Siren’s Song in the AI Ocean:

A Survey on Hallucination in Large Language Models

及

Towards Mitigating Hallucination in Large Language Models

via Self-Reflection

阅读报告

班级：人工智能（一）2102班

姓名：孙澈

学号：20216491

目录：

1.背景及意义

1.1背景

1.2任务

1.3现状分析

1.4挑战和问题

1.5价值和意义

2.幻觉问题评估方法

2.1基准方法

2.2创新

2.3理论，算法

3．实验分析与讨论

3.1实验设置，数据

3.2实验对比结果与分析

3.3结果讨论

4阅读心得

4.1和课堂的哪一部分知识相关

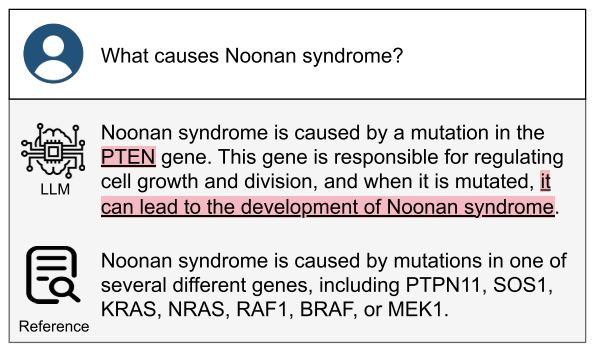
4.2方法的优势和劣势

4.3如何进行改进

**1.背景及意义**

**1.1背景**

在大语言模型出现之前，"幻觉"（hallucination）一词已经在NLP界被广泛使用，通常指生成无意义或者不忠于所提供来源内容的内容。由于大语言模型功能的多样性，现在幻觉的定义似乎已经被大大扩展，因此，论文中将大语言模型范畴内的幻觉分成了三类。第一类，是和输入相冲突的幻觉；第二类，语境冲突性的幻觉；第三类，与事实相冲突的幻觉。为了应对幻觉问题，本文中提出了一种包含知识获取和答案生成的交互式自我反思方法。通过这样的反馈过程，该方法稳步提高了生成答案的真实性、一致性和内涵。



作者举的例子：大语言模型中三种幻觉类型

幻觉问题示例： PTEN突变通常与努南综合征无关，因此答案听起来似乎合理，但却是幻觉

我将从对幻觉的定义，如何评估幻觉，大语言模型幻觉的来源，如何在训练和生成推理阶段减少幻觉，以及其他可以减少幻觉的方法来介绍该论文。

**1.2任务**

在研发大模型时，会涉及到预训练、微调、强化等多个阶段，每个阶段都会是幻觉的引入来源。首先，在预训练阶段，大语言模型会从大量训练数据中积累大量知识，然后将其存储在模型参数中。当被要求回答问题或者完成任务时，如果缺乏相关知识，或者内化了训练语料中的错误知识，模型就会产生幻觉。比方说，模型有时会将虚假的相关性误解为事实知识，也就是说幻觉与训练数据的分布之间存在很强的相关性，由于模型偏向于肯定测试样本，而样本的语料中本身也存在幻觉的现象，因此模型很容易复制甚至放大这种幻觉行为。其次，大模型有时会高估自己的能力。对于非常大的大语言模型来说，正确答案和错误答案的分布熵可能是相似的，这表明模型在生成错误答案时，与生成正确答案时同样自信，甚至是过度自信。第三，有问题的对齐过程可能会误导模型产生幻觉。大语言模型在预训练之后，通常都会经过微调和对齐的阶段，如果模型在预训练阶段，没有获得相关知识。那么在训练指令时，实际上是一个错误的对齐过程，促使模型产生幻觉。另一个潜在问题是"谄媚"（sycophancy）即模型可能会生成偏向用户观点的回答，而不是提供正确或者真实的答案，从而也会导致幻觉。第四点，模型采用的生成策略本身就存在潜在得风险。我们都知道大语言模型按照顺序来生成响应文本，每次输出一个token，但是即使模型意识到自己出现了早期错误，它们有时也会过度承诺，换句话说，大语言模型可能更喜欢用“滚雪球”来实现自我一致性，而不是从错误中恢复。有研究认为，局部最优并不一定能确保全局最优，早期的局部预测可能会导致模型难以形成正确的结果。同时，基于抽样的生成策略，比如top-p和top-k，它们所引入的随机性也可能是幻觉的一个潜在来源。

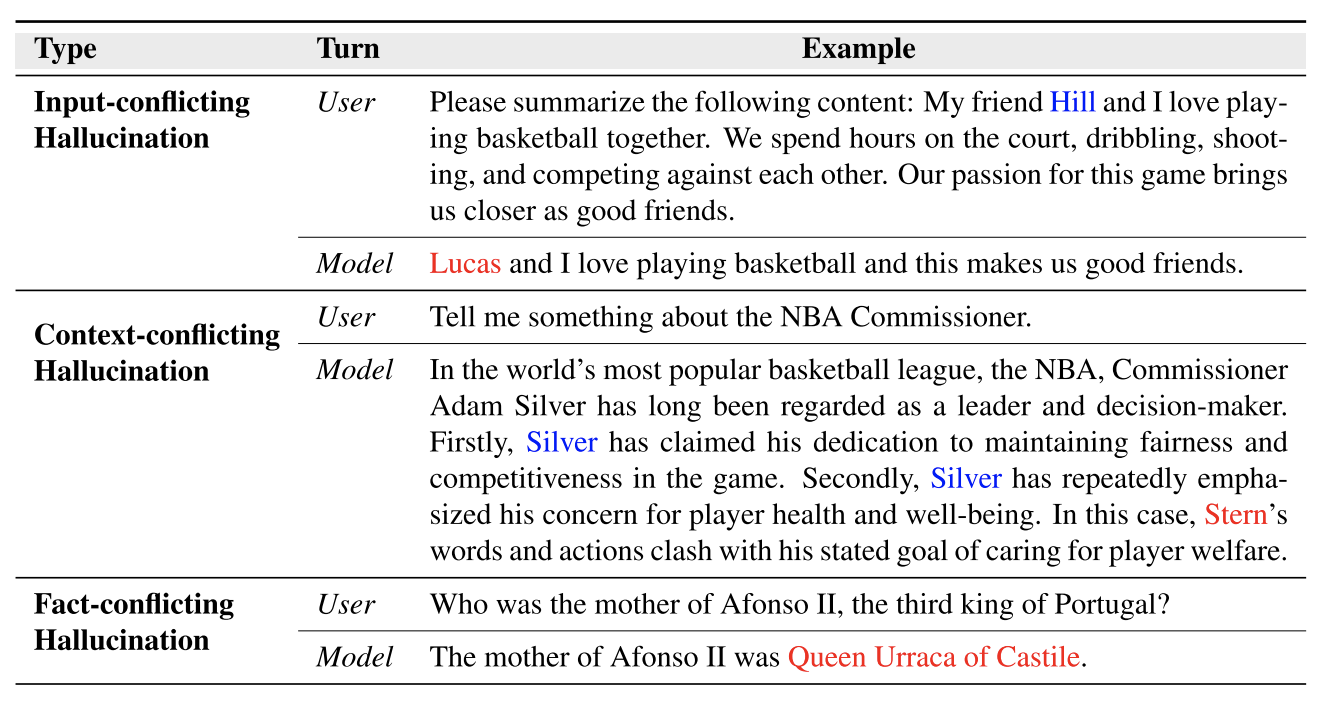
**1.3现状分析**

大语言模型偶尔会生成与用户输入不符的内容，或者是与之前生成的内容相矛盾的内容，又或者与事实不一致的内容。这种现象通常被称为"幻觉"，这极大地降低了大语言模型在真实世界场景中的可靠性，比如，大语言模型可能会编造出错误的医疗诊断或者治疗方案从而对实际生活造成风险。如果进行分类，我们会发现大模型的回答存在着多种不同类型的问题：

第一类，是和输入相冲突的幻觉，也就是大语言模型生成的内容与用户提供的源输入相背离，从而产生这种幻觉；

第二类，语境冲突性的幻觉，指的是大语言模型生成的内容与之前生成的信息本身相冲突。模型在生成冗长的、或者多轮回答的时候可能会表现出自我矛盾。这种幻觉产生的原因，可能是模型在整个对话过程中失去了对上下文的跟踪，或者无法保持长期记忆的一致性所造成的。当大语言模型的反应模棱两可，可以有多种解释时就会出现上下文无关的问题，这种回答不一定是错误的，但是它无法为用户问题提供有用的答案；

第三类，与事实相冲突的幻觉。这种幻觉指的是模型生成的内容与现有的世界知识相矛盾，这种幻觉问题比较严重，很容易误导知识不足的用户。同时，大语言模型中的偏见与生成文本中的不公平或者偏见表现有关，这些偏见可能来源于训练数据，训练数据通常包括历史文本、文学作品、社交媒体内容和其他来源，这些来源可能从本质上反映了种族偏见、性别偏见、刻板印象或者歧视性信念等等。由于RLHF的过程中，也有可能会导致大语言模型的过度优化，导致信息不足的状态。



作者定义的三种幻觉问题

**1.4挑战和问题**

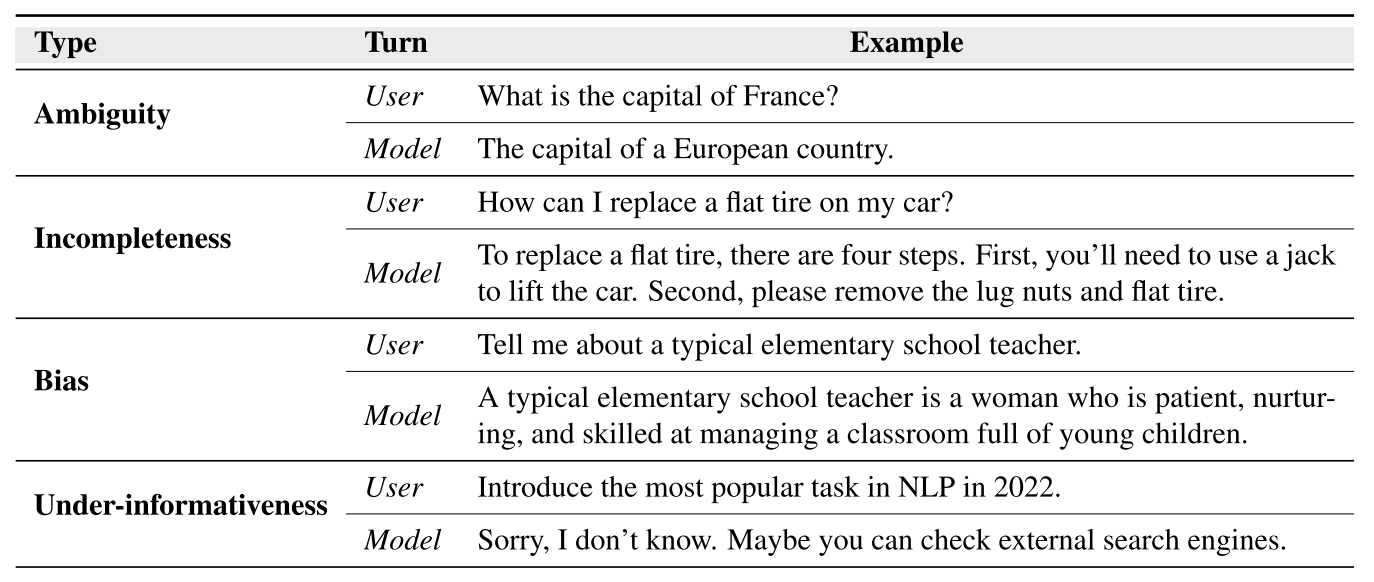
虽然传统自然语言生成（NLG）环境中的幻觉已被广泛研究，但在大语言模型领域中，理解和解决幻觉问题还有更困难的挑战：

1. 大量训练数据：LLM 预训练使用从网络获得的数万亿个token，不但很难定位、移除脏数据，也很淡分析大语言模型是学什么数据而导致的幻觉，因此很难消除捏造的、过时的或有偏见的信息；

2. 大语言模型的多功能性：通用型大语言模型有望在跨任务、跨语言和跨领域环境中表现出色，同时大语言模型出现了涌现能力，让综合评估和缓解幻觉的难度更上一层。

3. 错误难以察觉：大语言模型功能强大也有副作用，LLM生成的文本通顺、幻觉隐蔽，生的可能会生成最初看起来非常可信的虚假信息，这使得模型甚至人类都难以检测幻觉。

此外，相较于小模型，大语言模型生产周期长，一般要经过pre-training、SFT、RLHF过程、模糊的知识边界和LLM的黑箱特性也使检测、解释、解释变得复杂。



除了幻觉之外，LLM可能会暴露的各种问题的示例

**1.5价值和意义**

1. 信息准确性：解决幻觉问题可提升生成文本的信息准确性，减少误导用户或传播错误信息的风险。

2. 提升LLM用户信任度：用户对于大型语言模型的信任至关重要。如果生成的内容经常不准确或虚构，可能会失去用户的信任。通过解决幻觉问题，可以增强用户对模型的信任感，使其更愿意接受生成的信息。

3. 涉及重要决策的场景中，生成的信息必须是可靠的。解决幻觉问题有助于提高模型在商业应用中的可用性和价值，减少因误导而导致的潜在风险。

4. 社会责任：大型语言模型在社会中的广泛应用需要负责其影响。解决幻觉问题有助于确保模型的应用不会对社会产生负面影响。

5. 法律和伦理合规：解决幻觉问题有助于使模型在法律和伦理层面上合规，避免可能的法律责任问题。

**2.幻觉问题评估方法**

**2.1基准方法**

在本节中，作者利用LLM的零样本能力，直接从五个数据集中向LLM提出医学问题。然后全面评估和分析生成的答案，重点检查幻觉的发生。我们评估五个法学硕士生成的答案，包括三个普通LLM和两个在医学领域微调的LLM。Vicuna通过在 ShareGPT 的用户共享对话上微调 LLaMA 进行训练。 AlpacaLoRA采用低秩适应（LoRA）来复制斯坦福大学 Alpaca 模型的结果。 ChatGPT使用人类反馈强化学习（RLHF）来解释提示并提供全面的响应。 MedAlpaca建立在 LLaMA 框架之上，并在指令调整格式的医学对话和 QA 文本上进行了微调。 Robin-medical是使用 LMFlow 在医疗领域微调的 LLaMA。

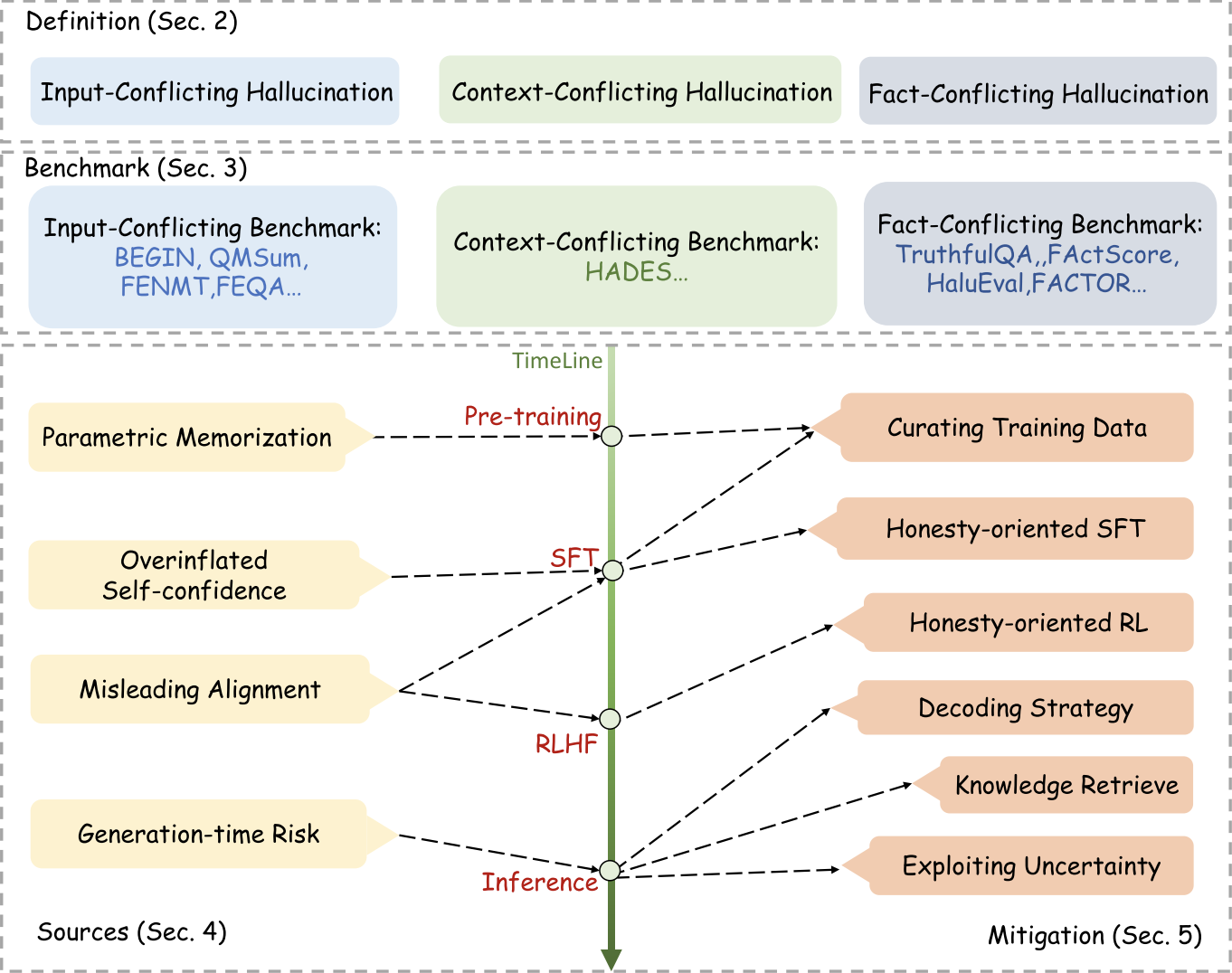
PubMedQA是一个生物医学QA数据集，包含1k个专家标记的实例，其中包括来自研究文章标题的问题、作为上下文的摘要、来自抽象结论的长答案以及简洁的是/否/可能的答案。MedQuAD包含来自美国国立卫生研究院网站的47,457个QA，涵盖各种医学主题，包括疾病、药物和诊断测试。作者使用 MEDIQA2019挑战赛的医学QA数据集，并将得分3和4的答案视为黄金答案。LiveMedQA2017包含用于问题分析和回答系统的带注释的医学QA。MASH-QA包括来自消费者健康领域的34k QA，专为多答案跨越医疗保健QA而设计。除了PubMedQA之外，这些数据集中的答案注释都涉及从真实的Web内容中手动提取和复制。虽然答案是相关且可验证的，但在上下文连贯性和问题关联方面还有改进的空间。

针对不同类型的幻觉，采用的评估方式是不一样的。首先，论文一中的作者提出了各种基准来评估大语言模型的幻觉。其中比较具有代表性的基准包括TruthfulQA、FactScore等现有的基准主要根据大语言模型的两种不同的能力来评估幻觉。分别是生成事实陈述，或者区分事实陈述与非事实陈述的能力，对应着生成式基准和判别式基准。这两种评估形式是有一定区别的，具体来说，生成式基准将幻觉视为一种生成特征，类似于流畅性和连贯性，并对模型生成的文本进行评估。例如TruthfulQA考察的是大模型回答问题的真实性，而FactScore考察的是大模型为某个人生成传记的事实准确性。

与此相反，判别式基准考虑的是大模型辨别真实陈述和幻觉陈述的能力。具体来说，HaluEval要求模型能够确定状态信息中是否包含幻觉信息，而FACTOR则研究模型是否能够赋予事实陈述比非事实陈述更高的可能性。同时也要说明的是，TruthfulQA也支持辨别形式，它提供了另一个替代方案，可以测试模型辨别真实陈述的能力。

有了评估基准之后，接下来就是确定评估的任务形式。分成三种，分别是问题解答、指令提示和文本补全。而以上大多数基准都需要人类注释者来创建数据集或者保证质量。评估指标的方法也主要分为人工评估和模型的自动评估。在自动评估方面，TruthfulQA训练了一个基于GPT-6.7B模型，可以根据问题的注释对答案的真假进行分类；AlignScore建立了一个统一的函数来评估两个文本之间的事实一致性。而FactScore首先使用通道检索器来收集相关信息，然后采用一个评估模型，比方说LLaMA-6B，使用检索到的知识来确定状态的真实性，并进一步采用微观F1分数和误差率，来评估自动指标与人工评估相比的可靠性。

减少幻觉现象的产生需要分成两个阶段，分别是训练阶段和生成推理阶段。训练阶段又可以分为三个时期，分别是预训练时期、SFT时期和RLHF时期。现有的工作认为，大语言模型的知识大多是在预训练阶段获得的。如果在预训练语料库中存在类似于错误信息之类的噪声数据，就可能会破坏大语言模型的参数知识。因此减少幻觉的直观方法可以是人工或自动整理预训练语料库，尽可能减少无法验证或者不可靠的数据。其实在大语言模型时代之前，已经有了一系列致力于人工消除噪声训练数据来减轻幻觉的工作。比方说专注于"数据到文本"（data-to-text）的任务，并邀请人工根据给定的知识库手动编写干净准确的回复。结果表明，使用这种经过编辑的训练数据可以有效减少幻觉。同样，在现有的表格到文本的数据集中，对文本进行人工提炼这一过程也可以大大减少事实幻觉。不过，随着现在预训练语料库的规模越来越大，在预训练期间整理训练数据变得越来越具有挑战性，例如Llama对大约两万亿个token进行预训练。因此，与人工整理相比，目前更实用的方法是自动选择可靠数据或者过滤掉噪声数据，例如GPT-的预训练数据就是基于与一系列高质量参考数据的相似性进行清理。Falcon通过启发式规则从网络中仔细提取高质量数据，并证明了经过适当分级的相关语料库可以产生强大的语言模型，Llama在构建预训练语料库时从维基百科等高度事实性的来源中向上抽取数据。



首先将LLM幻觉分为三种不同的类型，然后引入相应的评估基准。以及LLM 整个生命周期的缓解策略（预训练→SFT→RLHF→推理）

**2.2创新与理论**

在预训练过程中，鉴于现有的预训练语料库规模庞大，目前的研究主要采用简单的**启发式规则**来选择和过滤数据，而未来研究的方向是如何设计更有效的选择或过滤策略；其次，当前的大语言模型一般都会经历一个被称为"监督微调"（SFT）的过程，用来从预训练中获取所需的知识并学习如何与用户互动。一般来说，SFT要首先标注或者收集海量的任务指令跟踪数据，然后使用最大似然法（MLE）在这些数据上对预训练过的基础模型进行微调。与预训练类似，要减少SFT阶段的幻觉也需要对SFT的训练数据进行整理，但是SFT数据量相对较小，因此可以选择手动或者自动整理。要注意的是，在SFT过程中，可能会因为行为克隆而引起大语言模型产生幻觉，行为克隆是强化学习中的一个概念。简单来说，这种方法只是简单地模仿行为而不是为了学习实现最终的目标。大语言模型的SFT过程，可以被视为行为克隆的一个特殊案例，即模型通过模仿人类的行为来学习互动的形式和风格。对于基础模型来说，尽管已经将大量知识编码到了自己的参数中，但是仍有一些知识会超出它们的能力范围，通过克隆SFT过程中的人类行为，模型学会了以肯定的语气回答所有问题，而不去评估这些问题是否超出了它们的知识边界。因此，在推理过程中，如果被要求回答与没有学习过的知识相关的问题，他们很可能会自信地产生幻觉。

解决这个问题的方法之一，是采用**以诚实为导向的SFT**，即在SFT数据中引入一些诚实的样本，诚实样本指的是承认自己没有能力做出回答，比如"对不起、我不知道"等等。Moss项目开源的SFT数据中就包括这类诚实样本，使用这些样本的模型可以学会拒绝回答特定问题，从而帮助减少幻觉，不过，总体来说，由于SFT数据量可以接受，因此由人类专家进行人工整理依然是首选方案，同时由于标注人员不了解模型的真实知识边界，诚实样本也存在知识的盲区，因此这种方法也只能作为SFT阶段的次优方案。

考虑到SFT阶段的局限性，Schlman提出**在RLHF期间来解决幻觉问题**，他设计了一种专门用于减轻幻觉的特殊奖励函数，这里的"Unhedged/HedgedCorrect/Wrong"指的是大语言模型用肯定或者犹豫的语气，提供正确或错误的答案。它的核心理念是鼓励学习者挑战前提、表达不确定性，并通过从特殊的奖励中学习来证明自己的无能，这种方法称为"诚信导向强化学习"。它与"诚信导向SFT"相比有几个优点，首先它允许大语言模型自由探索知识的边界，从而增强对分布以外案例的概括能力，其次它还减少对大量人工标注的需求，并且消除了对标注者猜测模型知识边界的要求。这种方法的挑战在于，经过强化学习调整的模型，可能会表现出过度保守，比如ChatGPT拒绝回答它已经明确知道答案的问题，这可能是由于奖励函数的设计不合理，或者由于训练数据的质量不高所导致，那除了在训练时减少幻觉意外，在推理时减少幻觉，可能会具有更好的成本效益和可控性。

**3．实验分析与讨论**

**3.1实验设置，数据**

如今，现有研究大多数大模型幻觉问题都集中在在RLHF阶段减少幻觉。第一类就是设计解码策略，解码策略，比如贪婪解码和波束搜索解码，他们决定了我们如何从模型生成的概率分布中，选择输出TOKEN，通过评估发现，在事实性方面，核采样效果不如贪婪解码，可能是因为top-p采样引入的随机性，会无意中的导致幻觉。因为模型往往会编造信息来产生多样化的反应，有研究人员就提出了一种称为事实核采样的解码算法，希望利用top-p和贪婪解码的优势，在多样性和事实性之间取得更有效的平衡。还有研究团队提出了一种推理时干预方法ITI，他们发现Transformer模型中的某些注意力头，对于模型生成内容的真实性至关重要，在推理阶段，通过在注意力头上使用一种特殊的指令干预激活方式，可以有效的提升大模型生成内容的真实性，而且这种方法跟RLHF相比，所需要的成本非常低。

除此之外，有研究表明，大语言模型有时在处理下游任务，尤其是当检索到的知识与模型的参数知识相冲突，无法充分关注检索到的知识。为了解决这个问题，可以采用**直接的上下文感知解码策略CAD**，CAD方法能够让模型更多的关注上下文信息，而不是过度的依赖于自身的参数知识来做出决策。实验结果表明，CAD能够有效的激发模型利用检索知识的能力，从而减少下游任务中的事实幻觉。总的来说，设计解码策略可以减少大语言模型在推理过程中的幻觉，而且这种方法通常是即插即用的易于部署，不过大多数现有研究，都要求访问token级别的输出概率，而目前大部分模型只支持有限的API。

第二个研究方向就是**借助于外部知识**，作为补充证据来帮助大语言模型提供真实的回复，这也是最近兴起的一种解决方案，这种方法呢通常包括两个步骤，第一是准确的获取与用户指定相关的知识，一旦获取了有用的知识，第二步需要利用这些知识来指导应答的生成。我们知道，通过大量的预训练和微调，大语言模型已经将大量的知识内化到了自己的参数中，这些知识可以称为参数知识，但是不正确或者过时的参数知识就很容易的会导致幻觉。

为了解决这个问题，从可靠的来源获取可靠的最新知识就成为了模型的一种热补丁。而对于如何获取外部知识，又可以分成多种不同的类型。一种是通过外部知识库，如大规模的非结构化的语料库、结构化的数据库，甚至是整个互联网。证据检索过程通常会采用各种稀疏或者密集的检索器，搜索引擎也可以被视为一种特殊的信息再提取器，此外，还包括其他一些参数知识指导框架，可以从微调的白盒模型的参数存储器中重新汲取知识；另一种是通过外部的工具，比如FacTool和CRITIC。一旦获得了相关的知识，就可以在不同阶段加以利用来减轻模型的幻觉。现有的知识利用方法也可以大致分为两类，一种是生成式补充，即把知识和用户的查询先联合起来，再作为提示发给模型，这种方法既有效又易于实施，也被称为上下文知识，context knowledge，而大语言模型本身就具有很强的上下文学习能力，所以能够从中提取和利用有价值的信息；另一种是事后纠正，也就是在后处理阶段，构建一个辅助的矫正器来纠正幻觉。矫正器可以是另外一个大语言模型，也可以是一个特定的小模型。这种矫正器首先与外部知识源互动，收集足够的证据，然后纠正幻觉。这种利用外部知识来减轻幻觉的方法，不需要修改大语言模型，是一种即插即用的高效的解决办法，其次便于向模型传输专有的知识和实时更新信息，利用这种方法允许生成的结果追溯到源头的证据，从而提高了模型生成的信息的可解释性。但是，这个方法也存在一些现实的问题。首先是知识验证，如果外部知识来源是基于互联网的，那么如何验证这些知识的真实性是一个有待于解决的开放性和挑战性的问题。其次，检索器矫正器的性能和效率如何提升和优化，最后检索到的知识可能与模型存储的参数知识相冲突，当冲突发生的时候，模型可能就无法充分利用外部的知识，而且如果外部知识的上下文过长，可能就会造成模型性能的明显下降。

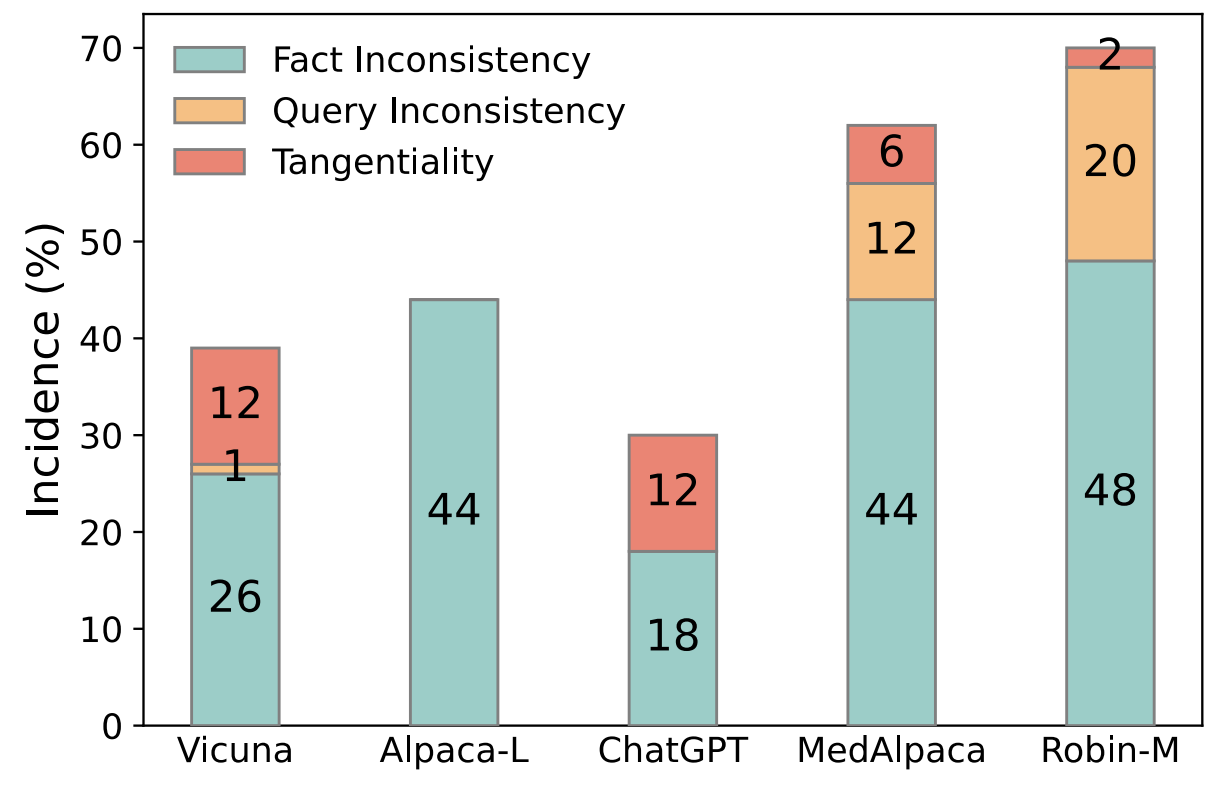
除了解码策略和外部知识以外，第三个研究方向就是利用不确定性。不确定性是推理过程中，保护和减少幻觉的重要指标，通常它指的是模型结果的置信度，不确定性可以帮助用户确定什么时候应该信任大语言模型，只要能够准确描述模型响应的不确定性，用户就能够过滤或者纠正模型的高不确定性的声明，因为这类声明更容易是捏造的。一般来说，估算大于模型不确定的方法可以分为以下三种，首先是基于logit的估计，这是一种基于对数的方法，它需要获取模型的对数，通常是通过计算token级的概率或者是熵来确定不确定性；其次是基于口头的估计，直接要求模型表达它的不确定度，这种方法之所以有效，是因为大语言模型的语言表达能力和服从指令的能力很强，也可以使用思维链提示来加强这种方法；第三种是基于一致性的估计，这种方法基于这样一个假设：当模型犹豫不决并且对事实产生幻觉的时候，他们很可能会对同一个问题做出逻辑上不一致的回答。

**3.2实验对比结果与分析**

1. 事实不一致：指答案提供的信息与事实不一致或相冲突。即模型在回答问题时未能正确回忆相关知识。

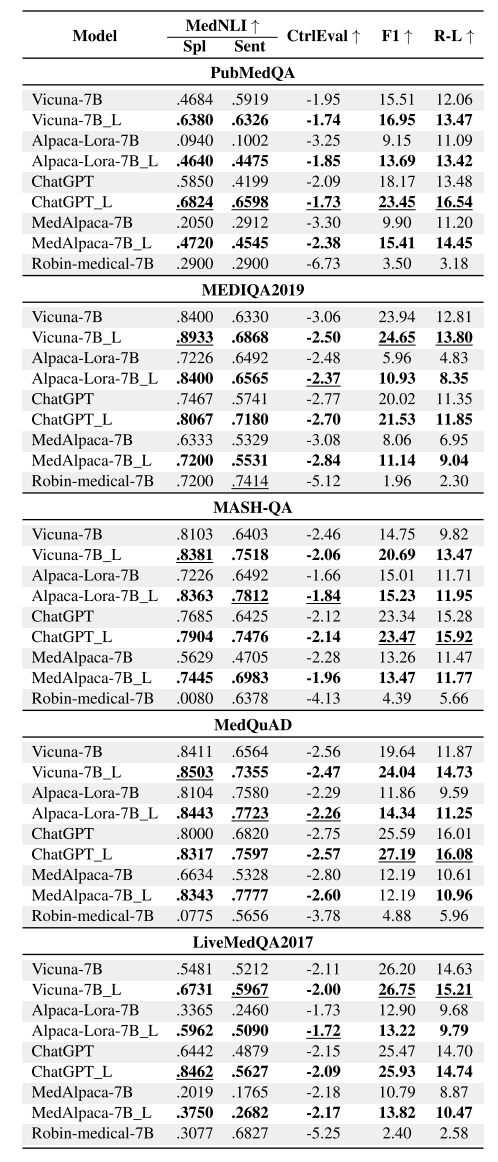
2. 查询不一致：指答案与查询无关或无意义。即模型既没有回答问题也没有适当地调用相关知识。

3. 切题：指答案提供与主题相关的信息，但不直接解决问题。即模型没有进一步处理掌握的知识（例如归纳、演绎和逻辑推理）。

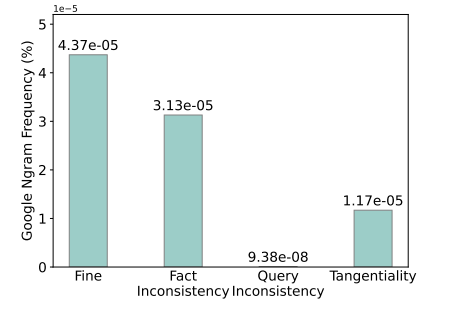


每个模型中每一类有问题的答案的发生率。

**3.3结果讨论**

****

使用交互式自我反思循环 (\_L) 和基线自动评估LLM



每个类别有问题的答案的 Google Ngrams 频率

有问题的答案的出现频率低于好的答案。这表明低频可能是产生幻觉的潜在原因，这需要更多的探索来证明。

**4阅读心得**

**4.1和课堂的哪一部分知识相关**

在探讨大语言模型的幻觉问题之前，我先系统学习大语言模型的相关知识，为此专门深入研读了中国人民大学的LLM文献综述。同时我学习了instructGPT，深入了解其中的基于人类反馈的强化学习方法（RLHF）。通过这些学习，我逐渐理解了大型语言模型的基本工作原理和训练技术。

除了理论学习外，我还积极尝试通过服务器进行Llama2模型的微调。看了很多视频并且尝试实操后，虽然没有配置成功，但是这一过程让我更深入地了解了模型微调的步骤和技巧，同时也使我更好地理解模型在特定任务上的表现和调整方法。

与课堂中所学的知识相对应，这一部分的学习帮助我建立了对大语言模型更深层次理解的基础。课堂上获得的理论知识如机器学习、深度学习、逻辑回归知识等为我后续学习生成式AI提供了指导，使我能够更加全面地把握大语言模型的关键概念和技术。通过结合课堂学习和实际操作，我在这个领域的知识体系更加丰富，为我深入研究大语言模型的幻觉问题奠定了基础。

**4.2方法的优势和劣势**

利用不确定性来识别和减轻大语言模型的幻觉，是当今很有前途的一个研究方向，然而这三种主要方法，也面临着一些挑战。

基于对数的方法越来越不适用于现代的商用大语言模型，因为这些模型通常都是闭源和黑盒的，导致它输出的对数无法访问。其次关于基于口头表达的方法，模型在表达它的信心的时候，往往会表现出高度的过度自信，第三在一致性的方面，如何有效的测量不同回答的一致性，仍然是一个尚没有解决的问题。

**4.3如何进行改进**

生成任务中的幻觉对人工智能的责任感和可信度提出了重大挑战。作者在一般和特定领域大语言模型的医学GQA背景下彻底、系统地研究了这个问题。为了应对这一挑战，作者提出了一种迭代自我反思方法，对背景知识和答案采用生成-分数-细化策略。经验证明该方法在减少幻觉方面是有效的、可推广的和可扩展的。

而我认为，我们还需要继续调查幻觉的根本原因，并且在大模型的各种下游任务中检查这种现象，也可以扩展作者所用方法来解决与这些任务相关的挑战。

文章从对幻觉的定义、如何评估幻觉、幻觉的来源以及如何减少幻觉几个方面较为完整的介绍了有关于大语言模型幻觉的各个方面。在我看来，幻觉跟大模型就是一体的关系，就像机遇和风险并存一样，我们只能够想办法尽量去控制减少风险却无法完全消除风险。或者说幻觉本身就是知识的一部分，也是人类认知的一部分。我们需要的是如何学会与幻觉并存，或者如何更好的区分幻觉与事实**。**