否还有其它药物适用于糖尿病？
```

- 将上述prompt输入训好的问题-答案对生成模型，得到如下结果

```
模型生成结果

问题：糖尿病一般用什么降血糖？
答案：胰岛素。

```

- 将上述生成的问题输入大模型，在计算大模型生成答案与上述问题答案的相似性即可得到结果。

二、如何缓解幻觉问题

理论上来说，只有创建高质量无噪声的数据集才是最关键的解决方案；但清洗、验证大规模数据，还有保证各个来源数据的质量难度太大了；因此多数论文还有一些工业界落地还是会去探索一些"治标不治本"的其他方法，下面列举一些比较实用性比较高的：

1. 通过使用外部知识验证主动检测和减轻幻觉

- 论文：A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation

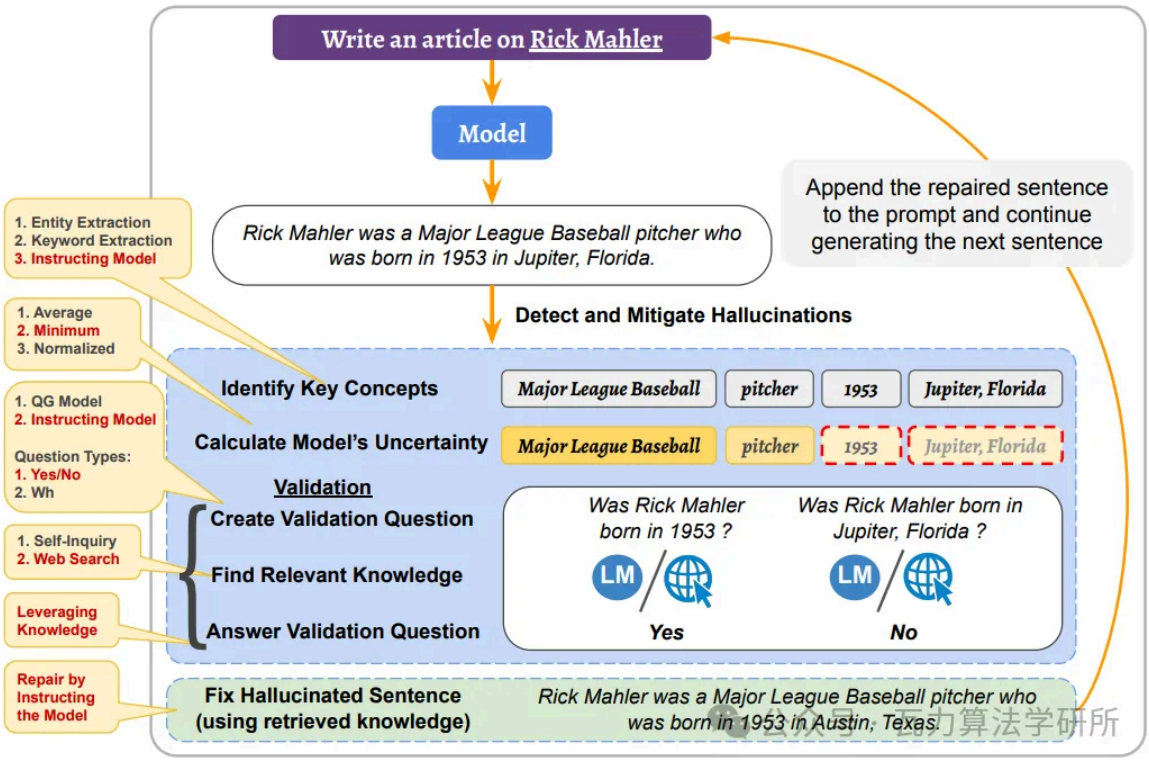
- 原理：

作者发现了两个问题：

1） 幻觉的生成是会传播的，比如一句话出现幻觉，后续生成的文本可能也会出现幻觉甚至更严重。这意味着，如果能够"主动"检测并减轻幻觉，那么也可以阻止其在后续生成的句子中的传播。

2）logit输出值（输出词汇表上的概率分布）可以用来获取幻觉的信号。具体地说，可以计算一个概率得分，当这个得分很低时，模型更容易产生幻觉。因此，它可以作为幻觉的一个信号， 当得分很低时，可以对生成的内容进行更具体的信息验证。

- 方法



检测阶段：首先确定潜在幻觉的候选者，即生成句子的重要概念。然后，利用其logit输出值计算模型对它们的不确定性并检索相关知识。

减轻阶段：使用检索到的知识作为证据修复幻觉句子。将修复的句子附加到输入（和之前生成的句子）上，并继续生成下一个句子。这个过程不仅减轻了检测到的幻觉，而且还阻止了其在后续生成的句子中的传播。

2. 修改解码策略

- 论文：Factuality Enhanced Language Models for Open-Ended Text Generation

- 原理：

作者认为，采样的随机性对生成句子的后半部分比生成前半部分影响更大，因此对事实性的损害也比在句子的开头更大。

因为在句子的开始没有前文，所以只要它在语法和上下文上是正确的，大模型就可以生成任何内容。然而，随着生成的进行，前提变得更为确定，候选单词会更少，从而导致句子更容易生成不符合事实的结果。

- 方法

引入了事实核心采样算法，该算法在生成每个句子时动态调整“核心”p。
在事实核心采样中，生成每个 句子的第t个标记的核心概率pt为

$$p_t = \max\{\omega, p \times \lambda^{t-1}\}$$

其中，

- λ是top-p概率的衰减因子，随着生成的token数量t 的增加，逐渐衰减p 的取值；
- ω是概率的下限衰减，为了避免p 衰减后过小，设置一个lower bound；
- p-reset ： 当一个句子生成完毕后，p 的值会因为t的增加而变得很小，当生成新的句子时，期望p能够恢复到原始的值

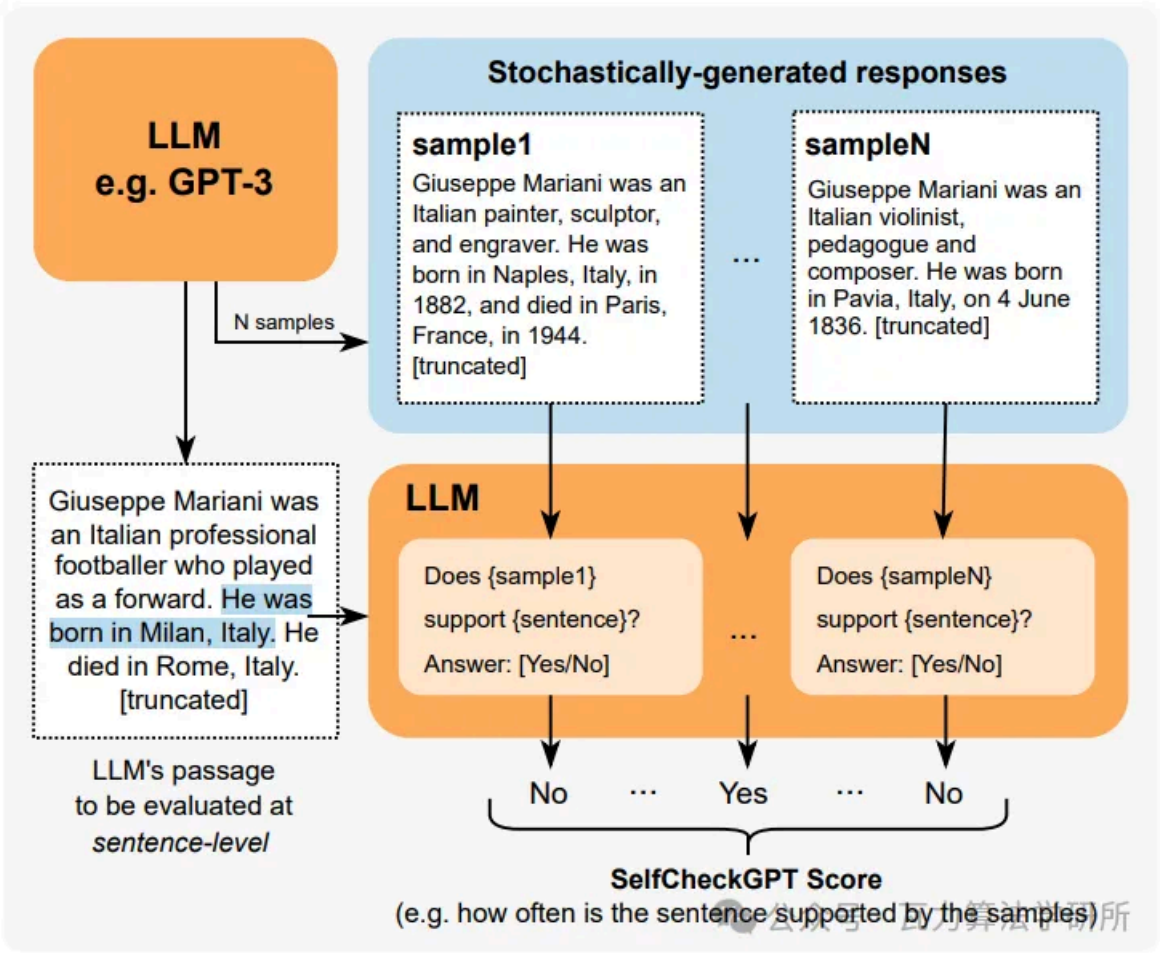
3. 采样多个输出并检查其一致性

- 论文：SELFCHECKGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models

- 原理：

这篇论文的主要思想是：如果模型真的掌握某个事实，那么多次生成的结果应该是相似的且事实一 致的；相反，如果模型在胡扯，那么随机采样多次的结果会发散甚至矛盾。

- 方法



从模型中采样多个response（比如通过变化温度参数）并测量不同response之间的信息一 致性，以确定哪些声明是事实，哪些是幻觉。

信息一致性可以使用各种方法计算，比如可以使用神经方法计算语义等价（如BERT Score）或使用IE/QA-based方法。

三、大模型在哪些问题上最容易出现幻觉

1. **数值混淆**：当LLM处理与数字有关的文本，如日期或数值时，容易产生幻觉。
2. **处理长文本**：在需要解读长期依赖关系的任务中，例如文档摘要或长对话历史，模型可能会生成自相矛盾的内容。
3. **逻辑推断障碍**：若模型误解了源文本中的信息，它有可能产生不准确的结论。因此，模型的逻辑推理能力至关重要。
4. **上下文与内置知识的冲突**：模型在处理信息时，可能会过度依赖于预训练阶段获取的知识，而忽略实际上下文，导致输出不准确。
5. **错误的上下文信息**：当给定的上下文包含错误信息或基于错误的假设时（如：“为什么高尔夫球比 篮球大？”或“氦的原子序数为什么是1？”），模型可能无法识别这些错误，并在其回答中产生幻觉。

四、 幻觉一定有害吗?

最后来探讨一个开放性的问题，可以想象一下幻觉在什么时候是无害的。

对幻觉的容忍度取决于具体的应用场景：在一些需要创造力或灵感的场合，比如写电影剧情，幻觉的存在可能带来一些奇思妙想，使得生成的文本充满想象力。当然也可能有其他更多的场景，可能需要小伙伴们一起想想。

想要获取技术资料的同学欢迎关注公众号，进群一起交流~

参考文献：

[1] A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation

[2] Factuality Enhanced Language Models for Open-Ended Text Generation

[3] SELFCHECKGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models