大模型的幻觉问题篇

来自： AiGC面试宝典

宁静致远 2023年09月29日 13:34

一、什么是 大模型幻觉问题?

1.1 大模型幻觉问题定义

- 定义：当模型生成的**文本不遵循原文（Faithfulness）或者不符合事实（Factualness）**，我们就可以认为模型出现了幻觉的问题。

1.2 何为 Faithfulness and Factualness?

- Faithfulness：是否遵循input content；
- Factualness：是否符合世界知识；

1.3 针对不同任务，幻觉的定义有何差异?

1. **数据源（source）不一致问题**

eg：摘要的数据源是document，data-to-text的数据源是data table，对话的数据源是对话历史，而开放域对话的数据源可以是世界知识；

1. **容忍幻觉的程度不一致问题**

在摘要、data-to-text任务中，非常看重response的Faithfulness，因此这些任务对幻觉的容忍程度很低；

而像开发域对话任务中，只需要response符合事实即可，容忍程度较高；

1.4 传统任务中的模型幻觉 vs LLMs 中模型幻觉

- **在传统任务里，幻觉大都是指的是Faithfulness**：
  - **Intrinsic Hallucination（信息冲突）**：LMs在生成回复时，与输入信息产生了冲突，例如摘要问题里，abstract和document的信息不一致；
  - **Extrinsic Hallucination（无中生有）**：LMs在生成回复时，输出一些并没有体现在输入中的额外信息，比如邮箱地址、电话号码、住址，并且难以验证其真假。（PS: 按照此定义，Extrinsic Hallucination有可能是真的信息，只是需要外部信息源进行认证）
- **而面向LLMs，我们通常考虑的幻觉则是Factualness**：
  - 因为我们应用LLM的形式是open-domain Chat，而不是局限于特定任务，所以数据源可以看做任意的世界知识。LLMs如果生成了不在input source里的额外信息，但是符合事实的，这种情况也可能是对我们有帮助的。

二、为什么 会 出现 大模型幻觉问题?

2.1 从 数据角度 进行分析

在 数据构建过程中，由于以下问题，导致 模型幻觉 的 发生：

1. 训练数据可信度问题。由于 大模型 的 训练数据 都是 通过 众包/爬虫检索 方式 收集得到的，这种数据构建方式的优点是量比较大，但是缺点是 包含 大量虚假信息。这种虚假信息 直接导致的问题就是使 模型出现错误认知；
2. 重复数据问题。过多的重复信息也可能导致模型的知识记忆出现bias，从而导致幻觉；

引用至 [3] Deduplicating training data makes language models better

2.2 从 模型角度 进行分析

不止是 数据角度问题，大模型幻觉问题 出现的原因 还 表现在 模型角度。

- **模型结构**：如果是较弱的backbone（比如RNN）可能导致比较严重的幻觉问题，但在LLMs时代应该不太可能存在这一问题；
- **解码算法**：研究表明，**如果使用不确定性较高的采样算法（e.g.，top-p）会诱导LMs出现更严重的幻觉问题。**甚至可以故意在解码算法中加入一些随机性，进一步让LMs胡编乱造（可以用该方法生成一些negative samples）

引用至 [4] Factuality enhanced language models for open-ended text generation

- **暴露偏差**：**训练和测试阶段不匹配的exposure bias问题可能导致LLMs出现幻觉，特别是生成long-form response的时候。**

引用至 [5] On exposure bias, hallucination and domain shift in neural machine translation

- **参数知识**：**LMs在预训练阶段记忆的错误的知识，将会严重导致幻觉问题。**

引用至 [6] Entity-based knowledge conflicts in question answering

三、如何 评估 大模型幻觉问题?

现有的传统幻觉评估指标和人类结果的相关性往往较低，同时大都是task-specific的 [7]。

3.1 Reference-based

Reference-based的指标有两类：

1. **基于Source Information和Target Reference**：利用一些统计学指标，比如**ROUGE、BLEU来评估输出结果和Source/Target信息的重叠度**：

2. **基于Source Information**：由于NLG任务里，Target输出往往是多种多样的，因此许多工作**只基于Source信息进行幻觉的评估**。比如Knowledge F1。

基于Reference的评价指标，**基本上只能评价Faithfulness，而无法评价Factualness，因此通常不适用于LLMs**。

3.2 Reference-Free

3.2.1 基于IE

- 介绍：**将知识限定于可以用三元组形式表示的关系和事件**，基于额外的IE模型进行抽取，接着使用额外模型进行验证；
- 缺点：
    - 可能存在IE模型的错误传播问题；
    - 知识被限定在三元组形式。

3.2.2 基于QA

- 介绍：

1. 第一步先**基于LM生成的回复**，使用一个QG(question generation)模型生成一系列QA pairs；
2. 第二步**给定Source Information，让QA模型对上一步生成的Question进行回复**；
3. 第三步则是**通过对比第一步的answers和第二步的answers，计算匹配指标，衡量模型的幻觉问题**；

- 缺点：
    - 可能存在IE模型的错误传播问题；
    - 难以评估Factualness，因为上述第二步里面，Source Information不可能包含全部的世界知识，因此对于一些问题难以生成可靠的回复。

引用至 [8] FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization

3.2.3 基于NLI

- 介绍：基于NLI的方法**通过NLI模型评估是否Source Information可以蕴含Generated Text，从而评估是否出现了幻觉现象**。
- 缺点：

1. Off-the-shelf NLI模型用于核查事实效果不是很好

引用至 [9] Evaluating groundedness in dialogue systems: The BEGIN benchmark.

1. 无法评估需要世界知识的幻觉问题：**仅能依赖于Source进行核查**；
2. 都是sentence-level的，**无法支撑更细粒度的幻觉检查**；

引用至 [10] Evaluating factuality in generation with dependency-level entailment.

1. 幻觉问题和蕴含问题实际并不等价：
    a. 例子：Putin is president. -> Putin is U.S. president (可以蕴含，但是是幻觉)

3.2.4 基于Factualness Classification Metric

- 介绍：标注/构造一批和幻觉/事实有关的数据，训练检测模型，利用该模型评估新生成文本的幻觉/事实问题。

引用至 [11] Knowledge-powered conversational agents

3.2.5 人工评估

- 介绍：目前为止最靠谱的，此外还可以依靠LLM打分（比如利用GPT4，但是GPT4也存在着严重的幻觉问题，即使经过retrival-augment，检索回来的信息也有可能是错误的）

四、如何 缓解 大模型幻觉问题?

4.1 基于数据的工作

4.1.1 构建高质量数据集

1. 人工标注

- **训练数据**：LLM上不可行，只适用于task-specific的幻觉问题；
- **评测数据**：构建细粒度的幻觉评估benchmark用于分析幻觉的严重程度和原因

引用至 [12] GO FIGURE: A meta evaluation of factuality in summarization

引用至 [13] Q2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering

1. 自动筛选：

- 利用**模型筛选出可能导致幻觉的数据并剔除**；
- 预训练时给更faithful的**数据加权（wiki vs. fake news）**，或者不使用可靠来源的数据（比如只选用经过人工审查的数据源，如wiki或者教科书，预训练）

引用至 [14] Vectara：让你的LLM应用告别幻觉!

4.2 模型层面的工作

4.2.1 模型结构

- 模型结构层面的工作往往focus在设计更能充分编码利用source information的方法，比如融入一些人类偏置，如GNN网络。
- 在解码时减少模型的生成随机性，因为diversity和Faithfulness往往是一个trade-off的关系，减少diversity/randomness可以变相提升Faithfulness/Factuality。

引用至 [15] Factuality enhanced language models for open-ended text generation

- **检索增强**被证明可以显著减少幻觉问题，e.g., LLaMA-index。

引用至 [16] Check your facts and try again: Improving large language models with external knowledge and automated feedback.

4.2.2 训练方式

- **可控文本生成**：将幻觉的程度作为一个可控的属性，利用可控文本生成技术进行控制。

引用至 [17] Increasing faithfulness in knowledgegrounded dialogue with controllable features

引用至 [18] A controllable model of grounded response generation

- **提前规划骨架，再生成**：sketch to content

引用至 [19] Data-to-text generation with content selection and planning

- **强化学习**：假设是基于word的MLE训练目标，只优化唯一的reference，可能导致暴露偏差问题。现有工作将减轻幻觉的指标作为强化学习的reward函数，从而减轻幻觉现象。

引用至 [20] Slot-consistent NLG for task-oriented dialogue systems with iterative rectification network

引用至 [21] improving factual consistency between a response and persona facts

- **多任务学习**: 通过设计合适的额外任务，可以达到减轻幻觉的效果。
- **后处理**：设计一个小模型专门用于fix幻觉错误。

引用至 [22] Improving faithfulness in abstractive summarization with contrast candidate generation and selection

4.3 可能的后续方向

1. 更细粒度的幻觉评估：
    a. token/phrase level instead of sentence level
    b. 更精细的幻觉分类体系：
        i. Intrinsic
        ii. Extrinsic
        iii. 其他类别：
            1. 按幻觉产生的原因分类（调用知识出错，还是缺少相应知识）
            2. 主观/客观幻觉
            3. 幻觉可能和时间（temporal）有关
2. 知识的定义和诱导：
    a. 怎么知道模型是否具备某一类知识，只是没有调用好?
    b. 知识的定义：
        i. 传统工作大都将wikipedia视作知识库，但它仅仅是世界知识的很小一部分
        ii. 如果将整个互联网当做世界知识，又不可避免的会有虚假信息的问题
3. 幻觉消除：
    a. 检索增强：互联网/外挂知识库(LLaMA Index)
    b. 强化学习（RLHF）
    c. 知识诱导/注入
    d. 直接修改LLM中错误记忆的知识：Model Editing工作，如ROME，MEMIT等