Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

AI的幻觉陷阱与创造力

硕士研究生 刘佳

2025年6月8日

问题回溯

- 总结反思
  - 语速较快
  - 浅层的部分讲解过多，论文算法细节深入不足

- 相关内容
  - 杨宗源《文本生成中的幻觉》——2023.08.20
  - 张凌浩《基于图结构处理的文本生成》——2022.02.27
  - 高依萌《预训练语言模型GPT3》——2021.02.07

内容提要

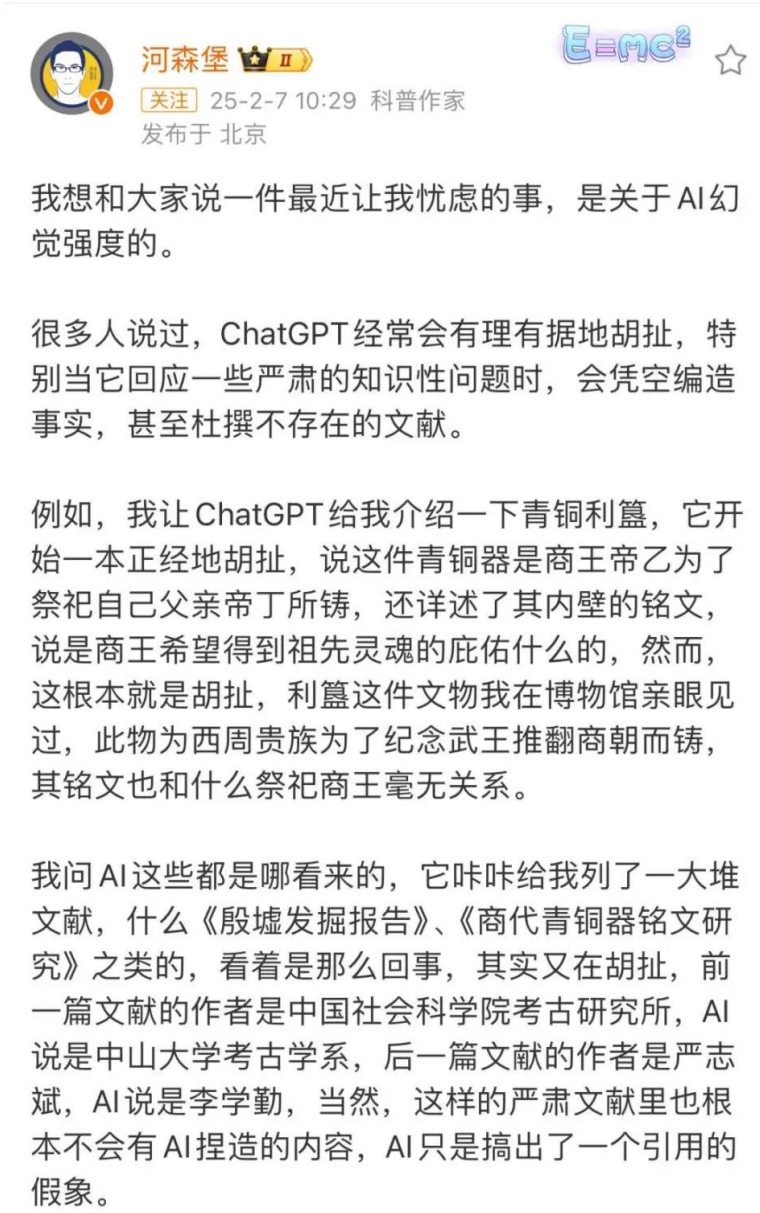
- 预期收获
- 案例引入
- 题目内涵解析
- 研究背景与意义
- 研究历史与现状
- 相关知识
- 算法原理
  - LLM-Check
  - VE
- 特点总结与工作展望
- 参考文献

预期收获

- 预期收获
  - 1. 什么是AI幻觉及大模型为什么会产生幻觉
  - 2. 现有AI幻觉检测的方法
  - 3. 如何减缓AI幻觉
  - 4. AI幻觉的创造力价值

案例引入

- 律师使用ChatGPT虚构案例
  - “马塔诉阿维安卡公司”案：
    - 乘客在飞行途中受伤，起诉航空公司索赔
    - 因无法访问联邦判例数据库，律师施瓦茨使用 ChatGPT 搜寻案例
    - 所引用的6个案例为伪造案例，内容不真实
  - 行业应对情况：
    - 各州法院和律师协会发布相关指导
    - 英国大律师公会发布使用规范
    - 使用GAI撰写法律文书必须人工验证，不得替代专业判断



河森堡 II
关注 25-2-7 10:29 科普作家
发布于 北京

我想和大家说一件最近让我忧虑的事，是关于AI幻觉强度的。

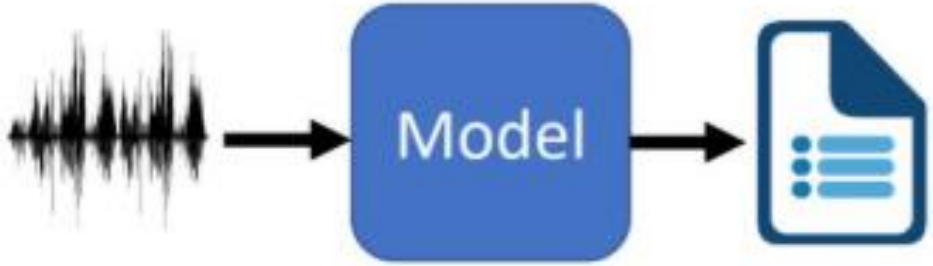
很多人说过，ChatGPT经常会有理有据地胡扯，特别当它回应一些严肃的知识性问题时，会凭空编造事实，甚至杜撰不存在的文献。

例如，我让ChatGPT给我介绍一下青铜利簋，它开始一本正经地胡扯，说这件青铜器是商王帝乙为了祭祀自己父亲帝丁所铸，还详述了其内壁的铭文，说是商王希望得到祖先灵魂的庇佑什么的，然而，这根本就是胡扯，利簋这件文物我在博物馆亲眼见过，此物为西周贵族为了纪念武王推翻商朝而铸，其铭文也和什么祭祀商王毫无关系。

我问AI这些都是哪看来的，它咔咔给我列了一大堆文献，什么《殷墟发掘报告》、《商代青铜器铭文研究》之类的，看着是那么回事，其实又在胡扯，前一篇文献的作者是中国社会科学院考古研究所，AI说是中山大学考古学系，后一篇文献的作者是严志斌，AI说是李学勤，当然，这样的严肃文献里也根本不会有AI捏造的内容，AI只是搞出了一个引用的假象。

案例引入

- **Whisper：OpenAI的自动语音识别（ASR）系统**
  - **行业应用：医疗系统中，将患者与医生的对话问诊过程音频，转写为文字病例，有超过30000名临床医生和40个医疗系统使用**
  - **发现：100多个小时的 Whisper 转录样本，其中约有一半内容存在幻觉**
    - **原音频：“嗯，她的父亲再婚后不久就去世了”**
    - **转录文本：“没关系。只是太敏感了，不方便透露。她确实在65岁时去世了”**
  - **结果：2.6W多份自动转录病例中，几乎每本都存在瞎编和幻觉问题，对患者健康和医疗系统产生严重负面影响**



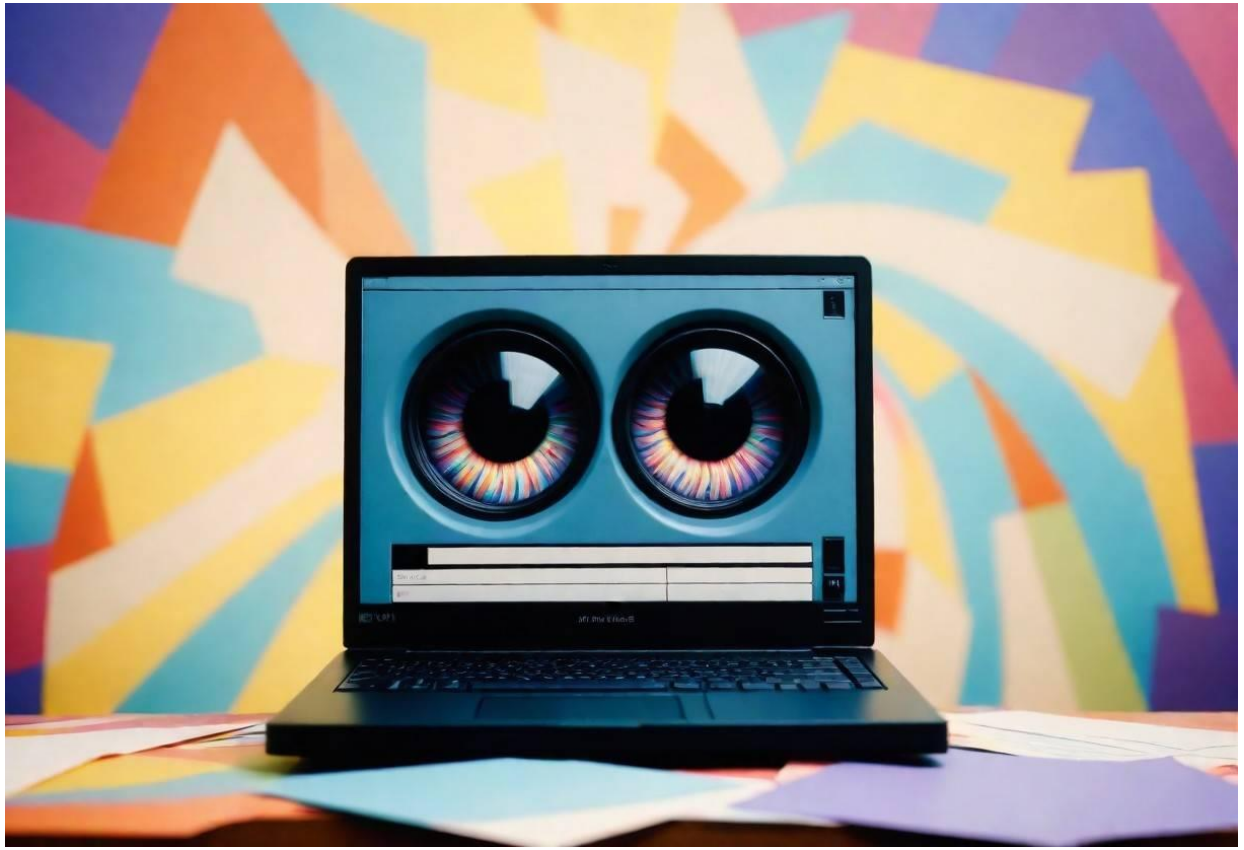
题目内涵解析 什么是AI幻觉？

- 人工智能幻觉（Artificial Intelligence Hallucination）
  - 指模型生成与事实不符、逻辑断裂或脱离上下文的内容，本质是统计概率驱动的“合理猜测”，即一本正经地胡说八道
  - 两个分类
    - 事实性幻觉：指模型生成的内容与可验证的现实世界事实不一致
    - 忠实性幻觉：指模型生成的内容与用户的指令或上下文不一致
    - 提问：糖尿病患者可以通过吃蜂蜜代替糖吗？

	回答	分析
事实性幻觉	是的，蜂蜜是天然的，可以帮助糖尿病患者稳定血糖水平。	错误：蜂蜜虽然是天然食品，但仍然含有大量果糖和葡萄糖，会升高血糖水平，不适合糖尿病患者代替糖使用。
忠实性幻觉	蜂蜜富含维生素和矿物质，对提高免疫力很有帮助，因此是一种健康的食品。	偏题：回答内容虽无事实错误，但与提问“糖尿病患者是否可以用蜂蜜代替糖”无关，未忠实于用户意图。

研究背景与意义

- 研究背景
  - 大模型生成内容质量难以保证
  - 应用场景对准确性要求高
  - 现有检测与缓解机制尚不成熟
- 研究意义
  - 保障AI系统的可信性与安全性
  - 推动AI内容生成规范化发展
  - 支撑AI监管、溯源与审计机制建设
  - 拓展多模态幻觉研究新方向



维护信息环境的安全、可信和公正

研究历史与现状

2022

Ji et al.提出经典的幻觉二分类框架，将幻觉分为输出与输入直接矛盾和输出无法验证两类，主要应用于摘要、对话、翻译等任务，为后续LLM幻觉研究打下分类的基础，是幻觉研究的开端

2022

Lin et al.提出TruthfulQA框架，构建对抗性问答数据集，诱导模型复述训练数据中的虚假但常见内容；首次提出提出模仿幻觉概念，设计人工+LLM联合评判标准揭示了大模型训练中“记错内容→生成错误”的隐性风险

2023

Manakul et al.提出多样性检测范式SelfCheckGPT，使用回答之间的自洽度判断是否存在幻觉，提出 “无需外部知识”的检测方式，开创后续基于行为一致性检测方法的思路

2023

Chern et al. 将LLM与工具链（搜索、维基百科等）组合，进行三阶段流程：事实检测→ 工具辅助→ 分数计算，奠定了LLM+检索工具（RAG-like）架构下的外部验证机制

2023

Li et al. 提出HaluEval系列统一检测基准，涵盖医疗、金融、教育等5个领域，提出事实错误、逻辑矛盾等分类型幻觉标签，支持多种模型评估、对比多种检测策略；建立了高质量、多样化、结构化的幻觉检测评估平台

2024

Luo et al.提出轻量级幻觉检测新范式LLM-Check，做到高效推理，使得推理开销<1秒/样本，且支持白盒与黑盒模型，开创了基于LLM内部信号+单轮推理的新检测范式

- 事实性检测
  - 事实检查
    - 事实提取
    - 事实验证
  - 不确定性估计
    - 基于LLM内部状态
    - LLM行为

- 忠实性检测
  - 基于事实的检测
  - 基于分类器的检测
  - 基于QA的检测
  - 基于不确定性的检测
  - 基于LLM的检测

相关知识

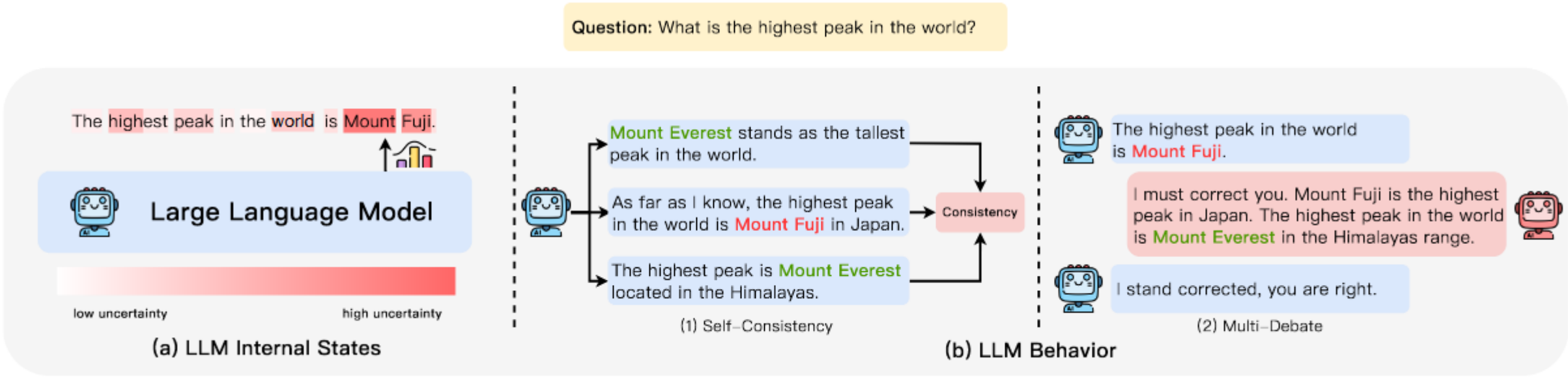
- AI为什么会产生幻觉
  - 数据驱动型幻觉：垃圾进，垃圾出
    - 医疗问答模型根据过时论文推荐禁用药物
    - 图像生成模型将“CEO”与“男性”强关联
  - 模型结构型幻觉：注意力机制的“盲区”
    - Transformer缺陷
    - 扩散模型缺陷
  - 推理链型幻觉：逻辑崩盘的“多米诺效应”
    - 数学解题：错误的第一步推导导致最终答案偏差
    - 法律分析：错误引用法条引发整套逻辑链失效

相关知识

- 推理与幻觉的关系
  - DeepSeek V3：提问 → 回答；DeepSeek R1：提问 → 思维链 → 回答
  - 推理增强 → 幻觉率降低
    - 逻辑准确性与错误减少
    - 上下文理解与信息关联
  - 推理增强 → 幻觉率提升
    - 逻辑过度外推
    - 认知置信度错位
    - 错误前提下的正确推理
- 普通用户应对幻觉的方式
  - 双AI验证/大模型协作
  - 联网搜索
  - 提示词工程

相关知识

- 应对AI幻觉的技术方案
  - RAG框架：利用检索增强生成（如先搜索权威数据库，再生成答案）
  - 外部知识库：结合外部知识库、通用知识，强化垂直领域
  - 精细训练：针对不同任务类型进行具体的微调或强化
  - 评估工具：开发高效的自动化AI幻觉识别工具，对生成内容进行及时验证



LLM-Check: Investigating Detection of Hallucinations in Large Language Models

算法原理 TIPO

T	目标	仅有单次响应、无外部知识条件下高效检测LLM生成文本中的幻觉
I	输入	用户输入的提示词、LLM生成的响应
P	处理	1.输入拼接： 将提示与模型响应拼接为完整序列 2.特征提取： 获取该序列在模型内部的隐藏状态与注意力矩阵 3.得分计算： 计算 Hidden Score、Attention Score、Perplexity 和 Logit Entropy 表征响应异常性 4.幻觉判断： 根据得分是否超出阈值判断该响应是否存在幻觉内容
O	输出	布尔值True/False是否存在幻觉

P	问题	摆脱依赖多响应或外部知识库的限制，实现低成本幻觉检测
C	条件	可访问 LLM 的隐藏状态和注意力结构 模型为自回归结构（如 Transformer） 可调用 softmax 概率输出接口（用于计算熵与困惑度）
D	难点	单响应、黑盒环境下的泛化能力、特征提取的稳定性与有效性
L	水平	NeurIPS 2024 CCF A

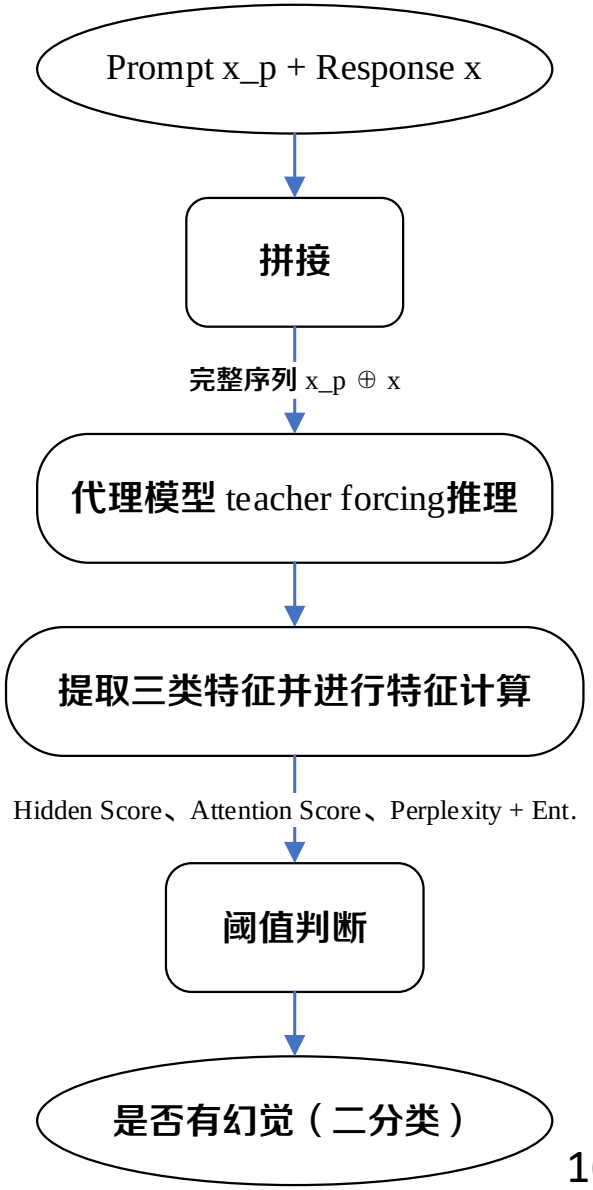
算法原理 创新说明

- LLM-Check
  - 仅依赖单个响应即可检测幻觉
    - 传统方法依赖多个生成结果之间的一致性（如 SelfCheckGPT），或依赖外部事实检索（如 RAGTruth），而 LLM-Check 实现了在无参考、无多响应条件下，仅凭单一生成结果判断是否存在幻觉
  - 融合内部表示与输出概率，构建新型得分体系
    - 首次将 Transformer 模型中的隐藏层状态、注意力矩阵和输出概率分布统一为四个检测信号，形成多维幻觉检测机制
  - 兼容黑盒环境
    - 支持幻觉检测部署于无模型权重的 SaaS 应用（如调用 GPT-4 生成的响应）

算法原理 LLM-Check

- **LLM-Check**
  - 检测目标：不借助外部数据库、不调用多次生成，仅凭单个模型输出，判断该输出中是否含有幻觉
  - 核心思想：利用LLM内部结构对幻觉的“敏感性”
    - 模型内部的表示结构、注意力机制和输出概率分布在生成内容是真实与幻觉之间存在差异
  - 四种得分方法、三类特征来源

特征来源（维度）	得分指标	描述
Hidden 表示层	Hidden Score	从隐藏状态提取的结构性表示差异
Attention 模块	Attention Score	从注意力分布提取的结构性焦点信息
Logit 输出概率	Perplexity、Windowed Entropy	从输出 token 的概率分布中推断生成的不确定性与局部不稳性



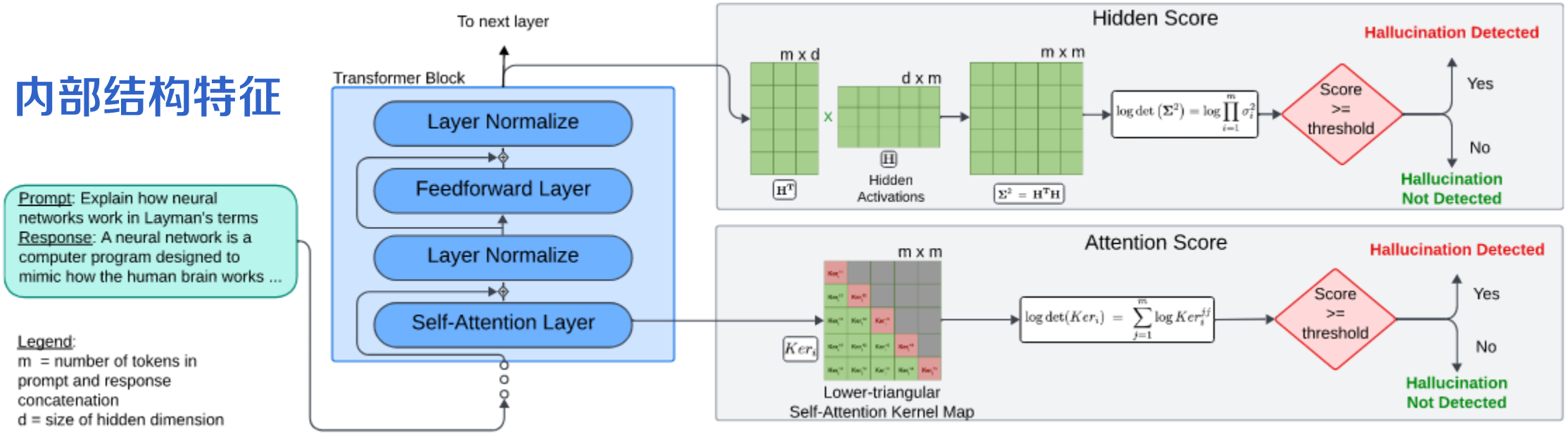
算法原理 LLM-Check

- 提取三类特征
  - Hidden Score —— 表征隐藏状态的“复杂性”
  - Attention Score —— 判断注意力模式的异常性
  - Logit-based Uncertainty —— 利用输出的不确定性估计幻觉
    - Perplexity（困惑度）
    - Windowed Logit Entropy（局部熵）

内部结构类

输出分布类

内部结构特征



算法原理 LLM-Check

- **Hidden Score：表征隐藏状态的“复杂性”**
  - 对每层隐藏状态矩阵 $H \in \mathbb{R}^{d \times m}$，构建协方差矩阵 $\sum = H^T H$
  - 计算 log-determinant 作为度量指标 $HiddenScore = \frac{2}{m} \sum_{i=1}^{m} \log \sigma_i$
    - 真实内容的隐藏表示更加分散、富含信息 → 协方差大、特征值丰富
    - 幻觉内容往往“偏离语义空间” → 表示结构异常，导致 log-det 降低
- **Attention Score：判断注意力模式的异常性**
  - 提取每层每个注意力头的下三角 attention kernel $A \in \mathbb{R}^{m \times m}$
  - 只取对角线元素并计算平均对角线 log 值 $AttenScore = \frac{1}{m} \sum_{i=1}^{m} \log A_{ii}$
    - 模型对真实生成内容有较明确的注意结构 → 注意力更集中合理
    - 幻觉区域 attention 更混乱 → 对角权重不正常，出现强自注意或异常偏移，模型可能对错误信息集中注意，或注意力稀疏无焦点

算法原理 LLM-Check

- **Logit-based Scores：输出概率不确定性**
  - **Perplexity（困惑度）**
    - $PPL(x) = \exp(-\frac{1}{m-n}\sum_{i=n+1}^{m} \log p(x_i | x_p \oplus x < i))$
    - 越高说明模型越“不确定” → 更可能是幻觉
  - **Windowed Logit Entropy（局部熵）**
    - 每个 token 输出概率的熵值 $Entropy_i = -\sum_{j=1}^{k} p_j \log p_j$ 滑动窗口选取最大局部值
    - 能够定位出局部片段中的“生成不确定性高”区域，常对应幻觉词组

特征类型	反应类型	幻觉类型
Hidden Score	表示是否异常集中	虚构事实、语义错乱
Attention Score	注意力是否异常	错误重点、无焦点描述
Perplexity	生成是否流畅	模型缺乏训练的知识领域
Window Entropy	局部是否高不确定性	短句伪造、实体错配等

多角度信号组合使得模型可以更可靠地识别各种幻觉模式

- 数据资源

数据集	来源	内容描述
FAVA-Annotation	Mishra et al., 2024	由人类标注的幻觉样本，覆盖多种类型
SelfCheckGPT Dataset	Manakul et al., 2023	WikiBio 数据，1908 条样本，每条有 20 个响应
FAVA-Train	作者构造（基于 FAVA）	在原始句子中插入人工标注幻觉 span，构造正负样本对
RAGTruth	Shi et al., 2023	来自 CNN/DailyMail，总结内容中有幻觉注释

- 对比方法

方法	不依赖模型训练	单响应	高效性	样本级	无外部知识条件
FAVA	✗	✓	✓	✓	✓
SelfCheckGPT	✓	✗	✗	✓	✓
INSIDE	✓	✗	✓	✗	✓
RAGTruth	✗	✓	✗	✓	✗
LLM-Check	✓	✓	✓	✓	✓

Model	Measure	AUROC	Accuracy	TPR @ 5% FPR	F1 Score
Llama-2-7B	Self-Prompt	50.30	50.30	-	66.53
Llama-2-7B	FAVA Model	53.29	53.29	-	43.88
Llama-2-7B	SelfCheckGPT-Prompt	50.08	54.19	-	67.24
Llama-2-7B	INSIDE	59.03	57.98	13.17	39.66
LLM-Check (Ours)					
Llama-2-7B	PPL Score	53.22	58.68	3.59	68.33
	Window Entropy	56.90	56.59	2.99	42.52
	Logit Entropy	53.80	55.99	2.99	56.73
	Hidden Score (LY 20)	58.44	58.08	11.98	59.66
	Attn Score (LY 21)	**72.34**	**67.96**	**14.97**	**69.27**
Vicuna-7B	PPL Score	53.96	56.89	3.59	64.20
	Window Entropy	55.24	58.38	5.99	66.02
	Logit Entropy	52.29	55.69	1.80	57.31
	Hidden Score (LY 15)	58.22	59.28	10.18	**66.99**
	Attn Score (LY 19)	**71.69**	**66.47**	**24.55**	62.00
Llama-3-8B	PPL Score	53.22	58.68	3.59	67.40
	Window Entropy	56.90	56.59	2.99	55.52
	Logit Entropy	53.80	55.99	2.99	56.27
	Hidden Score (LY 15)	57.10	57.78	10.78	65.38
	Attn Score (LY 23)	**68.19**	**65.87**	**15.57**	**70.53**

无外部知识条件

FAVA-Annotation 数据集

模型： GPT-3

Target Model	Measure	White-box	Black-box				
		Llama-2-7b	Llama-2-13b	Llama-2-70b	GPT-4	Mistral-7b	Overall
Hidden Score	AUROC	54.11	59.67	59.31	61.87	53.68	57.24
	Accuracy	56.33	59.66	58.42	68.52	54.15	57.62
	TPR@5%FPR	8.14	12.41	9.9	3.7	5.18	8.37
	F1 Score	61.51	50.42	66.14	67.86	32.58	47.45
Logit (Perplexity)	AUROC	53.73	52.46	56.97	52.13	52.11	53.27
	Accuracy	54.07	55.17	57.92	59.26	54.66	55.79
	TPR@5%FPR	7.69	8.97	6.93	0.00	4.15	6.01
	F1 Score	58.7	50.57	61.26	61.02	43.23	50.45
Logit (Win Entropy)	AUROC	52.08	55.71	56.38	55.83	52.61	54.58
	Accuracy	53.17	56.9	57.43	59.26	53.89	55.90
	TPR@5%FPR	4.98	15.86	1.98	7.41	10.36	10.08
	F1 Score	53.98	33.68	62.01	54.9	49.29	47.51
Logit (Log Entropy)	AUROC	53.95	51.18	55.14	50.34	50.43	51.68
	Accuracy	55.43	53.79	57.43	57.41	53.89	54.83
	TPR@5%FPR	7.24	9.66	4.95	0.00	6.22	6.65
	F1 Score	53.74	15.09	66.41	60	48.41	42.62
Attention Score	AUROC	54.19	60.05	60.01	63.51	55.37	58.30
	Accuracy	54.52	59.66	60.89	66.67	56.99	59.23
	TPR@5%FPR	5.88	14.48	12.87	7.41	5.18	9.87
	F1 Score	54.5	55.97	55.06	67.8	57.72	57.18

白盒、黑盒条件

RAGTruth 数据集

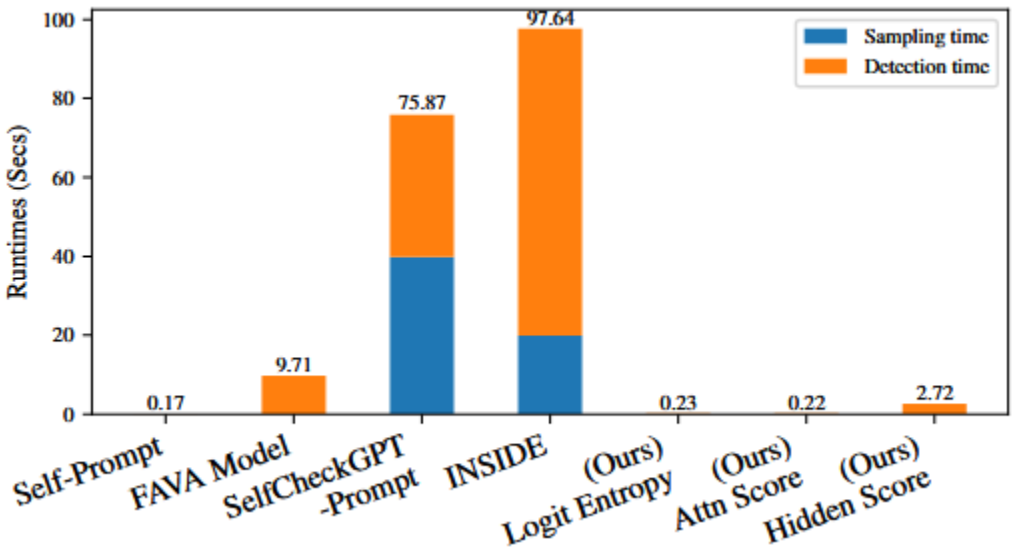
模型：Llama-2-7b

实验设计 实验结果

Model	Method	AUC-PR	Accuracy	TPR @ 5% FPR
Llama-2	SelfCheck	72.84	51.44	4.81
Llama-3	SelfCheck	75.06	54.84	5.10
LLM-Check (Ours)				
Llama-2	Attn Score	**80.04**	58.91	9.41
Llama-2	Prompt	79.46	**61.21**	8.76
Llama-3	Attn Score	79.96	58.92	**9.48**
Llama-3	Prompt	78.49	58.54	7.11

无外部参考、多响应 SelfCheckGPT数据集

Model	Measure	AUROC	Accuracy	TPR @ 5% FPR
Llama-2	PPL Score	74.20	70.00	26.00
	Window Entropy	**77.00**	**72.00**	**34.00**
	Logit Entropy	74.36	71.00	26.00
	Hidden Score	51.44	54.00	4.00
	Attn Score	69.57	66.60	11.60
Llama-3	PPL Score	73.48	68.80	13.20
	Window Entropy	78.44	72.00	**28.00**
	Logit Entropy	**79.24**	**73.60**	28.00
	Attn Score	71.91	68.20	19.60

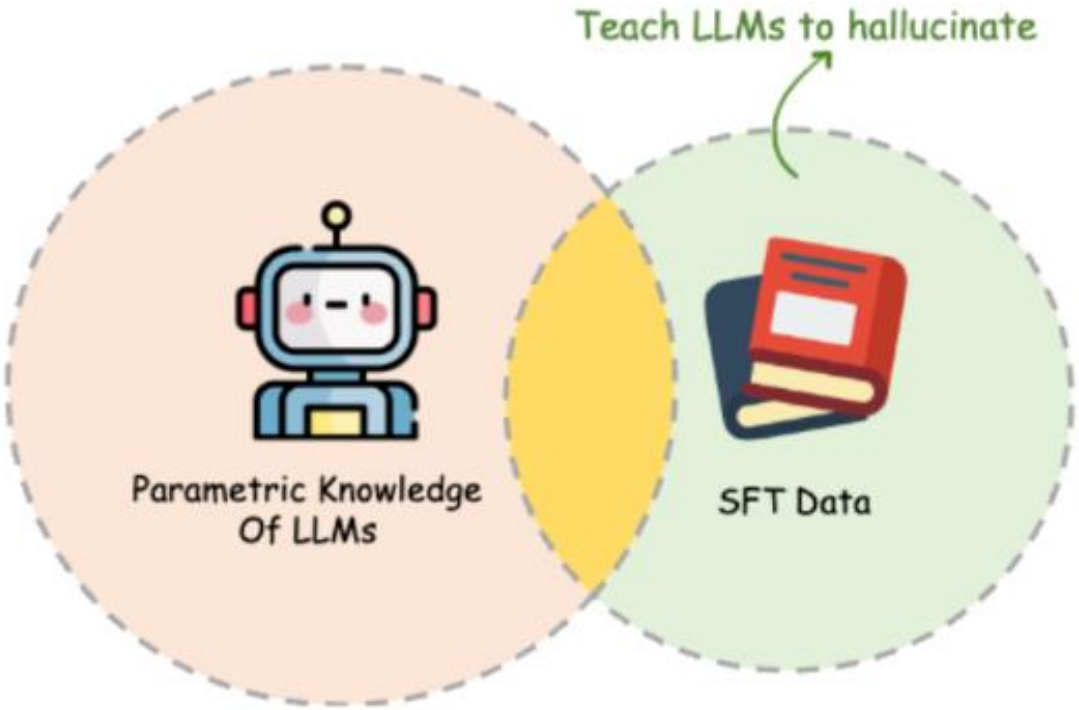


不同幻觉检测方法的平均运行时间分析

包含外部知识参考
FAVA Train Split数据
合成幻觉

算法总结

- 算法贡献
  - 首个支持“单响应”幻觉检测的高效方法
    - 不依赖多次采样或外部知识检索
  - 提出多源结构信号联合建模思路
    - 三类信号，四种特征
  - 算法简单、效率极高
    - Attention 得分 0.2 秒即可完成，比 SelfCheckGPT 快 45–450 倍
- 算法不足
  - 方法需要代理模型，非真正“完全黑盒”
    - 仍需一个可控的开源模型做 teacher forcing
  - 各特征得分目前未做深度融合建模
    - 评估仍以单一打分 + 阈值为主，尝试融合，未做进一步结构化集成



Verify-and-Edit: A Knowledge-Enhanced Chain-of-Thought Framework

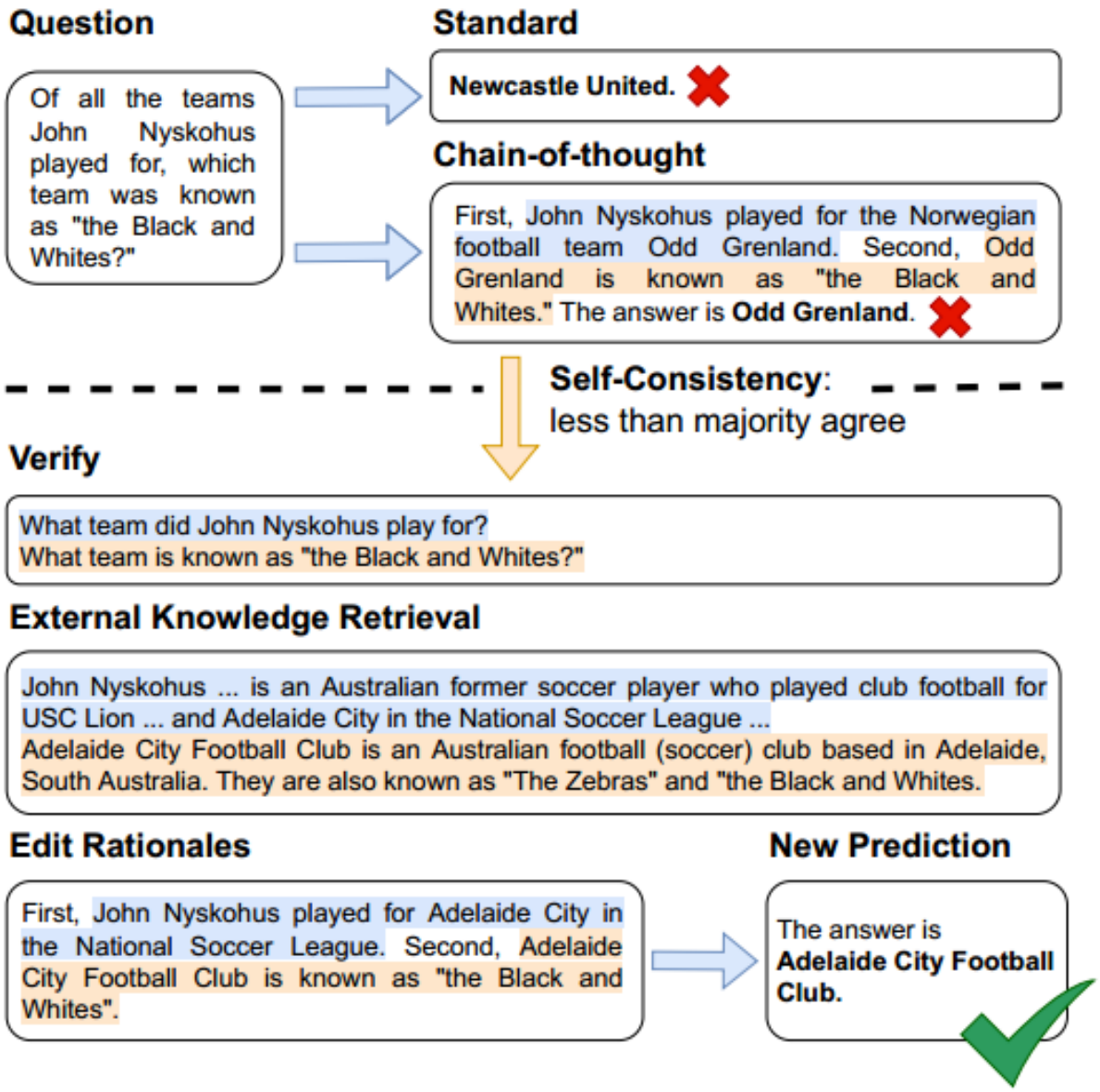
算法原理 TIPO

T	目标	在CoT推理过程中引入外部知识，对生成的推理链进行校验与修正，从而提升回答的事实准确性和最终预测表现
I	输入	用户问题、LLM推理链及回答
P	处理	1.使用自一致性方法，筛选出模型内部推理路径分歧较大的预测 2.对CoT中推理语句，利用LLM自动生成一个相关的验证性问题 3.通过开放领域检索系统查找与验证问题相关的事实性证据 4.利用检索的知识生成更具事实性的新推理语句，替换原始推理 5.将更新后的推理链重新输入LLM，生成新的最终答案
O	输出	校验编辑后的推理链、修改后的答案

P	问题	通过生成可解释的推理链来改善复杂推理任务上的可信度这一方法在知识密集的任务中仍存在事实性问题
C	条件	开放领域的知识检索系统、推理路径自一致性评估能力
D	难点	验证性问题生成、检索系统质量、编辑后推理链逻辑
L	水平	ACL 2023 CCF A

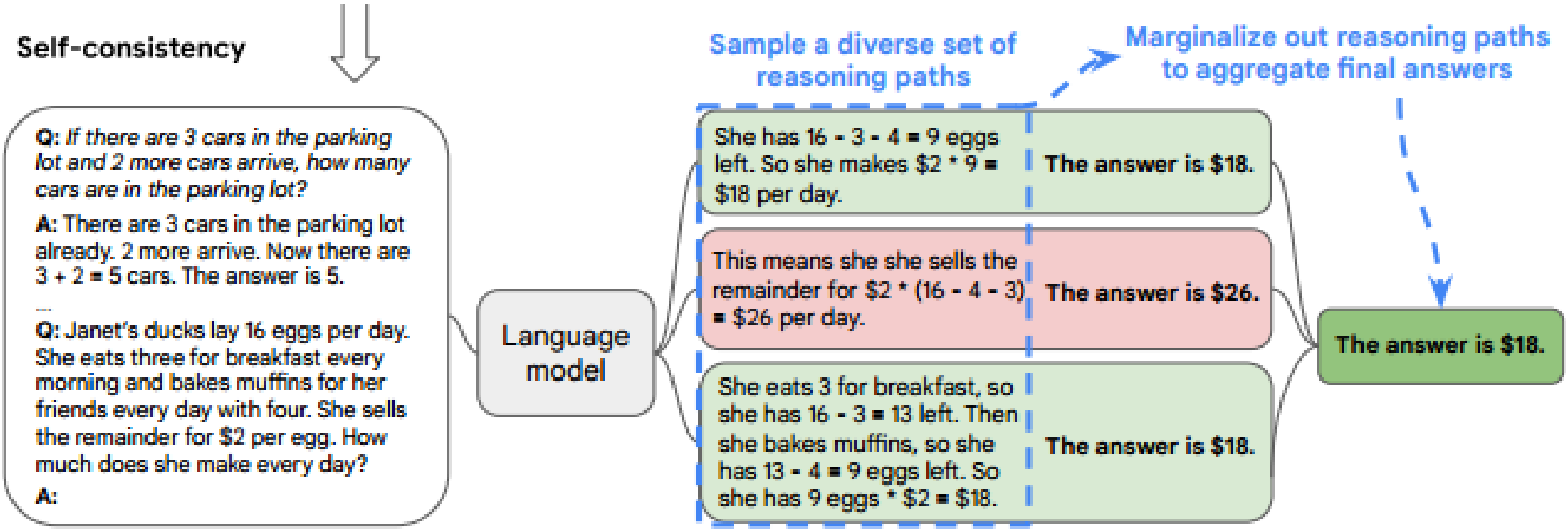
算法原理 创新说明

- **Verify-and-Edit**
  - 基于一致性判断的不确定样本筛选：借助 Self-Consistency（自一致性）方法，从多个 CoT 推理路径中判断当前回答是否可信
  - “验证式问题生成”+ 外部知识检索 + 推理编辑组合
  - 推理-答案重生成：使用修正后的 CoT作为 prompt 输入，重新生成最终答案



算法原理 VE

- 一致性判断
  - 目标：不对所有样本都编辑，而是“有选择地编辑”
  - 关键点：Self-Consistency
  - 若最终预测答案中最多的答案出现次数 < [n/2]，即“多数意见无法达成一致”，判定为“不确定”
  - 核心思想：一致性低的样本，更可能含有虚假推理链



算法原理 VE

- 思维链编辑
  - Facts：验证原始推理中涉及的事实
    - 生成验证性问题，注意避免搜索偏移
    - 检索外部知识，使用 Sentence-BERT 计算验证问题和句子的语义相似度，选取 top-3 相似度最高的句子作为检索证据
  - Reasoning：利用检索证据重写推理句
    - 正确的事实+原逻辑结构
    - 不直接替换句子，而是生成具有同等逻辑功能、但事实正确的语句
- 重新作答
  - 将新编辑后的推理链作为 Prompt 提交给 GPT，模型根据新的、更真实的 CoT 重新生成最终答案
  - 不是“修了推理就用旧答案”，而是让模型“读完新推理再重新思考”

算法原理 VE

Algorithm 1 Verify-and-Edit

Require: The original question q; An n-shot CoT prompt p_{cot}
Require: An LLM $f(\cdot)$; LM number of completions n; LM decoding temperature τ
Require: An external knowledge retrieval model $g(\cdot)$
Require: n-shot prompts for verifying question generation (p_{vq}) and answer generation (p_{va})

输入

$R, A \leftarrow f(p_{cot}, q, n, \tau)$ ▷ Generate a set of reasonings (R) and answers (A).

第一阶段：生成初始推理链和答案集合

$s^*_{sc} \leftarrow \max P(a|p_{cot}, q), a \in A$ ▷ The highest self-consistency score among all answers.
$r^*, a^* \leftarrow \arg\max P(a|p_{cot}, q), a \in A$ ▷ Reasoning and answer with highest self-consistency.
if $s^*_{sc} < \lceil \frac{n}{2} \rceil$ **then** ▷ Edit reasoning with a less-than-majority-agree consistency.

第二阶段：计算一致性，判断是否需编辑

  for $o_i \in r^*$ **do** ▷ Edit each sentence in the reasoning.
    $u \leftarrow f(p_{vq}, q, o_i)$ ▷ Generate verifying question.
    $v \leftarrow g(u)$ ▷ Retrieve external knowledge.
    $w \leftarrow f(p_{va}, u, v)$ ▷ Generate verifying answer.
    $o_i \leftarrow w$ ▷ Edit original reasoning sentence with verifying answer.

第三阶段：验证并编辑推理链

  end for
  $a^* \leftarrow f(p_{cot}, q, r^*)$ ▷ Generate final answer with edited reasoning.
  return a^*

第四阶段：基于更新后的推理链重新预测

else if $s^*_{sc} \geq \lceil \frac{n}{2} \rceil$ **then** ▷ Answer with high consistency is left as-is.
  return a^*
end if

无需编辑的情况

实验设计　数据资源

- 数据资源及指标

数据集	类型	样本数	难点	指标
HotpotQA	多跳问答	250	有干扰段落、需多步逻辑	EM, AUC
2WikiMultihop	多跳问答	1000	基于结构逻辑规则	EM, AUC
FEVER	事实验证	1000	推理浅但逻辑要求高	Accuracy

- 基线实验
  - Standard Prediction：直接使用 few-shot prompt 预测答案
  - Original CoT：使用 Chain-of-Thought 推理链生成答案
  - CoT with Self-Consistency：生成多条推理链 + 多答案后投票选择
  - Calibrator：使用预测解释的可信度对最终答案进行概率调整
  - ReAct：推理 + 动作结构，动态调用知识源

实验结果　对比实验

- 实验结果

Method	knowledge	EM	ΔEM	AUC
CoT-SC → ReAct	Wiki.	34.2%	+0.8%	-
ReAct → CoT-SC	Wiki.	35.1%	+1.7%	-
Standard	-	23.1%	-	43.24
CoT	-	31.8%	-	38.30
CoT-SC	-	31.2%	-	34.97
CoT-SC + Calib.	Dataset	-	-	49.00
CoT-SC + VE	Wiki.	35.7%	+4.5%	45.62
CoT-SC + VE	DRQA	36.0%	+4.8%	46.06
CoT-SC + VE	Google	37.7%	+6.5%	47.98
CoT-SC + VE	Dataset	**56.8%**	**+25.6%**	**60.94**

Adversarial HotpotQA数据集

原始CoT+SC本身提升不大：多路径并不能纠正事实错误，只能做投票，容易误导

Verify-and-Edit显著优于ReAct
Google检索源最优，说明搜索引擎对事实校验最有效

使用ground-truth（oracle检索）时性能大幅提升：检索质量是该框架性能的关键

AUC提升明显（一致性与准确率协同提升）：VE能拉高不一致样本的准确率，验证其“知道什么时候错了”

实验结果　对比实验

- 实验结果

Method	knowledge	EM	ΔEM	AUC
Standard	-	16.9%	-	35.89
CoT	-	28.4%	-	16.64
CoT-SC	-	27.7%	-	17.16
CoT-SC + Calib.	Dataset	-	-	24.13
CoT-SC + VE	Wiki.	33.1%	+5.4%	28.32
CoT-SC + VE	DRQA	31.1%	+3.4%	27.75
CoT-SC + VE	Google	33.6%	+5.9%	30.06
CoT-SC + VE	Dataset	**37.2%**	**+9.5%**	**32.28**

Method	knowledge	Accuracy	Δ Accuracy
CoT-SC → ReAct	Wiki.	-	+4.2%
ReAct → CoT-SC	Wiki.	-	+1.6%
Standard	-	46.8%	-
CoT	-	50.0%	-
CoT-SC	-	52.0%	-
CoT-SC + Calib.	-	33.7%	
CoT-SC + VE	Wiki.	53.6%	+1.6%
CoT-SC + VE	DRQA	53.3%	+1.3%
CoT-SC + VE	Google	53.9%	+1.9%

2WikiMultihopQA 数据集

VE在结构化推理任务中同样有效

Wikipedia和Google在开放领域结构问答中表现更优

FEVER 数据集

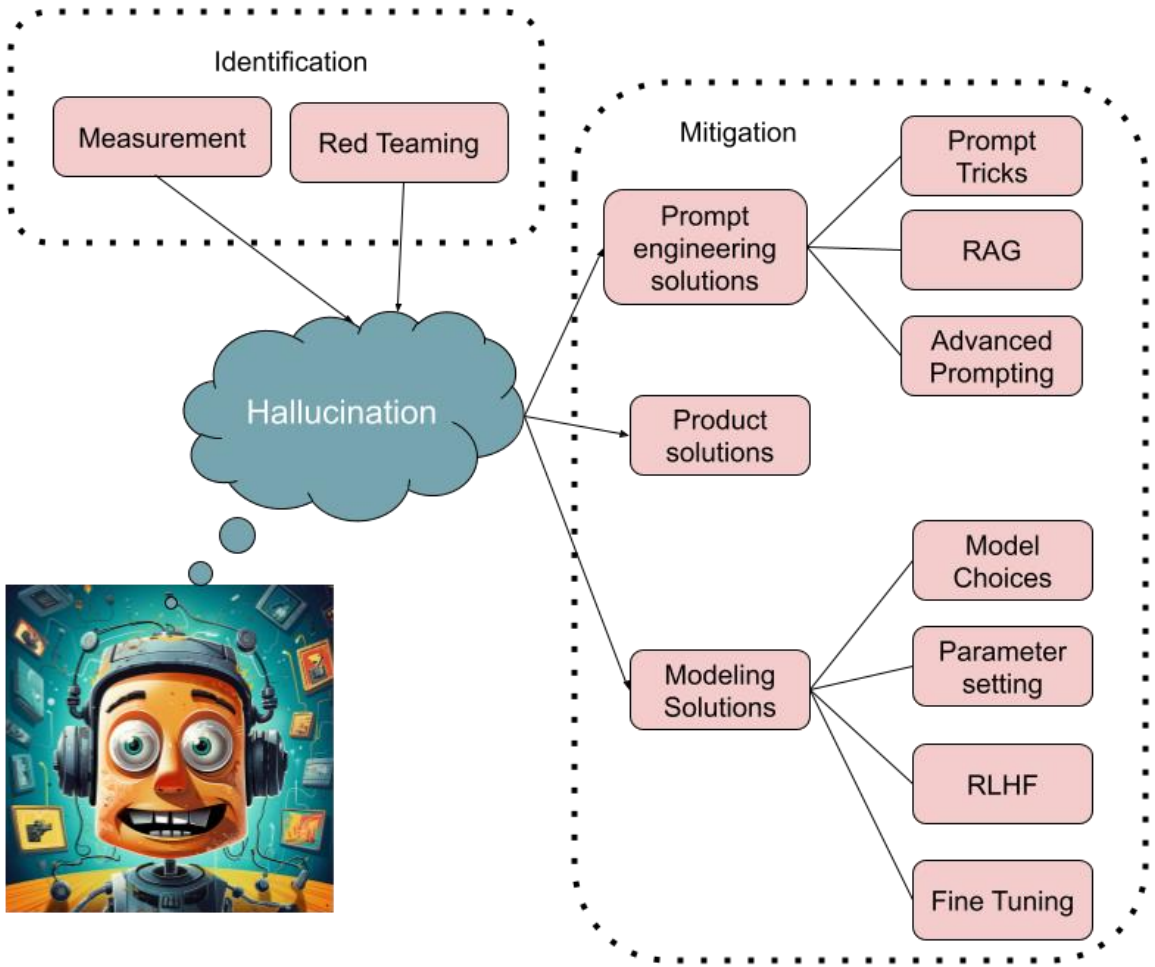
Calibrator 完全失效

VE在非问答类任务上仍有提升，但幅度小

Google 检索略优，体现泛化性与事实可靠性

算法总结

- 算法贡献
  - 提出“推理链事后编辑”框架
  - 引入一致性驱动的选择性编辑机制
  - 融合外部知识检索与自然语言生成
  - 多任务实验验证 + 显著性能提升
- 算法不足
  - 检索模块仍是性能瓶颈
  - 编辑推理句依赖 LLM 对 prompt 格式的掌握
  - 对一致性阈值敏感
  - 不适合所有任务（如简单 QA 或无需推理任务）



特点总结与未来展望

特点总结与未来展望

- 特点总结
  - AI幻觉检测：LLM-Check
    - 支持“单响应”幻觉检测的高效方法
    - 多源结构信号联合建模思路
    - 需要代理模型，非真正“完全黑盒”
    - 评估仍以单一打分 + 阈值为主，尝试融合，未做进一步结构化集成
  - AI幻觉缓解：Verify-and-Edit
    - 推理链事后编辑，一致性驱动的选择性编辑机制
    - 融合外部知识检索与自然语言生成
    - 检索模块仍是性能瓶颈
    - 不适合所有任务（如简单 QA 或无需推理任务）

特点总结与未来展望

- 未来展望——AI幻觉的创造力价值
  - 科学发现：从“错误”到突破的范式跃迁
    - 蛋白质设计：大卫·贝克团队利用 AI “错误折叠”启发新型蛋白质结构，获2024诺贝尔化学奖，认为AI幻觉是“从零开始设计蛋白质”的关键


Article | Published: 01 December 2021

De novo protein design by deep network hallucination

Ivan Anishchenko, Samuel J. Pellock, Tamuka M. Chidyausiku, Theresa A. Ramelot, Sergey Ovchinnikov, Jingzhou Hao, Khushboo Bafna, Christoffer Norn, Alex Kang, Asim K. Bera, Frank DiMaio, Lauren Carter, Cameron M. Chow, Gaetano T. Montelione & David Baker

  - 娱乐游戏：创造新的视觉和听觉体验
    - AI生成的虚拟环境和角色设计为游戏开发人员提供了无限的可能性，增强了玩家的沉浸感和探索欲
    - AI幻觉还被用于生成故事、对话和诗歌，为游戏和文学创作提供灵感

特点总结与未来展望

- 未来展望——AI幻觉的创造力价值
  - 文艺设计：突破人类思维定式的“超现实引擎”
  - 技术创新：从“缺陷”到方法论的转化
    - DeepMind团队发现，AI在图像分割任务中产生的“超现实边界”虽不符合真实场景，却意外提升了自动驾驶系统对极端天气（如浓雾、暴雨）的识别精度
  - 新型科研范式：科学界正构建“AI幻觉-实验验证-理论重构”的三阶段研究流程
    - 加州理工学院团队通过AI生成虚构导管设计，通过新型人工智能技术优化后的新设计，在实验中证实将向上游游动的细菌数量减少了100倍，形成“疯狂创意→理性筛选”的创新闭环

参考文献

[1] Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions[J]. ACM Transactions on Information Systems, 2025, 43(2): 1-55.

[2] Sriramanan G, Bharti S, Sadasivan V S, et al. Llm-check: Investigating detection of hallucinations in large language models[J]. Advances in Neural Information Processing Systems, 2024, 37: 34188-34216.

[3] Varshney N, Yao W, Zhang H, et al. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation[J]. arXiv preprint arXiv:2307.03987, 2023.

相关知识

- 提示词工程
  - 知识边界限定：通过时空维度约束降低虚构可能性（本质：约束大模型）
    - 时间锚定法：基于2023年之前的公开学术文献，分步骤解释量子纠缠现象
    - 知识锚定法：基于《中国药典》回答，若信息不明确请注明“暂无可靠数据支持”
    - 领域限定法：作为临床医学专家，请列举FDA批准的5种糖尿病药物置信度声明：如果存在不确定性，请用[推测]标签标注相关陈述
    - 上下文提示：根据《2024全球能源转型报告》（国际能源署，2024年1月发布）显示：2030年 光伏发电成本预计降至0.02美元/千瓦时，但储能技术突破仍是普及瓶颈。请基于此数据，分析中国西部光伏基地发展的三个关键挑战，并标注每个挑战与原文结论的逻辑关联
    - 生成参数协同控制：请以temperature=0.3的严谨模式，列举2024年《柳叶刀》发表的传染病研究

相关知识

- 提示词工程
  - 对抗性提示：强制暴露推理脆弱点，用户可见潜在错误路径（本质：大模型自我审查）
    - 植入反幻觉检测机制： 请用以下格式回答：- 主要答案（严格基于公开可验证信息）- [反事实检查] 部分（列出可能导致此答案错误的3种假设)
    - 预设验证条件，迫使模型交叉检查信息：请先回答“量子纠缠能否证明灵魂存在？”，然后从以下角度验证答案的可靠性： 1. 物理学界主流观点； 2. 近五年相关论文数量； 3. 是否存在可重复实验证据。
    - 链式验证：请完成以下验证链：1. 陈述观点：_____ 2. 列出支撑该观点的三个权威数据源 3. 检查每个数据源是否存在矛盾信息 4. 最终结论（标注可信度等级）

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！