



苏州大学

SOOCHOW UNIVERSITY

论文汇报

刘承伟

2023-11-21

工作

▶ 文献阅读:

1. 《Unveiling the Siren's Song: Towards Reliable Fact-Conflicting Hallucination Detection》 Zhejiang University/ 阿里
2. 《Can We Edit Multimodal Large Language Models?》 Zhejiang University/ 腾讯 (EMNLP2023)



Unveiling the Siren's Song: Towards Reliable Fact-Conflicting Hallucination Detection

在数字时代，大型语言模型（LLMs）如 ChatGPT/GPT-4 因其广泛的实际应用而受到瞩目。然而，它们在网络平台上产生的事实冲突幻觉问题限制了其应用范围。本文介绍了一个名为 **FACTCHD** 的事实冲突幻觉检测基准，能够在 LLMs 的“查询-响应”环境中评估事实性。

FACTCHD 集成了多个领域的事实性知识，涵盖了广泛的事实性模式，如原始事实、多跳推理、比较和集合操作模式。其独特之处在于，它旨在将基于事实的证据链相互结合，当预测一个声明的事实性或非事实性时，提供有说服力的理由。然而，全程依靠人工注释来收集大量数据不仅耗时而且资源消耗巨大，其可扩展性亦有限。

因此，本文建议采用现有的知识图谱（KG）和文本知识作为数据来源，提出了一种基于知识事实的数据构建策略，并结合半监督注释的方法，以促进上述基准的创建和发展。与此同时，结合领域 **KG** 构建幻觉检测数据集的策略拓展性较高，进一步为未来在高风险领域如金融、医疗和法律等领域应用生成性 **AI** 提供了可能



Unveiling the Siren's Song: Towards Reliable Fact-Conflicting Hallucination Detection

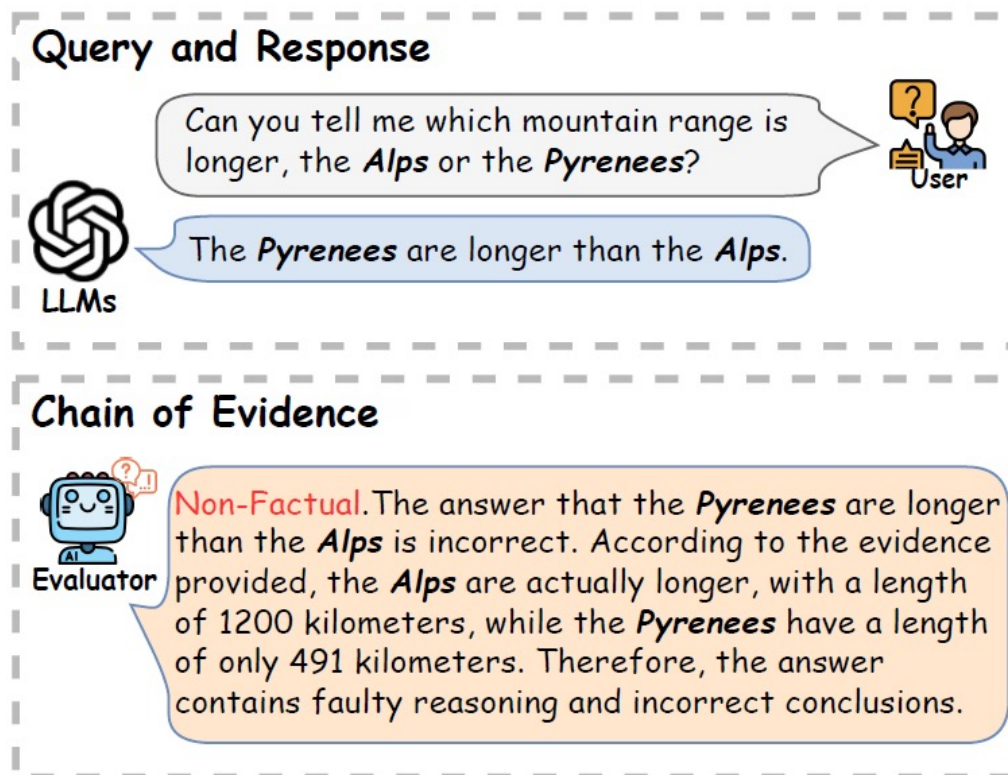


Figure 1: Illustrating Fact-Conflicting Hallucination Detection.

Unveiling the Siren's Song: Towards Reliable Fact-Conflicting Hallucination Detection

为了评估多种大型语言模型（如 Alpaca、ChatGPT 等）的效能，我们开展了一系列基准实验，利用我们的 FACTCHD 在不同设置下进行测试——零样本学习、上下文内学习、专门为检测特定专业知识进行调优，以及通过检索/工具进行知识增强。尽管调优和知识增强对事实冲突幻觉的评估产生了积极影响，但开源的大型语言模型和 ChatGPT 在精准和稳健地检测事实不准确性方面仍面临挑战。

因此，本文引入了一个“三角测量”框架进行幻觉辨别，其使用交叉参考生成器和搜索工具来裁决有问题的事实回答。初步实验验证了不同 LLM 在识别事实冲突幻觉方面的不同表现，并确认了本文提出方法的优越性。



Unveiling the Siren's Song: Towards Reliable Fact-Conflicting Hallucination Detection

FactCHD 基准的构建

基于上述定义，我们构建了涉及多个领域的 FactCHD 基准，其中包含了一系列全面的训练样本，并额外添加了 6,960 个经过精心筛选的样本，用于评估 LLM 生成的事实冲突幻觉。我们的数据集确保了 **factual** 或 **non-factual** 类别之间的平衡，为评估提供了一个坚实的平台。值得注意的是，FactCHD 具有以下三个显著特点：

(1) 如图 1 和 2 所示它包含了多样化推理模式，包括多跳、比较和集合操作，并涉及健康，医疗，科学，气候等多个领域；

(2) FactCHD 遵循现实场景，提供 “Query-Response” 对和相关证据来验证提供的信息；

(3) 该基准测试经过精心设计，在初始数据构建阶段利用知识图谱 (KGs)，经过细致的人工验证以确保质量。此外，该数据集本身允许通过基于知识图谱的更新进行扩展，从而在保持时代性和可扩展性方面具有独特优势。

接下来，本文将介绍事实冲突幻觉检测基准测试的设计原则。



Unveiling the Siren's Song: Towards Reliable Fact-Conflicting Hallucination Detection

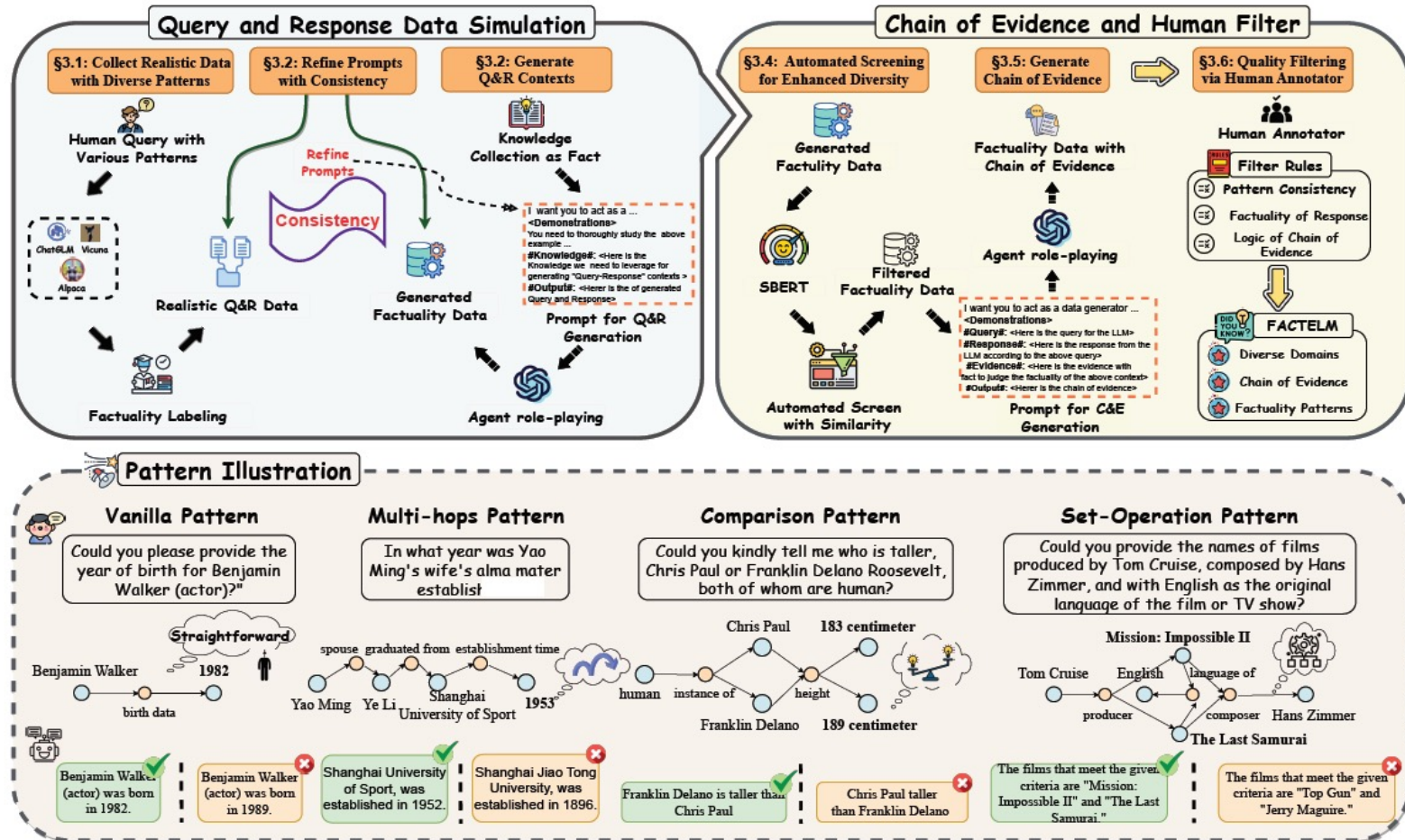


Figure 1: Overview of the construction and pattern illustration of FACTCHD

Unveiling the Siren's Song: Towards Reliable Fact-Conflicting Hallucination Detection

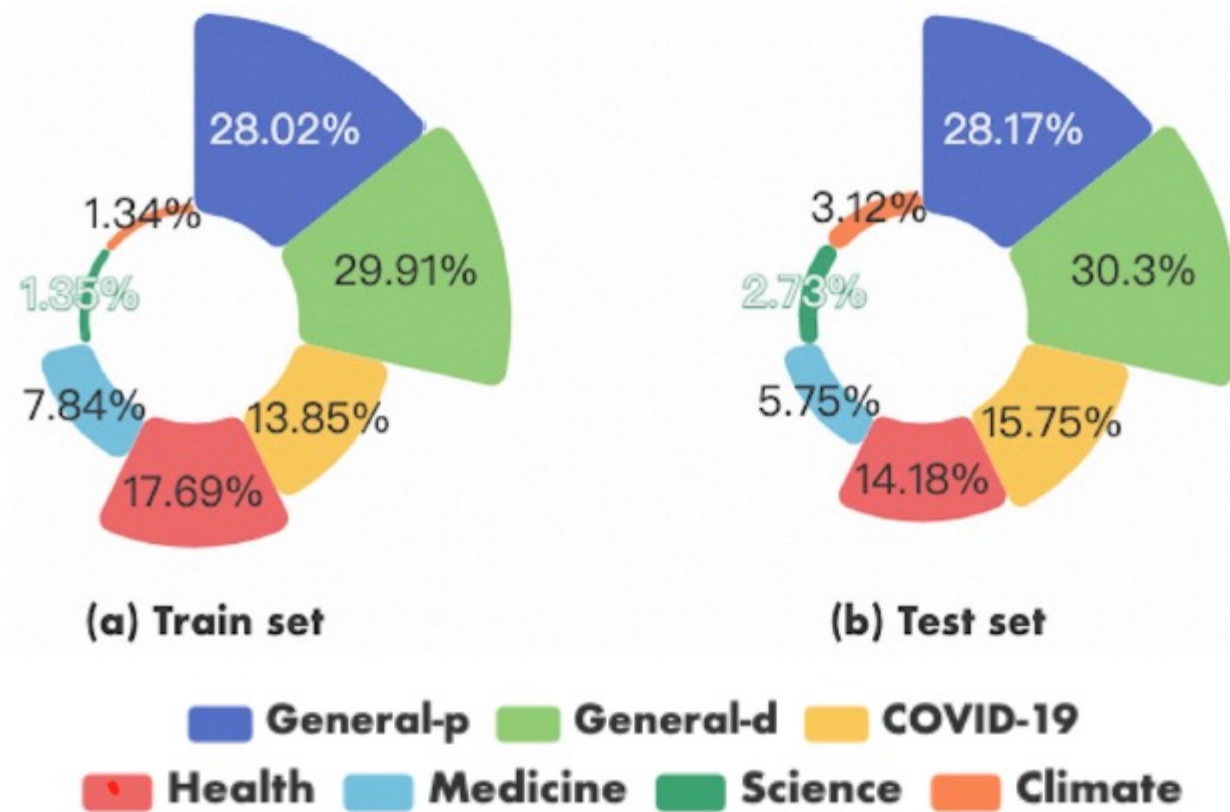


Figure2: Distribution of FactCHD across various domains

Unveiling the Siren's Song: Towards Reliable Fact-Conflicting Hallucination Detection

1.1 生成“查询-回答”上下文

将知识作为事实加以利用：如图 3 所示，知识图谱（KG）以其庞大的实体和关系数据存储库，不仅支持复合推理，而且为事实信息提供了基础结构。同时，文本知识在提供额外和细致信息方面起着关键作用。基于这一点，本文的目标是探索利用现有知识作为事实支撑半自动构建事实冲突幻觉中的能力，具体：

（1）本文从 Wikidata 和 PrimeKG 中提取 KG 数据，将其作为生成“查询-回答”数据的基础知识库。通过手动选择 438 个常见关系，并通过从不同的随机选择的起始实体进行 K-hop 遍历，重复 N 次以获取了多样化的子图集合，用于多跳推理、事实比较和集合操作。本文应用启发式规则以确保提取的子图之间的最小交集和一致性。

（2）本文采用各种事实验证数据集中的文本知识，包括 FEVER、Climate-Fever、Health-Fever、COVID-FACT 和 SCIFACT，以构建本文基准数据集 FactCHD 中的数据。本文仅选择具有相应证据的 factual 和 non-factual 样本，并最初使用 ChatGPT 直接评估这些数据集中的 claim，并选择模型难以提供错误答案的样本。

我们精心设计了有效的提示语，以引导 ChatGPT 生成“查询-回答”场景，这包含三个关键要素：角色描述、事实性模式和事实性展示。角色描述界定了系统的角色，并明确了生成过程的输入和目标。为了精确控制与事实性相关样本的类型和质量，我们提供了相应的事实性模式和展示的解释，以指导模型生成“查询-回答”场景。



Unveiling the Siren's Song: Towards Reliable Fact-Conflicting Hallucination Detection

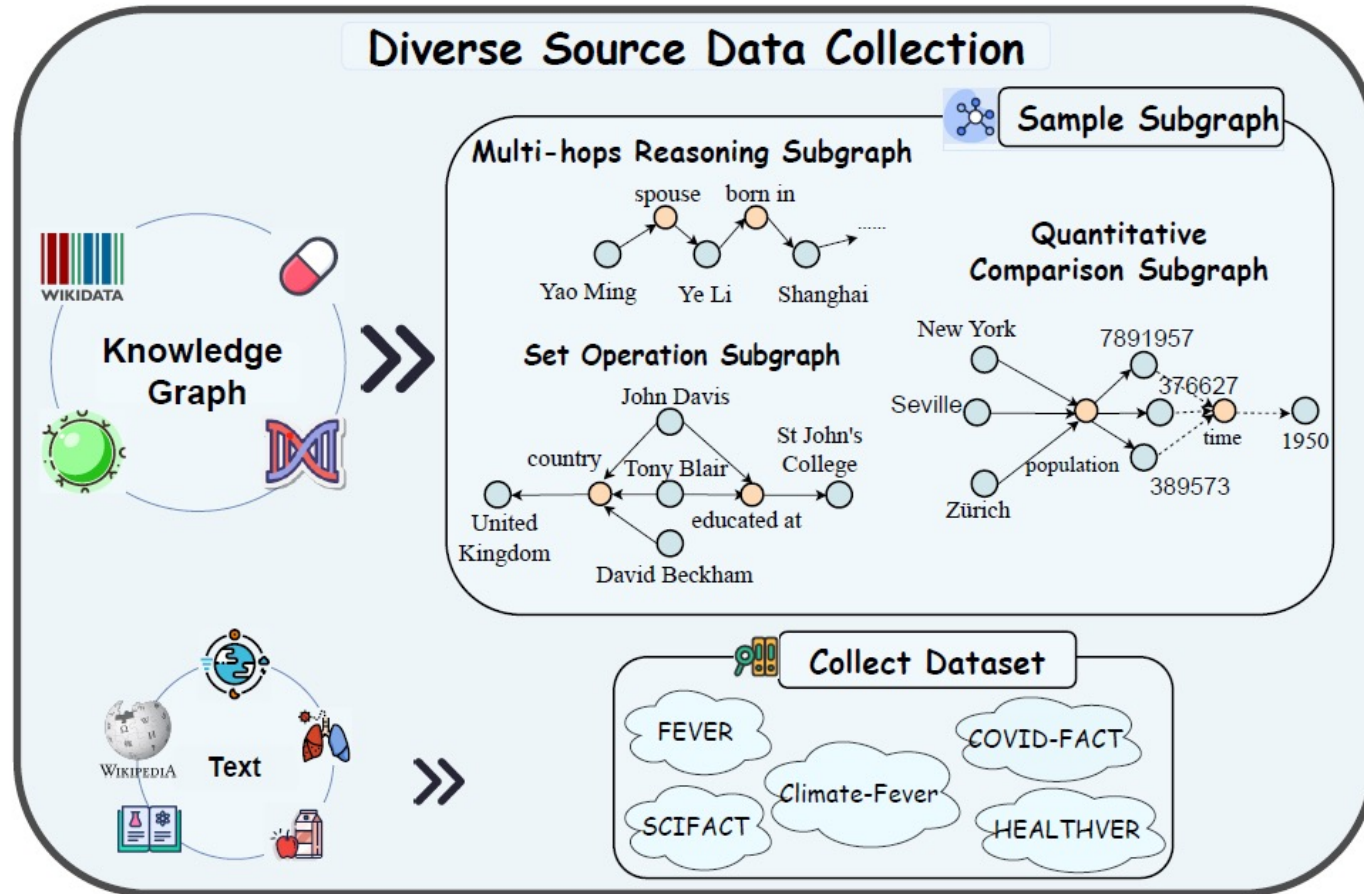


Figure3: Distribution of FactCHD across various domains

Unveiling the Siren's Song: Towards Reliable Fact-Conflicting Hallucination Detection

1.2 收集具备多个幻觉模式的真实数据

解决 LLM 输出中的幻觉问题需要对幻觉和非幻觉实例进行仔细检查。LLM 对事实冲突幻觉的敏感性源于它们受限的知识和欠优的推理能力。因此，本文将事实错误分为四种不同的模式，如图下半部分生动地展示：

（1）常规模式处理可以通过已建立的来源进行客观验证的事实陈述，通常更容易识别；

（2）多跳模式表示通过连接多个事实来得出结论的过程；

（3）比较模式涉及评估和比较不同事实之间的相对价值和关系；

（4）集合操作模式涉及操作组合元素集以分析不同事实之间的关系。随后，本文使用反映已确定的事实模式的具体指令查询开源 LLM。通过对幻觉回答进行手动注释，目标是收集真实的幻觉数据，以模拟这些现实幻觉作为示例。



Unveiling the Siren's Song: Towards Reliable Fact-Conflicting Hallucination Detection

1.3 根据一致性优化提示

本文利用 ChatGPT 生成“查询-回答”上下文，通过使用手动提示来使其充当角色扮演代理，这些提示是基于特定的示例。鉴于 ChatGPT 对提示的敏感性，本文进行了迭代的提示优化，以确保生成的数据与期望的真实模式密切匹配，从最初的固定 100 个示例开始，本文选择了五个样本进行人工相似性评估，并将其与示例进行对比。通过多数规则，本文评估每个上下文是否符合既定的相似性标准。这个迭代过程允许不断调整提示，确保生成的数据模式与期望的目标保持一致。

1.4 自动筛选以增强数据多样性

为了提高生成的“查询-回答”上下文的多样性，本文在自动筛选过程中采用 Sentence-BERT 计算上下文内的平均语义相似度。这有助于识别和排除高度相似的样本，保障数据集的多样性。在筛选过程中，总共删除了来自训练集的 1,538 个样本和来自测试集的 632 个样本，确保最终的样本集合是多样的。通过仔细消除语义上类似的条目，增强了基准数据集在评估多样的事实冲突幻觉实例方面的实用性。



Unveiling the Siren's Song: Towards Reliable Fact-Conflicting Hallucination Detection

1.5 生成证据链

本文的基准测试不仅仅是识别 LLM 生成的“查询-回答”上下文中的事实错误；它还需要生成连贯的证据链来支持这些判断，所有这些判断都根植于事实知识并由 ChatGPT 表达。

特别地，本文利用两种类型的知识源作为本文的事实基础。ChatGPT 在为“查询-回答”上下文分配适当标签时，可以使用子图事实或文本事实来提供全面的证明。值得注意的是，这些解释的有效性深刻影响着判断的可信度和可靠性，从而增强了对预测模型的信任。这些解释输出还可以增强用户的整体理解能力。

1.6 通过人工审查进行低质量数据过滤

本文设计了三个方面的过滤规则：模式一致性、回答的事实性和证据链的逻辑，利用这些规则指导标注员进行质量过滤。本文的团队由 21 名标注员组成，每个标注员都拥有熟练的英语阅读能力和本科以上的学历，并按照标准化的注释指南接受统一培训。标注员根据自己的意识，并在必要时利用搜索引擎，确保做出明智的过滤决策，排除违反既定规则的样本。

鉴于通过这些规则定义匹配质量存在固有的主观性，本文将标注员和样本分成了七组，每组三名标注员通过投票机制对同一批数据进行审查，标注员同时判断不匹配的样本被舍弃。



Unveiling the Siren’s Song: Towards Reliable Fact-Conflicting Hallucination Detection

| Evaluator | | VANILLA | | MULTI-HOPS | | COMPARISON | | SET-OPERATION | | AVERAGE | |
|--------------|----------------------------|---------|-------|------------|-------|------------|-------|---------------|-------|--------------|--------------|
| | | CLS. | EXP. | CLS. | EXP. | CLS. | EXP. | CLS. | EXP. | CLS. | EXP. |
| Zero-Shot | GPT-3.5-turbo | 55.12 | 22.79 | 59.54 | 29.84 | 16.66 | 18.89 | 55.46 | 28.23 | 52.82 | 24.03 |
| | text-davinci-003 | 52.06 | 17.72 | 59.92 | 25.30 | 25.50 | 16.09 | 48.58 | 25.71 | 50.98 | 19.57 |
| | Alpaca-7B | 29.66 | 11.72 | 5.20 | 25.60 | 8.88 | 17.95 | 13.08 | 21.37 | 23.10 | 13.66 |
| | Vicuna-7B | 35.26 | 24.62 | 17.54 | 34.39 | 9.34 | 24.88 | 14.96 | 31.41 | 28.84 | 26.84 |
| | Llama2-7B-chat | 3.57 | 26.78 | 5.49 | 33.87 | 10.53 | 35.25 | 12.61 | 33.27 | 5.77 | 29.41 |
| ICL (4-shot) | GPT-3.5-turbo | 62.02 | 37.29 | 65.66 | 51.85 | 32.2 | 48.11 | 64.74 | 50.14 | ↑8.22 61.04 | ↑17.93 41.96 |
| | text-davinci-003 | 56.52 | 39.36 | 55.02 | 58.22 | 8.50 | 48.53 | 50.34 | 51.82 | ↑1.9 52.88 | ↑24.88 44.45 |
| | Alpaca-7B | 35.82 | 31.01 | 18.12 | 40.16 | 8.86 | 29.28 | 6.70 | 31.52 | ↑5.24 28.34 | ↑16.76 32.23 |
| | Vicuna-7B | 41.36 | 42.51 | 29.24 | 58.35 | 19.36 | 41.55 | 13.46 | 53.60 | ↑6.3 35.14 | ↑19.14 45.98 |
| | Llama2-7B-chat | 31.00 | 39.08 | 39.13 | 54.38 | 10.50 | 41.83 | 27.96 | 51.73 | ↑24.48 30.25 | ↑13.61 43.02 |
| Det. (tune) | Alpaca-7B-LoRA | 73.14 | 49.00 | 63.34 | 70.83 | 69.92 | 59.88 | 68.18 | 63.75 | ↑42.32 70.66 | ↑22.73 54.96 |
| | Vicuna-7B-LoRA | 73.52 | 48.07 | 64.72 | 71.74 | 67.34 | 62.08 | 50.36 | 66.04 | ↑34.44 69.58 | ↑9.00 54.98 |
| | Llama2-7B-chat-LoRA | 77.41 | 47.91 | 67.70 | 67.30 | 62.27 | 57.03 | 78.68 | 65.94 | ↑44.48 74.73 | ↑10.69 53.71 |
| Knowledge | Alpaca-7B-LoRA (wiki) | 73.86 | 49.44 | 67.3 | 69.97 | 68.24 | 60.25 | 67.38 | 63.00 | ↑0.66 71.32 | ↑0.11 55.07 |
| | Vicuna-7B-LoRA (wiki) | 75.14 | 49.56 | 65.46 | 72.71 | 65.10 | 63.51 | 55.42 | 66.65 | ↑1.28 70.86 | ↑1.30 56.28 |
| | Llama2-7B-chat-LoRA (wiki) | 77.14 | 46.71 | 69.61 | 64.17 | 66.05 | 49.73 | 78.08 | 64.52 | ↑1.13 75.86 | ↑0.87 54.58 |
| | GPT-3.5-turbo (tool) | 69.71 | 38.60 | 69.92 | 48.43 | 44.08 | 47.26 | 74.21 | 45.65 | ↑7.59 68.63 | ↑0.41 42.37 |
| | TRUTH-TRIANGULATOR | 80.97 | 47.08 | 75.01 | 64.21 | 66.27 | 55.70 | 80.87 | 65.25 | 78.15 | 52.52 |

Unveiling the Siren's Song: Towards Reliable Fact-Conflicting Hallucination Detection

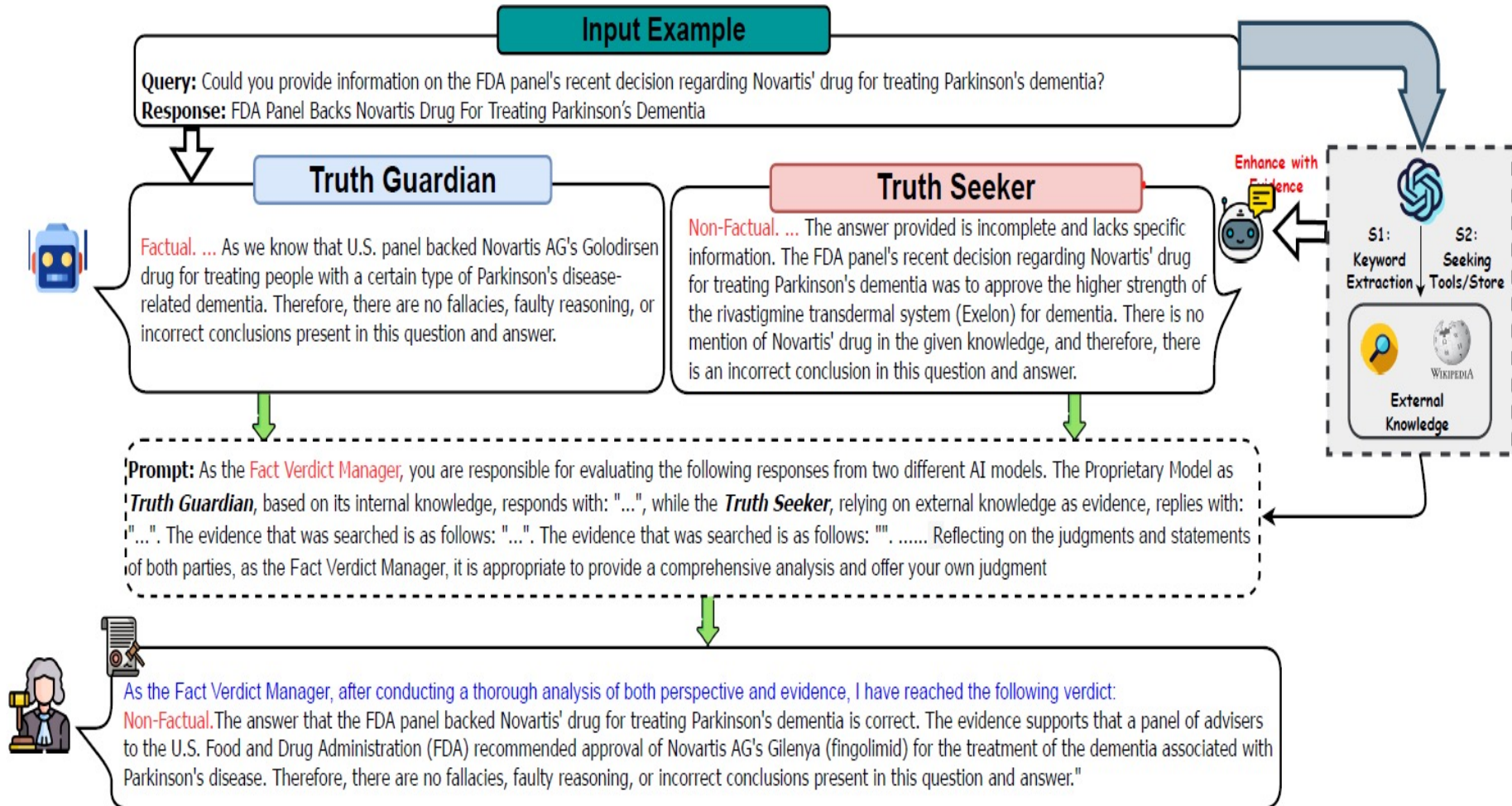
2.3 A Strong Baseline: Triangulation for Truth

如图 4 所示，我们将工具增强的 ChatGPT 称为“真相探寻者”，其目标是通过外部知识做出明智的判断。然而，外部知识源的信息可能不完整、错误或多余，可能误导大型模型。领域特定专家作为“真相守护者”则更依赖其自身知识和任务专长，倾向于更保守的预测。

为了解决这些问题，我们提出了 Truth-Triangulator 框架，灵感来自“真相三角验证”理论，通过交叉参考多个独立的源或视角来验证和确认信息。我们微调 ChatGPT 作为“事实判定管理者”，从不同角度收集证据，以提高真相或结论的可靠性和准确性。表 1 展示了我们的模型与 Llama2-7b-chat-LoRA 和 GPT-3.5-turbo（工具）相比的改进。这强调了三角验证在减少依赖单一来源或方法可能引起的错误和不一致性方面的有效性，从而促进对真相更全面和更强健的理解。



Unveiling the Siren's Song: Towards Reliable Fact-Conflicting Hallucination Detection



Unveiling the Siren’s Song: Towards Reliable Fact-Conflicting Hallucination Detection

3.3 案例分析

为了突显的广泛适用性，我们将其应用于超越 FactCHD 基准范围的真实世界幻觉数据。我们通过展示来自我们的超出分布案例分析的发现，阐明了我们模型的能力和限制，具体结果见表表 3。这些案例的结果证实了我们的方法可以做出熟练的判断，特别是在检测特定专家和工具增强型 ChatGPT 之间存在差异的情况下。这通过在真实的场景中实际应用进行了可靠性验证，增加了对于存在事实冲突的幻觉检测的可信度。。

| User Contexts and Returned Evidence | Detection Opinion from Both Sides | Prediction with Triangulation |
|--|---|--|
| <p>QUERY: Who was the mother of Afonso II, the third king of Portugal?</p> <p>RESPONSE:The mother of Afonso II, the third king of Portugal, was Queen Urraca of Castile.</p> <p>Correct Answer: Dulce Berenguer of Barcelona.</p> <p>Returned by the tool: 1.—died March 25, 1223, Coimbra), the third king of Portugal (1211–23), under whom the reconquest of the south from the Muslims was continued. born: 1185?, Coimbra, Port. Afonso II was the son of King Sancho I and Queen Dulcia, daughter of Ramón Berenguer IV of Barcelona. 2...</p> | <p>TRUTH GUARDIAN:FACTUAL. The answer that Queen Urraca of Castile was the mother of Afonso II, the third king of Portugal, is correct. This is supported by the evidence that Afonso II was the third king of Portugal and that Queen Urraca of Castile was his mother. Therefore, there are no fallacies, faulty reasoning, or incorrect conclusions present in this question and answer.</p> <p>TRUTH SEEKER: NON-FACTUAL. The answer stating that the mother of Afonso II, the third king of Portugal, was Queen Urraca of Castile is incorrect. According to the provided knowledge, the mother of Afonso II was Queen Dulcia, daughter of Ramón Berenguer IV of Barcelona. Therefore, there is an incorrect conclusion in this question and answer.</p> | <p>TRUTH-TRIANGULATOR: As the Fact Verdict Manager, after conducting a thorough analysis of both perspectives and evidence, I have reached the following verdict:</p> <p>NON-FACTUAL. The answer that Queen Urraca of Castile was the mother of Afonso II, the third king of Portugal, is incorrect. The evidence provided shows that Afonso II was the son of King Sancho I and Queen Dulcia, daughter of Ramón Berenguer IV of Barcelona. Therefore, the answer contains false assumptions and incorrect conclusions. The Truth Seeker’s response is supported by the provided evidence, while the Proprietary Model’s response is not accurate as it did not consider the external evidence.</p> |
| <p>QUERY: Who is the Jeff Probst Show-nominated television producer’s first wife?</p> <p>RESPONSE:His first wife was psychotherapist Shelley Wright.</p> <p>Correct Answer: Shelley Wright.</p> <p>Returned by the tool:1.Probst was married to his first wife, psychotherapist Shelley Wright, from 1996 to 2001. In 2004, after filming on Survivor: Vanuatu wrapped, he began dating one of the contestants, a 23 year old Julie Berry - he was about 20 years her senior. They broke up in early 2008. 2. ...</p> | <p>TRUTH GUARDIAN: FACTUAL. The answer that Jeff Probst’s first wife was psychotherapist Shelley Wright is correct. According to the evidence provided, Jeff Probst was previously married to Shelley Wright, who is a *psychotherapist* in addition to being a television producer. Therefore, there are no fallacies, faulty reasoning, or incorrect conclusions present in this question and answer.</p> <p>TRUTH SEEKER: FACTUAL. According to the provided knowledge, Jeff Probst’s first wife was psychotherapist Shelley Wright. Therefore, the answer is correct and there are no fallacies, faulty reasoning, or incorrect conclusions present in this question and answer.</p> | <p>TRUTH-TRIANGULATOR: As the Fact Verdict Manager, after evaluating the responses from both AI models and examining the provided evidence, I have reached the following verdict:</p> <p>FACTUAL. The answer that Jeff Probst’s first wife was psychotherapist Shelley Wright is correct. Both AI models correctly identified that there are no fallacies, faulty reasoning, or incorrect conclusions present in the question and answer. The evidence provided confirms that Jeff Probst was married to Shelley Wright from 1996 to 2001, and that she is a psychotherapist in addition to being a television producer. Therefore, the answer is factual and accurate.</p> |

Can We Edit Multimodal Large Language Models?

世界不但是文字的世界，我们生活的世界本质上是多模态的。我们需要不断处理和整合来自不同模态的信息，如视觉信息、听觉信息等，利用这些信息我们可以更好地与周围环境互动，提升认知世界的能力。

随着OpenAI开放了ChatGPT的多模态功能，多模态大语言模型再一次成为了焦点。现有的主流多模态大语言模型框架是利用训练好的大语言模型和图像编码器，用一个图文特征对齐模块进行连接，从而让语言模型能够理解图像特征并进行更深层的问答推理。代表作有VisualGLM、BLIP 2和MiniGPT-4等。



Can We Edit Multimodal Large Language Models?

但是目前复杂的多模态大语言模型都面临一个重大的挑战：对象幻觉（Object Hallucination）。就算是高质量的多模态语言模型，比如InstructBLIP，也存在高幻觉的文本率。多模态模型幻觉的主要原因可能有两点：

1. 多模态指令微调过程导致LVLMs 更容易在多模态指令数据集中频繁出现/共现的物体上产生幻觉；
2. 一些幻觉继承于原先的LLMs，由于使用的LLMs本来就存在一些错误/谬误知识，导致多模态语言模型也继承了这些错误知识，从而出现幻觉。



Can We Edit Multimodal Large Language Models?

最近随着一种可以精确修改模型中特定知识的范式出现，对解决模型幻觉问题提供了一个新的可行性思路，这种方法被称作模型编辑。模型编辑可以在不重新训练模型的基础上，去修改模型的参数知识，这可以节约大量的资源。但是现有的模型编辑技术大部分都是针对单模态的，那多模态的模型是否是可编辑的呢？

本文就是去探究编辑多模态大语言模型的可行性，作者构建了多模态语言模型知识编辑场景的benchmark，即设计了多模态模型编辑的指标和构建了相关数据集。并类比人类视觉问答场景，提出了编辑多模态语言模型的两种方式。其中多模态模型编辑的展示如下图所示：



Can We Edit Multimodal Large Language Models?

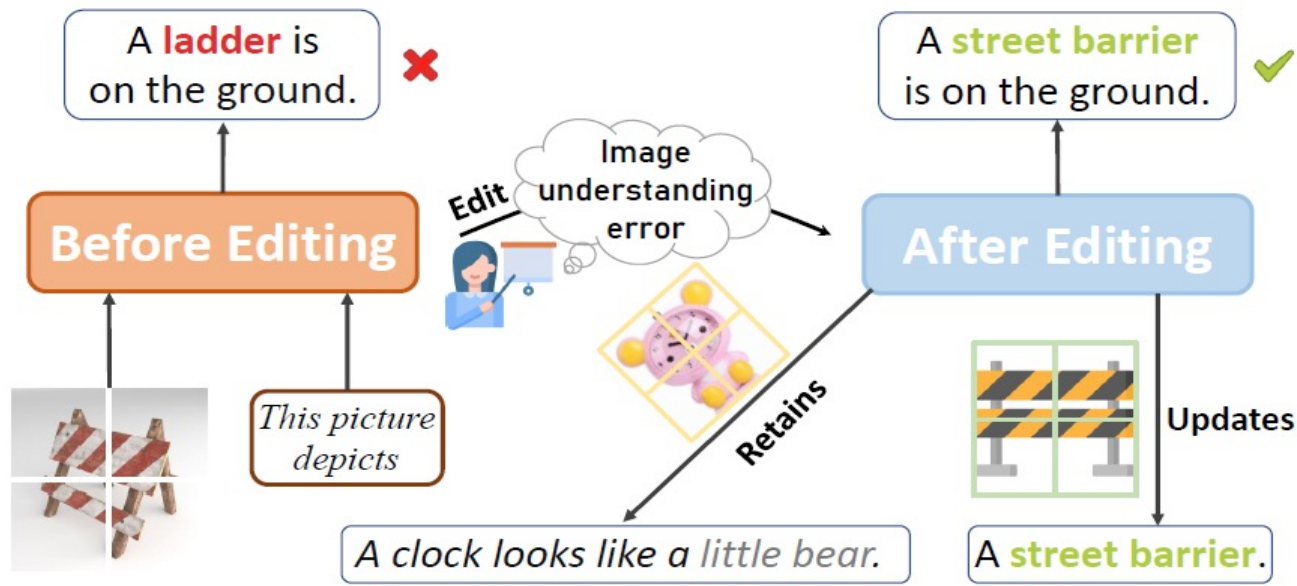


Figure 1: Overview of the **multimodal model editing** task. The editing target is to update the model's understanding of the edited input (e.g., image or text), while ensuring its interpretation of unrelated inputs remains as consistent as possible.

Can We Edit Multimodal Large Language Models?

不同于单模态模型编辑，多模态模型编辑需要考虑更多的模态信息。文章出发点依然从单模态模型编辑入手，将单模态模型编辑拓展到多模态模型编辑，主要从以下三个方面：可靠性（Reliability），稳定性（Locality）和泛化性（Generality）。

可靠性：模型编辑需要能够保证正确修改模型的知识，可靠性就是衡量编辑后模型的准确率。多模态模型编辑亦是如此。

稳定性：稳定性是判别模型编辑影响模型其余知识的程度。模型编辑希望在编辑完相关知识过后，不影响模型中其余的一些知识。多模态模型编辑与单模态不同，由于我们需要编辑多个模型区域，所以我们需要判断多模态模型进行编辑之后到底是对哪部分产生的影响多，哪部分少。所以作者提出了两种稳定性测试：T-Locality和M-Locality，一个测试纯语言模型的稳定性，一个测试多模态整体模型的稳定性。

泛化性：编辑需要对一定编辑范围内的数据都要具有编辑效应，单模态模型编辑泛化性只考虑一种数据形式，即同义语义集合。多模态模型需要考虑更多模态数据，VLMs多增加了一个图片模态数据。



Can We Edit Multimodal Large Language Models?

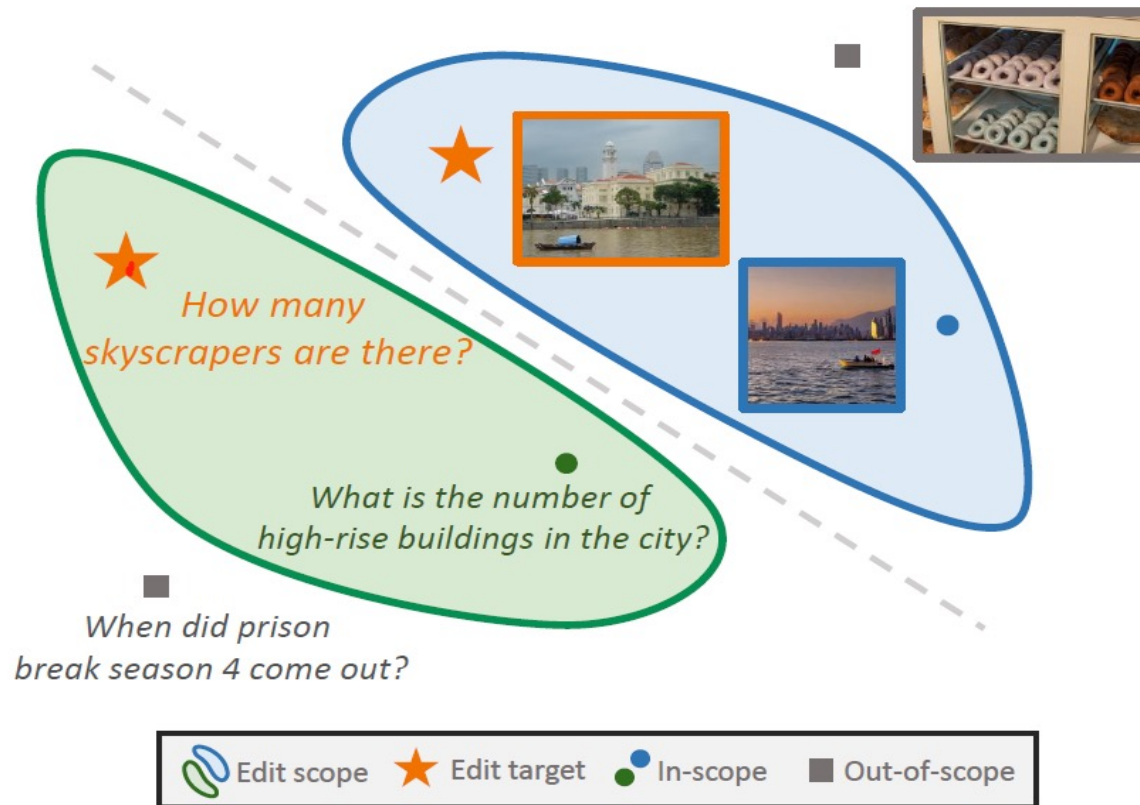
数据集

对于上述所有评估指标，本文作者都构造了对应的数据集来进行测试，其中针对可靠性数据集，作者收集了现有多模态大语言模型表现不佳的任务数据来作为编辑对象数据集，本文采用两个不同的多模态任务分别是VQA和Image Caption。并设计两种任务编辑数据集E-VQA和E-IC。

对于泛化性数据，多模态模型由于本身的数据也是多模态的，所以需要考虑更多模态的泛化数据情况。其中多模态泛化性数据例子如下：



Can We Edit Multimodal Large Language Models?



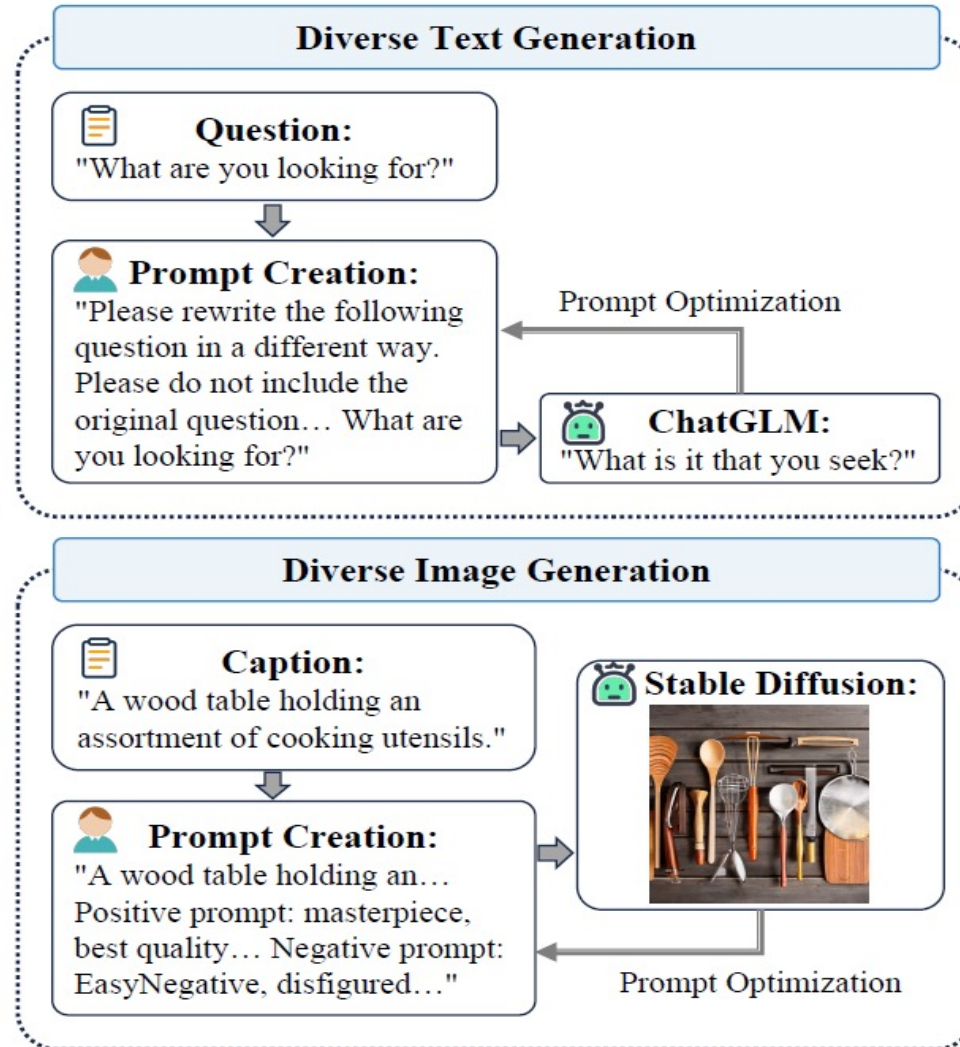
Can We Edit Multimodal Large Language Models?

对文本数据，本文作者利用不同的方法构造相关泛化数据集。首先对于VQA数据，文章作者使用ChatGLM去生成文本类的泛化数据集，通过构造相关的prompt，让对话模型吐出相似句子。Image Caption任务由于其本身的数据比较简单，生成效果并不佳，所以作者人工构建了几十条相似文本数据，然后通过随机替换的方式作为Image Caption任务的泛化数据集。

然后对于图片数据，作者利用COCO数据集中提供的图片描述。通过现有效果非常不错的图片生成模型Stable Diffusion 2.1来生成与图片描述相似的图片。具体构造流程如下图所示：



Can We Edit Multimodal Large Language Models?



Can We Edit Multimodal Large Language Models?

对于稳定性数据集，作者为了考量编辑不同区域对模型的影响，所以将稳定性数据分为了Text Stability测试数据和Vision Stability测试数据。这部分数据不用构造，作者直接使用了之前的已有数据集。对于文本，沿用MEND中的NQ数据集，对于多模态数据，文章使用了多模态中比较简单的问答数据集OK-VQA作为测试数据集。

最后数据集统计如下：

| TASK | Train | Test | L-Locality | M-Locality |
|-------|-------|-------|------------|------------|
| E-VQA | 6,346 | 2,093 | 4,289 | 5,046 |
| E-IC | 2,849 | 1,000 | 4,289 | 5,046 |

Can We Edit Multimodal Large Language Models?

多模态模型编辑

对于如何去编辑多模态语言模型，文章类比人类视觉问答场景出错场景，来设计多模态模型编辑实验。以VQA任务为例子，人类在做VQA题目时有两种出错的可能：

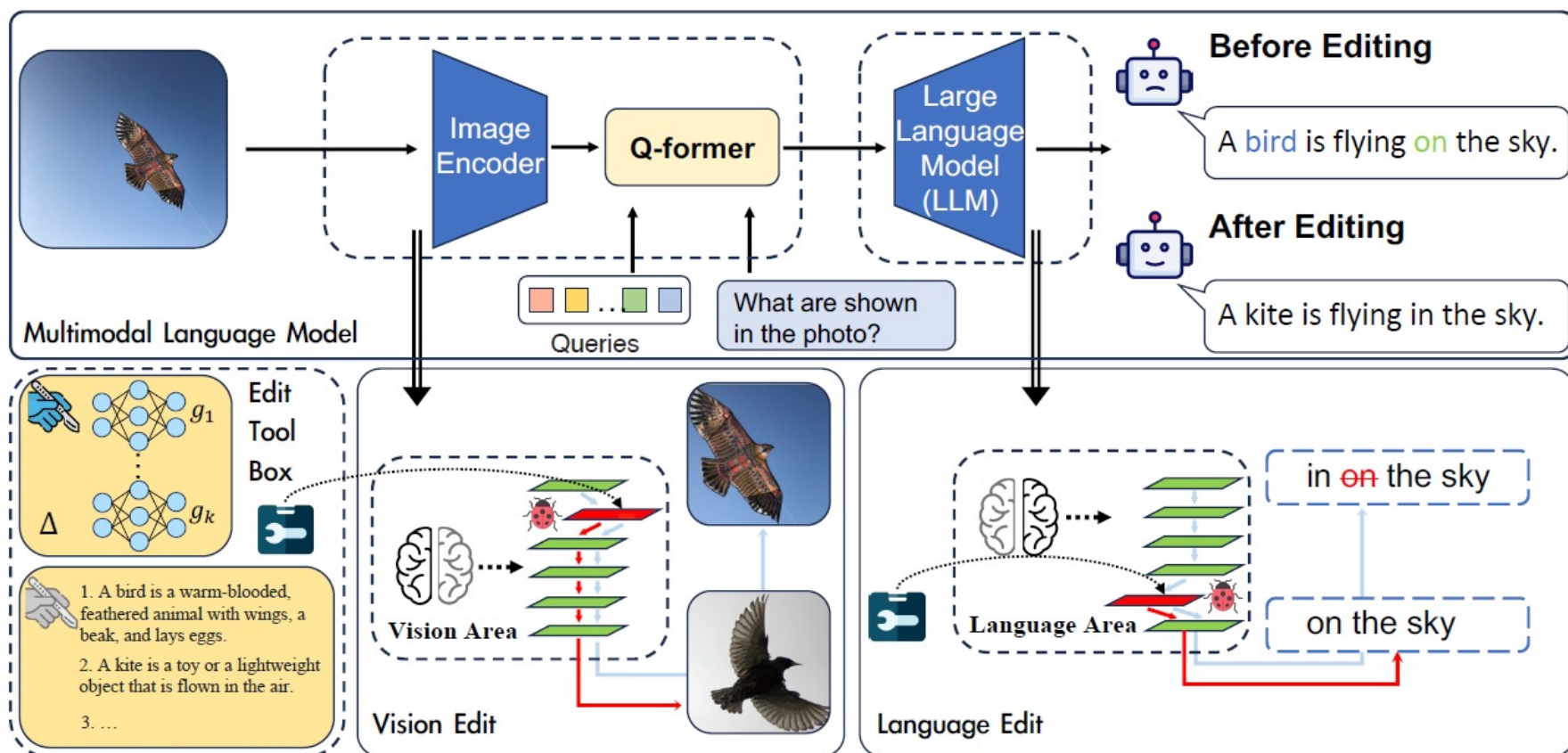
视觉出错：人类可能在图片识别这个阶段就出错，可能是看错，也有可能是视觉细胞本身就存在问题。例如人类色盲患者没有办法正确识别图片的颜色特征，就会在颜色识别的任务上出错。针对这个，文章作者提出了Vision Edit，针对VLMs的视觉模块进行编辑。

知识出错：人类可能正确识别了图片中的关键特征，但是本身的知识库里却没有相关特征的知识，这就导致人犯“指鹿为马”的失误。针对这个问题，作者提出了Language Edit，由于多模态语言模型的知识库都来自于LLMs，所以这部分编辑也就是针对语言模型。



Can We Edit Multimodal Large Language Models?

多模态模型编辑的主要流程图作如下图所示：








Can We Edit Multimodal Large Language Models?

| | | EDITING VQA | | | | EDITING IMAGE CAPTION | | | |
|---------------|--------------------|---------------|----------------|--------------|--------------|-----------------------|----------------|--------------|--------------|
| Method | | Reliability ↑ | T-Generality ↑ | T-Locality ↑ | M-Locality ↑ | Reliability ↑ | T-Generality ↑ | T-Locality ↑ | M-Locality ↑ |
| BLIP-2 OPT | | | | | | | | | |
| Size: 3.8B | | | | | | | | | |
| Base Methods | Base Model | 0.00 | 0.00 | 100.0 | 100.0 | 0.00 | 0.00 | 100.0 | 100.0 |
| | FT (vision block) | 56.28 | 29.88 | 100.0 | 11.32 | 0.08 | 0.00 | 100.0 | 7.31 |
| | FT (last layer) | 58.70 | 15.33 | 78.86 | 2.86 | 0.24 | 0.10 | 67.67 | 3.91 |
| Model Editing | Knowledge Editor | 67.80 | 63.00 | 97.32 | 45.89 | 69.00 | 62.80 | 96.21 | 45.55 |
| | In-Context Editing | 99.95 | 91.59 | 13.16 | 1.88 | 96.70 | 78.20 | 13.36 | 2.17 |
| | SERAC | 91.20 | 91.40 | 100.0 | 0.33 | 94.40 | 96.00 | 100.0 | 0.47 |
| | MEND | 92.60 | 90.80 | 96.07 | 65.15 | 65.00 | 38.00 | 92.67 | 55.72 |
| MiniGPT-4 | | | | | | | | | |
| Size: 7.3B | | | | | | | | | |
| Base Methods | Base Model | 0.00 | 0.00 | 100.0 | 100.0 | 0.00 | 0.00 | 100.0 | 100.0 |
| | FT (vision block) | 39.58 | 0.98 | 100.0 | 3.96 | 0.63 | 0.00 | 100.0 | 5.13 |
| | FT (last layer) | 39.57 | 0.58 | 72.01 | 16.42 | 2.75 | 0.00 | 35.52 | 9.28 |
| Model Editing | Knowledge Editor | 87.77 | 86.62 | 97.15 | 55.77 | 35.10 | 24.20 | 96.78 | 52.22 |
| | In-Context Editing | 71.72 | 40.23 | 13.46 | 2.00 | 68.60 | 59.80 | 12.51 | 2.96 |
| | SERAC | 87.20 | 84.60 | 100.0 | 0.33 | 40.20 | 36.60 | 100.0 | 0.97 |
| | MEND | 95.51 | 95.27 | 98.73 | 71.33 | 87.10 | 84.10 | 98.34 | 59.53 |

可以看到微调的效果都比较一般，而且会对于模型中的其他知识造成灾难性遗忘。模型编辑在可靠性上表现的都还不错，并且对于模型的稳定性也维持的比较好，不会造成模型的过拟合和灾难性遗忘

Can We Edit Multimodal Large Language Models?

作者觉得这可能和模型的架构有关，编辑语言模型部分可以直接影响模型的输出，而编辑视觉部分只能影响模型输入。而且大部分的知识都是保存在 LLMs 中的，所以编辑视觉模块的效果不佳。最后展示几组编辑case：

| | | |
|--|---|---|
| <p>Before Editing</p>  <p>What is the man doing?</p> <p>Boarding.</p> | <p>Before Editing</p>  <p>What are shown in the photo?</p> <p>A photo getting on a bus that has bicycles on the rack.</p> | <p>Before Editing</p>  <p>What is the train number?</p> <p>17788.</p> |
| <p>After Editing</p>  <p>What is the man doing?</p> <p>Skateboarding.</p> | <p>After Editing</p>  <p>What are shown in the photo?</p> <p>A person getting on a bus that has bicycles on the rack.</p> | <p>After Editing</p>  <p>What is the train number?</p> <p>18688.</p> |

Case of successful VQA editing (By SERAC)

Case of successful Image Caption editing (By SERAC)

Case of failure VQA editing (By IKE)

Can We Edit Multimodal Large Language Models?

总结：

多模态模型是非常重要的领域，如何解决目前面临的幻觉问题是非常关键的问题。模型编辑技术为解决模型幻觉提供了一个不错的思路，但是在多模态模型上依然有许多不足的地方，比如如何能够更有效地进行不同模态之间的协同编辑？如何解决编辑OOD数据？如何做到多模态的连续编辑？这些都是未来值得探讨的方向。



Thank you

