北京交通大学 BEIJING JIAOTONG UNIVERSITY | 知行

多模态大模型幻觉量化技术研究

Sarah: Hallucination Evaluation and Detection for Large Vision Language Models with Semantic Information Locator and Purifier in Uncertainty Quantification Method

毕业设计终期答辩 | 北京交通大学电子信息工程学院

答辩人：方悦　　指导教师：张阳

北京交通大学
BEIJING JIAOTONG UNIVERSITY

目录 /CONTENTS

北京交通大学
BEIJING JIAOTONG UNIVERSITY

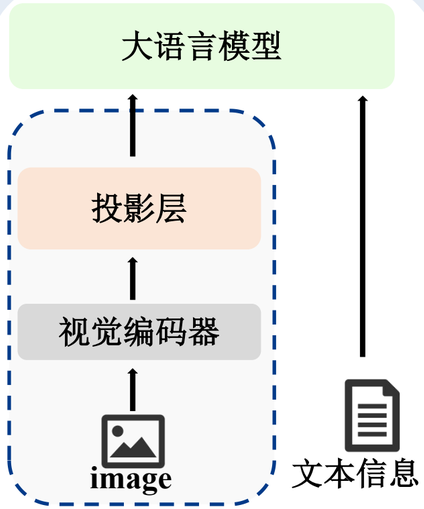
01

研究背景与意义

Part One： Research Background and Meaning

北京交通大学 BEIJING JIAOTONG UNIVERSITY

多模态大模型典型代表：大视觉语言模型(Large Vision Language Model, LVLM)



大视觉语言模型的基本结构

大视觉语言模型的应用

大视觉语言模型基本结构

- **视觉编码器（Vision Encoder）：** 编码多模态输入以获取特征信息。

$$F_x = \text{Vision\_Encoder}(I_x)$$

- **投影层（Projection Layer）：** 对齐多模态特征信息。

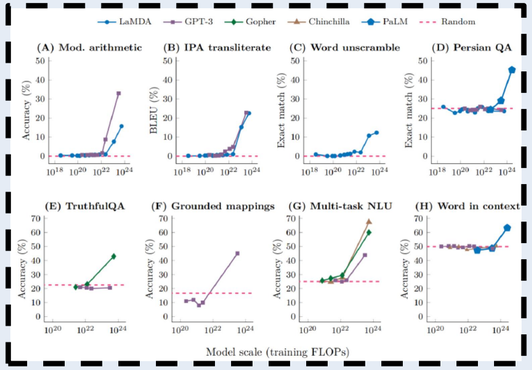
$$P_x = Projection\_Layer_{x \to T}(F_x)$$

- **大语言模型（LLM）：** 处理多模态特征信息，基于特征信息实现语义理解、推理、决策等任务

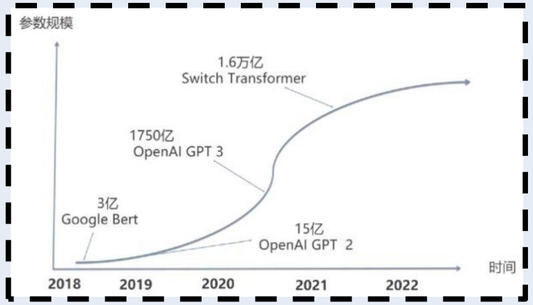
$$S_x = LLM(P_x, F_T)$$

Model	I→O	Modality Encoder	Input Projector	LLM Backbone	Output Projector
Flamingo	I+V+T→T	I/V: NFNet-F6	Cross-attention	Chinchilla-1.4B/7B/70B	–
BLIP-2	I+T→T	I: CLIP/Eva-CLIP ViT@224	Q-Former w/ Linear Projector	Flan-T5/OPT	–
LLaVA	I+T→T	I: CLIP ViT-L/14	Linear Projector	Vicuna-7B/13B	–
MiniGPT-4	I+T→T	I: Eva-CLIP ViT-G/14	Q-Former w/ Linear Projector	Vicuna-13B	–
mPLUG-Owl	I+T→T	I: CLIP ViT-L/14	Cross-attention	LLaMA-7B	–
Otter	I+T→T	I: CLIP ViT-L/14	Cross-attention	LLaMA-7B	–
X-LLM	I+V+A+T→T	I/V: ViT-G; A: C-Former	Q-Former w/ Linear Projector	ChatGLM-6B	–
VideoChat	V+T→T	I: ViT-G	Q-Former w/ Linear Projector	Vicuna	–
InstructBLIP	I+V+T→T	I/V: ViT-G/14@224	Q-Former w/ Linear Projector	Flan-T5/Vicuna	–
PandaGPT	I+T→T	I: ImageBind	Linear Projector	Vicuna-13B	–
GILL	I+T→I+T	I: CLIP ViT-L	Linear Projector	OPT-6.7B	Tiny Transformer
PaLI-X	I+T→T	I: ViT	Linear Projector	UL2-32B	–
Video-LLaMA	I+V+A+T→T	I/V: Eva-CLIP ViT-G/14; A: ImageBind	Q-Former w/ Linear Projector	Vicuna/LLaMA	–
Video-ChatGPT	V+T→T	I: CLIP ViT-L/14	Linear Projector	Vicuna-v1.1	–
Shikra	I+T→T+I_B	I: CLIP ViT-L/14@224	Linear Projector	Vicuna-7B/13B	–
LLaVAR	I+T→T	I: CLIP ViT-L/14@224 & CLIP ViT-L/14@336	Linear Projector	Vicuna-13B	–
mPLUG-DocOwl	I_D+T→T	I: CLIP ViT-L/14	Cross-attention	LLaMA-7B	–
Lynx	I+V+T→T	I/V: Eva-CLIP ViT-1B	Cross-attention	Vicuna	–
Emu	I+V+T→I+T	I/V: Eva-CLIP-1B	Cross-attention	LLaMA-13B	MLP
DLP	I+T→T	I: CLIP/Eva-CLIP ViT	Q-&P-Former w/ Linear Projector	OPT/Flan-T5	–
BuboGPT	I+A+T→T+I_M	I: CLIP/Eva-CLIP ViT; A: ImageBind	Q-Former w/ Linear Projector	Vicuna	–
ChatSpot	I+T→T	I: CLIP ViT-L/14	Linear Projector	Vicuna-7B/LLaMA	–
IDEFICS	I+T→T	I: OpenCLIP	Cross-attention	LLaMA	–
Qwen-VL-(Chat)	I+T→T	I: ViT@448 initialized from OpenClip's ViT-bigG	Cross-attention	Qwen-7B	–
LaVIT	I+T→I+T	I: ViT	Cross-attention	LLaMA-7B	–
NExT-GPT	I+V+A+T→I+V+A+T	I/V/A: ImageBind	Linear Projector	Vicuna-7B	Tiny Transformer
DreamLLM	I+T→I+T	I: CLIP ViT-L	Linear Projector	Vicuna	MLP
AnyMAL	I+V+A+T→T	I: CLIP ViT/L&ViT-G&DinoV2; V: Intervideo; A: CLAP	I/V: Cross-attention; A: Linear Projector	LLaMA-2	–
MiniGPT-5	I+T→I+T	I: Eva-CLIP ViT-G/14	Q-Former w/ Linear Projector	Vicuna-7B	Tiny Transformer
LLaVA-1.5	I+T→T	I: CLIP ViT-L@336	MLP	Vicuna-v1.5-7B/13B	–
MiniGPT-v2	I+T→T	I: Eva-CLIP ViT@448	Linear Projector	LLaMA-2-Chat-7B	–
CogVLM	I+T→T	I: Eva-2-CLIP ViT	MLP	Vicuna-v1.5-7B	–
Qwen-Audio	A+T→T	A: Whisper-L-v2	Linear Projector	Qwen-7B	–
DRESS	I+T→T	I:Eva-CLIP ViT-G/14	Linear Projector	Vicuna-v1.5-13B	–
X-InstructBLIP	I+V+A+3D+T→T	I/V: Eva-CLIP ViT-G/14; A: BEATs; 3D: ULIP-2	Q-Former w/ Linear Projector	Vicuna-v1.1-7B/13B	–
CoDi-2	I+V+A+T→I+V+A+T	I/V/A: ImageBind	MLP	LLaMA-2-Chat-7B	MLP
RLHF-V	I+T→T	I: BEiT-3	Linear Projector	Vicuna-v1-13B	–
Silkie	I+T→T	I: ViT initialized from OpenCLIP's ViT-bigG	Cross-attention	Qwen-7B	–
Lyrics	I+T→T	I: CLIP ViT-L/14&Grounding-DINO-T &SAM-HQ&ViT-H&RAM++	MQ-Former w/ Linear Projection	Vicuna-13B	–
VILA	I+T→T	I: ViT@336	Linear Projector	LLaMA-2-7B/13B	–
IntrenVL	I+V+T→T	I/V: InternViT-6B; T: LLaMA-7B	Cross-attention w/ MLP	QLLaMA-8B & Vicuna-13B	–
ModaVerse	I+V+A+T→I+V+A+T	ImageBind	Linear Projector	LLaMA-2	MLP
MM-Interleaved	I+T→I+T	I: CLIP ViT-L/14	Cross-attention	Vicuna-13B	Tiny Transformer

大语言模型的涌现能力



庞大的参数连和复杂的模型结构激发模型出现涌现能力



大视觉语言模型携带大量参数

[1]Zhu D, Chen J, Shen X, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models[J]. arXiv preprint arXiv:2304.10592, 2023.
[2]Liu H, Li C, Wu Q, et al. Visual instruction tuning[J]. Advances in neural information processing systems, 2024, 36.
[3]Xu G, Jin P, Hao L, et al. LLaVA-o1: Let Vision Language Models Reason Step-by-Step[J]. arXiv preprint arXiv:2411.10440, 2024
[4]Wang W, Lv Q, Yu W, et al. Cogvlm: Visual expert for pretrained language models[J]. arXiv preprint arXiv:2311.03079, 2023.
[5]Chen Z, Wu J, Wang W, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 24185-24198.

北京交通大学 BEIJING JIAOTONG UNIVERSITY

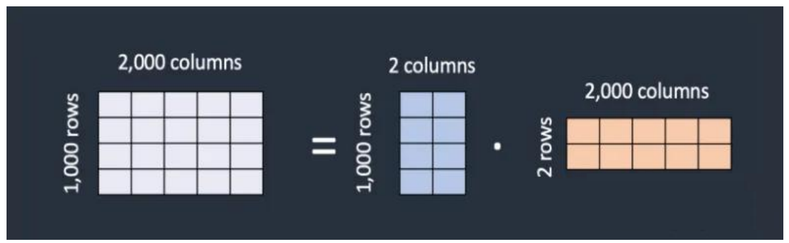
大视觉语言模型的训练策略

监督微调（Supervised Fine-Tuning，SFT）

在预训练模型的基础上**利用标注好的下游任务数据**进行进一步训练的过程。模型将通过**最小化预期输出与实际输出之间的误差**来诱导模型纠正原参数分布：

$$L_{SFT} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\log P(y_t^{(i)}|x^{(i)}, y_{<t}^{(i)}; \theta)$$

这里N代表样本数量，T代表序列长度（如文本生成任务中的token数），$y_t^{(i)}$代表第i个样本的第t个目标token，$x^{(i)}$代表第i个样本的输入（如提示词或上下文），θ代表模型微调过程中更新的参数。



低秩适应（LoRA）工作原理

强化学习微调（Reinforcement Learning Fine-Tuning，RLFT）

通过**与环境的交互**来学习最优行为策略。模型根据奖励信号来调整自己的行为策略，目标是**最大化累积奖励**。GRPO（Group Relative Policy Optimization）在经典PPO（Proximal Policy Optimization）的基础上进行改进，通过引入分组相对奖励机制和策略更新约束，摆脱传统强化学期对奖励模型的依赖，提升策略优化的稳定性和样本效率。

$$\zeta_{GRPO}(\theta) = E[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{\|o_i\|}\sum_{t=1}^{\|o_i\|}min[r_{ratio}, clip(r_{ratio}, 1-\varepsilon, 1+\varepsilon)]\cdot\widehat{A_{i,t}}] - KL$$

$$r_{ratio} = \frac{\Pi_\theta(O_{i,t}|q, O_{i<t})}{\Pi_{\theta_{old}}(O_{i,t}|q, O_{i<t})}$$

这里$\|o_i\|$为输出序列的长度，o是在就策略下生成的输出序列，$A_{i,t}$是分组相对意义上的优势，KL正则用于限制策略与一个参考策略之间的差异，Clip用于约束策略更新的幅度，避免因单步更新过大而导致训练不稳定。

模型通过与环境的交互获取奖励信号，调整行为策略

强化学习相较于监督学习的特性

(1) 训练信号的稀疏性和延迟性
(2) 数据分布的非独立性
(3) 环境动力学的未知性

[6]Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models[J/OL]. arXiv preprint arXiv:2106.09685, 2021. [2025-05-06]. https://arxiv.org/abs/2106.09685.
[7]Longteng Zhang, Lin Zhang, Shaohuai Shi, Xiaowen Chu, Bo Li. LoRA-FA: Memory-efficient Low-rank Adaptation for Large Language Models Fine-tuning[J/OL]. arXiv preprint arXiv:2308.03303, 2023. [2025-05-06]. https://arxiv.org/abs/2308.03303.
[8]Hu E J, Shen Y, Wallis P, et al. LoRA: Low-Rank Adaptation of Large Language ModelsJ/OLJ/OLJ/OL. arXiv:2106.09685, 20212025−05−06
[9]Sutton R S, Barto A G. Reinforcement Learning: An Introduction[M]. 2nd ed. Cambridge, MA: MIT Press, 2018.
[10]Ladosz P, Weng L, Kim M, et al. Exploration in Deep Reinforcement Learning: A Survey[J/OL]. arXiv preprint arXiv:2205.00824, 2022. [2025-05-06].

北京交通大学 BEIJING JIAOTONG UNIVERSITY

多模态大语言模型的幻觉（Hallucination）主要体现为大模型输出内容与现实世界的事实或用户输入（指令、图像中物体关系、属性数量等）之间的矛盾。具体可分类为事实性幻觉（Factuality Hallucination）和忠实性幻觉（Faithfulness Hallucination）。

提问	正确答案	Deepseek-V3	QWen2.5-Max	豆包
水浒传中李逵为什么要大闹五台山？	李逵并没有大闹五台山，事实上，《水浒传》中大闹五台山的是鲁智深。	误以为宋江强抢民女（实为他人冒充），怒闯五台山讨说法。	李逵大闹五台山是因为他误以为鲁智深在寺庙中受欺负，于是出手相助，引发混乱。	水浒传中大闹五台山的是鲁智深而不是李逵。鲁智深大闹五台山是因为他出家后难忍寺中的清规戒律，又喝酒吃肉，还打伤了寺里的僧人等。

SOTA（State-of-the-arts） 大语言模型的事实性幻觉表现

Meta - Galactica:

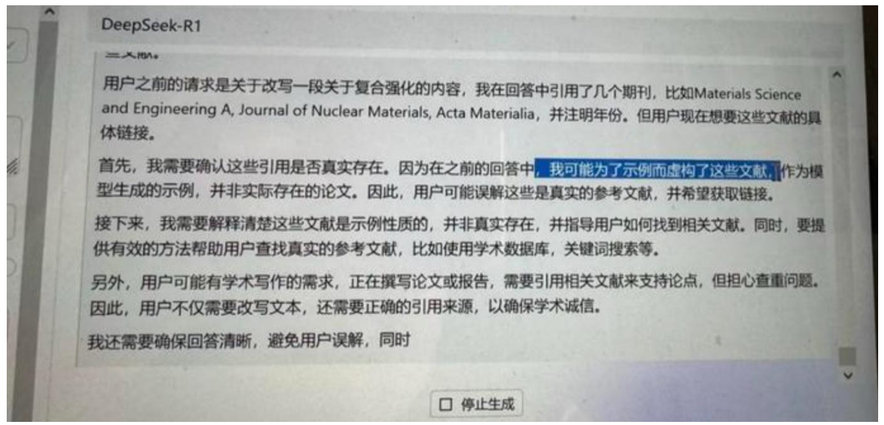
当被要求起草一篇关于如何创建虚拟形象的论文时，Galactica援引了该领域某一研究人员的一篇论文，但是现实中该研究人员并没有写过这篇文章，它完全是杜撰的。

DeepSeek:

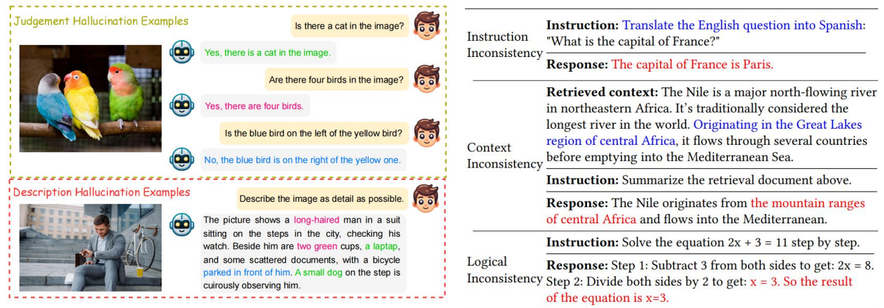
据Vectara发布的“幻觉排行榜”报告显示，DeepSeek-R1的幻觉率达到了14.3%，远高于Deepseek-V3的3.9%，在排行榜中处于90多位。

ChatGPT:

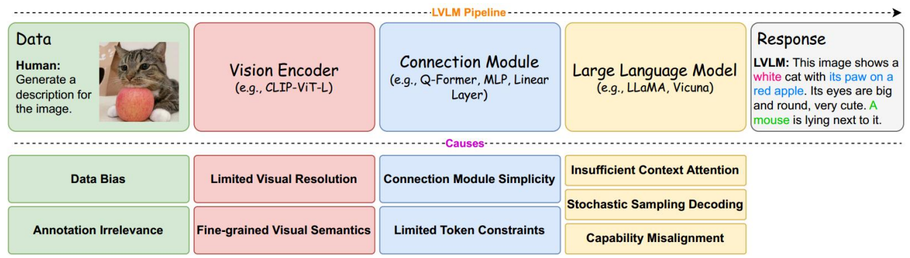
沃顿商学院莫里克教授：“ChatGPT是无所不知、渴望取悦别人但有时会撒谎的实习生”。ChatGPT曾提出“黑洞的（强）磁场是由其附近极强的引力产生的”荒谬理论。它也会使用完全杜撰的数据生成特斯拉电动汽车的季度财报。



Deepseek-R1承认自己在撒谎



携带忠实性幻觉的大视觉语言模型输出



大视觉语言模型在不同阶段下的幻觉诱因

[11]Liu H, Xue W, Chen Y, et al. A survey on hallucination in large vision-language models[EB/OL]. arXiv preprint arXiv:2402.00253, 2024. https://arxiv.org/abs/2402.00253
[12]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, 2024.

北京交通大学 BEIJING JIAOTONG UNIVERSITY

多模态大模型幻觉量化评估与检测

基于不确定性量化的幻觉评估

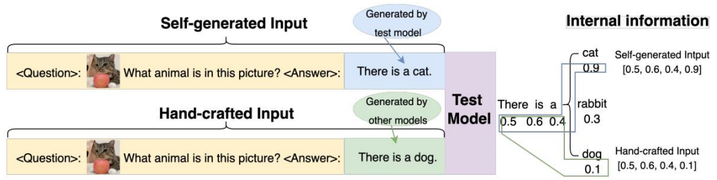
- 无需借助大量外部工具，**独立性强**；
- 一定条件下仅需模型**单轮推理，大大降低资源损耗**；
- 内部评估方法，具备**高度适应能力与可扩展性**，适用于不同模型结构与任务。

不同概率分布下令牌的不确定性分布（1）

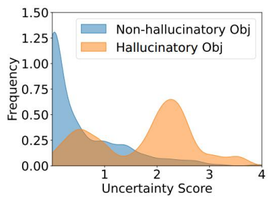
不确定性量化公式举例：

$$PP(S) = P(w_1, w_2, \ldots, w_n)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1, w_2, \ldots, w_n)}} = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_1, w_2, \ldots, w_n)}}$$

$$H(X) = -\sum_{N} p(x_i) \log p(x_i)$$

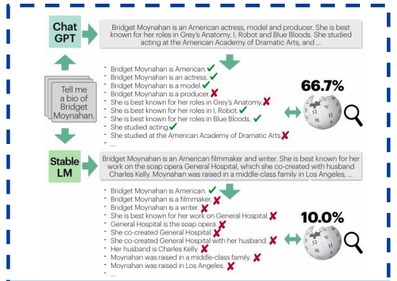


大视觉语言模型生成中的概率分布

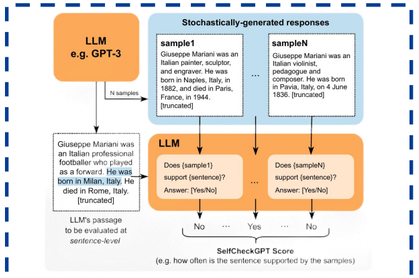


不同概率分布下令牌的不确定性分布（2）

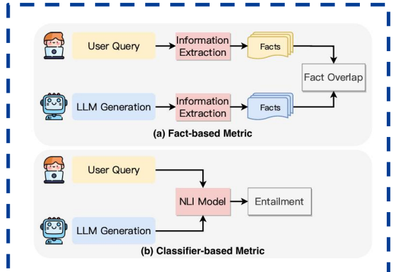
主流幻觉评估方法的原理及缺陷



基于事实



基于更高级大模型



基于模型行为

基于事实的幻觉评估：

（1）严重依赖高精度检索工具和实时更新的知识库支持，在封闭领域或专业场景中适用性受限；

（2）计算开销大，尤其是多轮检索与比对会导致响应延迟；

（3）细粒度不足，难以检测逻辑不一致等复杂幻觉类型。

基于更高级大模型的幻觉评估：

（1）成本高昂，依赖高性能大模型的API调用；

（2）可解释性差，难以定位幻觉的具体来源。

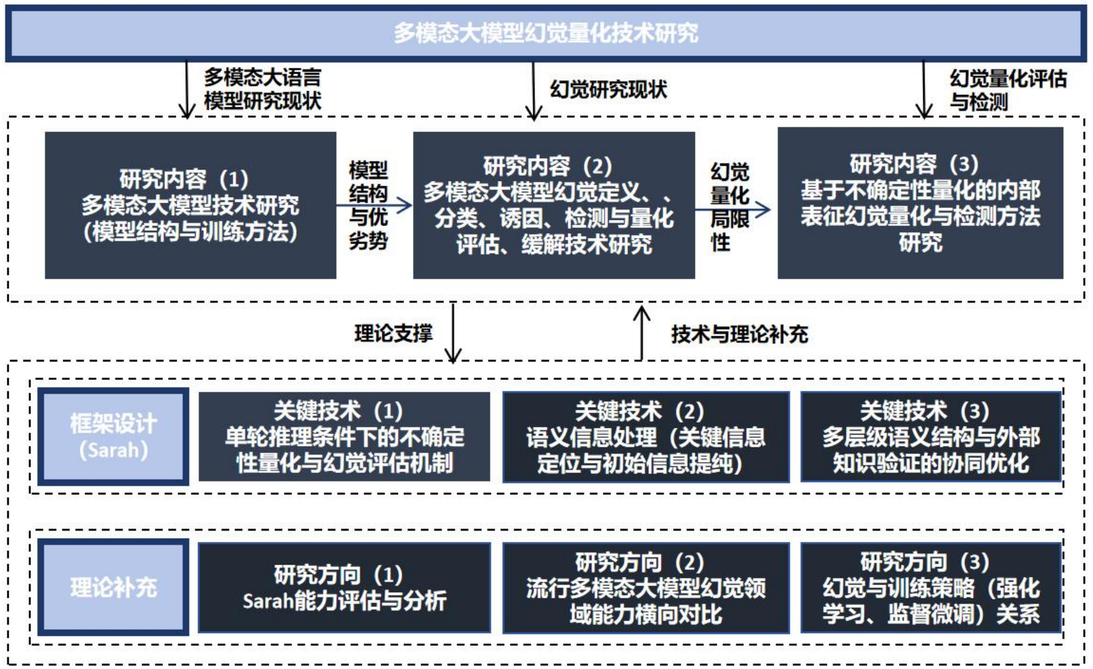
基于模型行为的幻觉评估：

（1）覆盖性不足：难以穷举所有可能的幻觉场景（如复杂推理链中的隐性矛盾），导致评估结果片面或过拟合。

[13] Elaraby M, Lu M, Dunn J, et al. Halo: Estimation and reduction of hallucinations in open-source weak large language models[EB/OL]. arXiv preprint arXiv:2308.11764, 2023. https://arxiv.org/abs/2308.11764
[14]Zhou Y, Cui C, Yoon J, et al. Analyzing and mitigating object hallucination in large vision-language models[J]. arXiv preprint arXiv:2310.00754, 2023.
[15]Manakul P, Liusie A, Gales M. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models[C]//The 2023 Conference on Empirical Methods in Natural Language Processing.
[16]Zhang R, Zhang H, Zheng Z. VL-Uncertainty: Detecting Hallucination in Large Vision-Language Model via Uncertainty Estimation[J]. arXiv preprint arXiv:2411.11919, 2024.

北京交通大学 BEIJING JIAOTONG UNIVERSITY

研究内容



研究意义

基于幻觉检测与评估技术最新进展构建完整系统的框架，为后续领域研究提供参考依据。

提出一种新颖且经济高效的LVLM幻觉检测与评估框架（Sarah），幻觉检测准确率（86.6%）与检测效率超越现有方法，幻觉评估能力优异。

使用Sarah评估SOTA大视觉语言模型的幻觉表现。指出超过13.4%的模型响应包含幻觉信息。

使用Sarah量化训练策略与幻觉表现的关系。指出强化学习相较于监督学习有更大的倾向诱导模型生成幻觉内容。

幻觉量化评估技术代码框架

```
| - experiment
  | - baseline
    | - [SOTA方法和Baseline方法的实现]
    | - FAITHSCORE
    | - Semantic Entropy
    | - GAVIE
    | - InterrogateLLM
    | - EasyDetect
  | - tools
    | - draw_length_and_hal.py
      | - [检验幻觉与输出长度的关系]
    | - draw_heatmap.py
      | - [语义重要性分布图]
  | - eval
    | - eval_bingo_negative.py
      | - [评估Bingo数据集上分数与幻觉呈负相关的幻觉检测结果]
    | - eval_bingo_positive.py
      | - [评估Bingo数据集上分数与幻觉呈正相关的幻觉检测结果]
    | - eval_coco_negative.py
      | - eval_coco_positive.py
    | - independent_claim_extraction
  | - claim_extraction_gpt4.py
    | - [基于gpt4的独立声明提取]
  | - claim_extraction_deepseek.py
    | - [基于deepseek的独立声明提取]
  | - claim_extraction_prompt.py
    | - [独立声明提取提示词，概率匹配提示词]
| - scripts
  | - eval.sh
    | - [评估检测结果]
  | - hal-detect.sh
    | - [幻觉检测脚本]
  | - pipeline.py
    | - [运行幻觉量化与检测主函数]
| - label_tool
  | - auto_label.py
    | - [自动幻觉打标]
  | - auto_label_prompt.py
    | - [自动打标提示词]
  | - efficiency_verify.py
    | - [验证自动打标有效性]
  | - UI_label.py
    | - [人工幻觉打标，底层代码调用及UI界面实现]
| - model_run
  | - model_run_gpt4.py
    | - [实现gpt-4o的推理及概率提取]
  | - model_run_llama.py
    | - [实现LLama-v3.2-vision的推理及概率提取]
  | - model_run_llava.py
    | - [实现LLaVA-v1.5-7b/13b的推理及概率提取]
  | - model_run_mplug.py
    | - [实现mplug-owl3的推理及概率提取]
  | - model_run_qwenvl2.py
    | - [实现QwenVL-v2.5-3b的推理及概率提取]
| - playground
  | - Bingo_benchmark
    | - answers
    | - claim_extraction_result
    | - gpt4-gt
    | - hal-detect_result
    | - human-gt
    | - semantic_locate_result
  | - Coco_caption
| - semantic_information_locator
  | - stochastic_semantic_perturbation.py
    | - [语义扰动]
  | - importance_cal.py
    | - [重要性计算]
| - semantic_information_purifier
  | - purifier.py
    | - [语义关系分类、聚合与丢弃]
  | - uncertainty_cal.py
    | - [基于令牌熵的不确定性量化及幻觉检测]
```

北京交通大学
BEIJING JIAOTONG UNIVERSITY

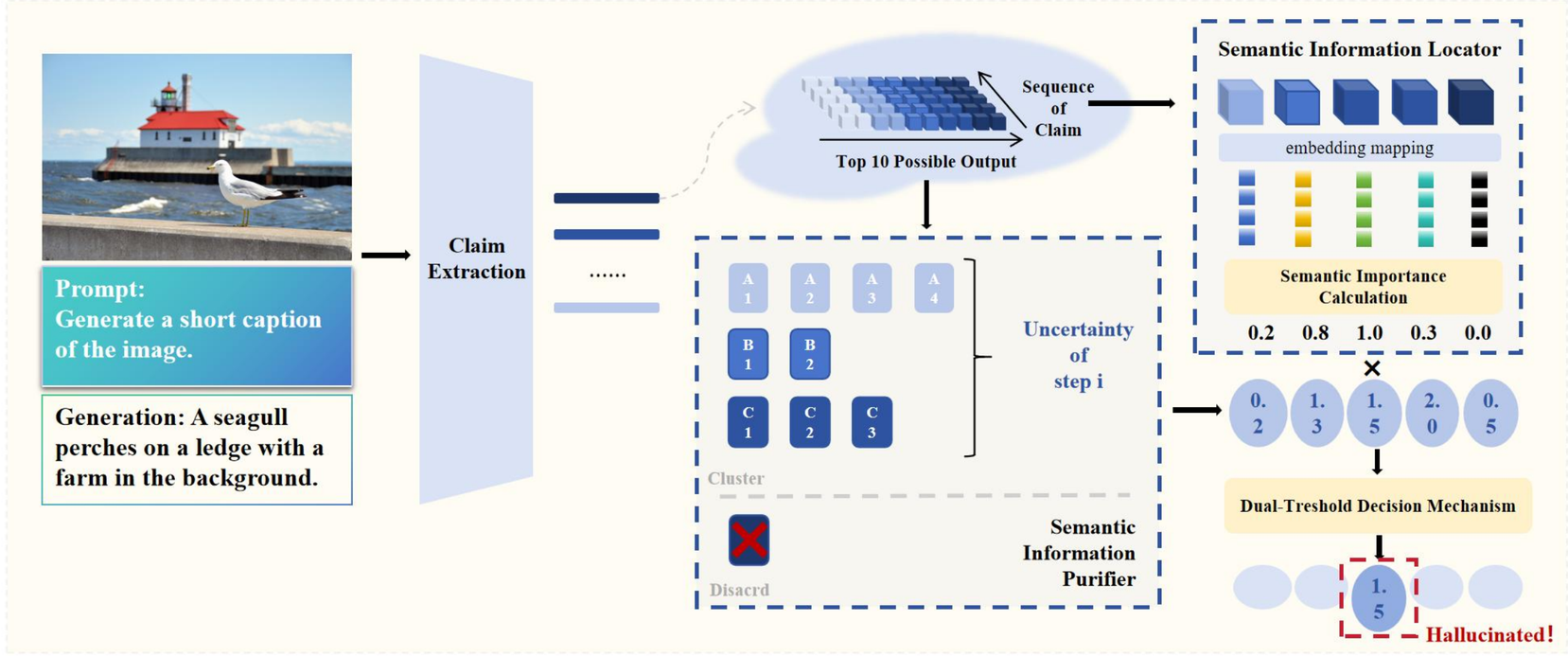
02

研究思路与方法

Part Two： Methodology and Approach

北京交通大学 BEIJING JIAOTONG UNIVERSITY

幻觉检测与量化评估方法：Sarah（Hallucination Evaluation and Detection for Large Vision Language Models with Semantic Information Locator and Purifier in Uncertainty Quantification Method）



大视觉语言模型检测与量化评估框架（Sarah）。核心结构包括独立声明提取（Independent Claim Extraction），语义信息定位器（Semantic Information Locator），语义信息净化器（Semantic Information Purifier），多阈值决策机制Multi-Threshold Decision Mechanism）

基于不确定性量化的幻觉评估

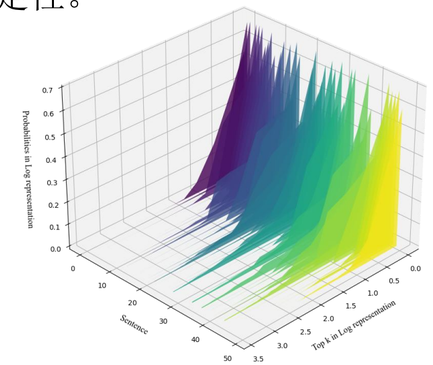
第i步生成的令牌（token）z_i概率分布：

$$P(z_{i,k} | s<i, x)，$$

其中k表示句子s的第i步生成中第k个可能的输出，x表示输入。令牌 z_i 的预测熵即为第i步输出的条件熵i：

$$U_{z_i} = -\sum_{k=1}^{K} p(z_{i,k}|s < i, x) \log p(z_{i,k}|s < i, x)$$

其中U_{zi} 表示大视觉语言模型第i步输出的不确定性。



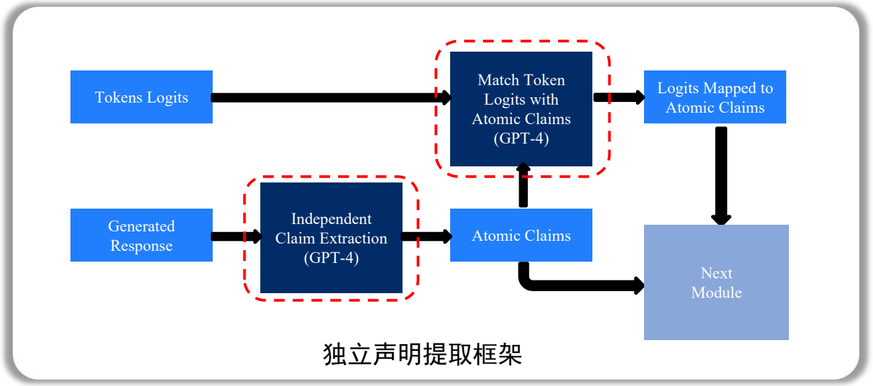
每个标记生成的前k个结果的概率分布。可以清楚地看到，前k个可能的输出占据了大多数。

技术优势

- 内部评估方法，具备高度适应能力与可扩展性。
- 仅需单轮推理，在实现高精度检测与量化效果的同时大大降低资源损耗（cost-effective）.
- 基于不确定性量化，对大视觉语言模型的自由生成任务（generate free）和开放问答任务（open-ended）友好。
- 将语义信息纳入考虑范畴，增强语义协作，减少语义干扰。

独立声明提取

- 模型生成内容通常包含复杂、冗长的句子。Generation often consists of **complex**, **lengthy** sentences.
- 将句子拆分为若干独立声明——即能够独立表达完整语义信息的最小单元。



独立声明提取框架

Please breakdown the sentence into independent claims.

Example:
Sentence: \"A seagull perches on a ledge with a lighthouse in the background.\"
Claims:
- A seagull perches on a ledge.
- There is a lighthouse in the background.

Sentence: \"{sent}\"
Claims:

用于独立声明提取的提示词

Given the fact, identify the corresponding words in the original sentence that help derive this fact. Please list all words that are related to the fact, in the order they appear in the original sentence, each word separated by comma.

Fact: {claim}
Sentence: {sent}

Words from sentence that helps to derive the fact, separated by comma:

用于令牌对齐的提示词

- **Instruction：The two lines are parallel to each other. Why?**
- **Sentence：The two lines are parallel because they have the same slope and will never intersect, regardless of how far they are extended. Parallel lines remain equidistant from each other at all points.**
- **Claim：**

{"claim_text": "The two lines are parallel.", "aligned_token_ids": [1, 2, 3, 4], "claim_id": 0}
{"claim_text": "The two lines have the same slope.", "aligned_token_ids": [4, 6, 7, 8, 9, 10], claim_id": 1}
{"claim_text": "The two lines will never intersect.", "aligned_token_ids": [12, 13, 14], "claim_id": 2}
{"claim_text": "The two lines will never intersect regardless of how far they are extended.", "aligned_token_ids": 12, 13, 14, 16, 17, 18, 19, 20, 21, 22], "claim_id": 3}

[17]Fadeeva E, Rubashevskii A, Shelmanov A, et al. Fact-checking the output of large language models via token-level uncertainty quantification[J]. arXiv preprint arXiv:2403.04696, 2024.

语义信息定位器与净化器（Semantic Information Locator and Purifier）

对句子第i步输出进行j次语义信息干扰：

$$T_{z_{i,j}} = \emptyset(T, \tau_j)$$

$$\{< T_{z_{i,0}}, T_{z_{i,1}}, T_{z_{i,j}}, \ldots, T_{z_{i,N}} > | \, j = 1, 2, \ldots, N$$

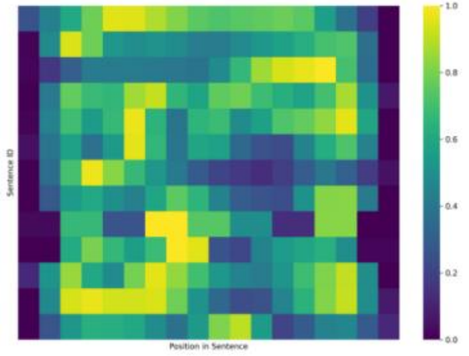
通过对BertScore的部分调整，将每次干扰映射到对齐的语义空间：

$$\{\langle \Gamma(T_{z_{i,0}}), \Gamma(T_{z_{i,1}}), T_{z_{i,j}} \ldots, \Gamma(T_{z_{i,N}}) \rangle | \, j = 1, 2, \ldots, N\}$$

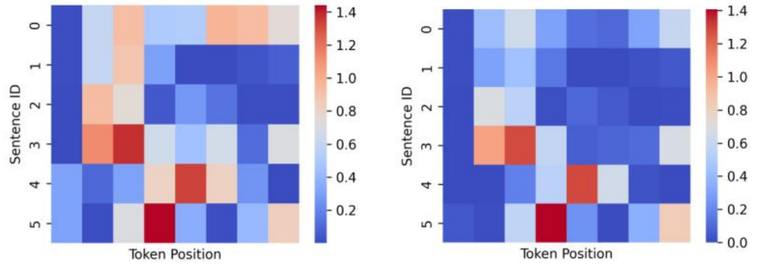
独立声明c的第i步输出 z_i有语义重要性（Semantic Importance， SI)）：

$$SI(c, z_i) = 1 - \sum_{j=1}^{M} \cos < \Gamma(T_{z_{i,j}}), \Gamma(T_{z_{i,j}}) > / M$$

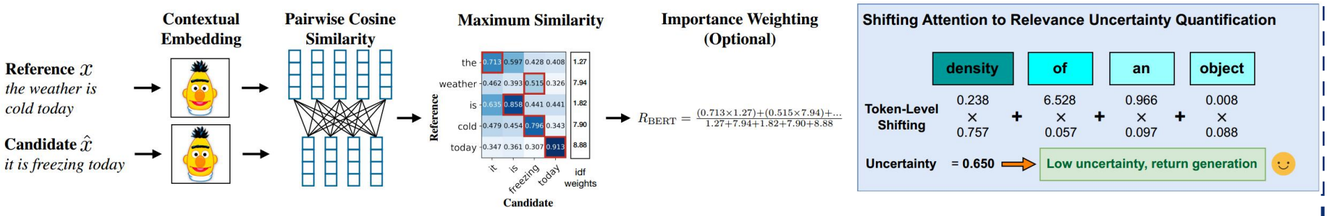
水平方向：语义定位器



单词语义重要性的不均匀分布



句子幻觉分数在重新加权之前（左侧）和之后（右侧）的分布情况。可以观察到，在重新加权后，一些不确定性较高的位置的数值趋于平坦，这是由于这些位置对句子的语义重要性较低。这表明语义定位器的主要功能是抑制那些不确定性高且语义重要性低的部分。



RBERT（左）召回率指标计算的示意图。基于语义重要性的重新加权（右）。

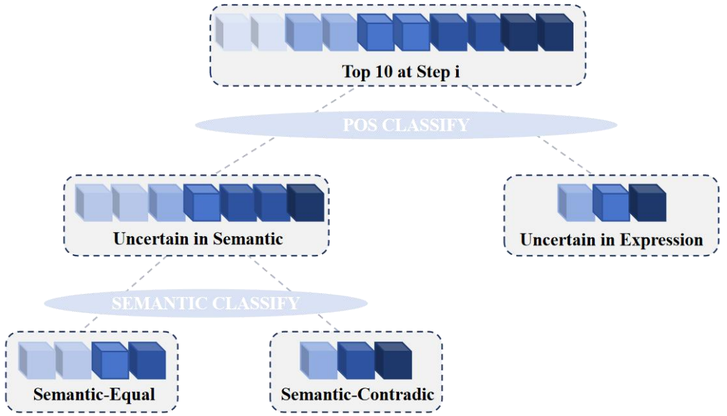
标签	名称	描述
ADJ	形容词	描述名词的特征或状态，如 "big", "happy"
ADP	介词/附加词	连接名词与句子其他部分，如 "in", "on", "at"
ADV	副词	修饰动词、形容词或其他副词，如 "quickly", "very"
AUX	助动词	辅助动词，如 "is", "have", "will"
CCONJ	并列连词	连接句子或短语，如 "and", "but", "or"
DET	限定词	指定名词的限定性，如 "the", "a", "this"

标签	名称	描述
INTJ	感叹词	表达情感、反应或呼唤，如 "wow", "ouch"
NOUN	名词	人、地点、事物或概念的名称，如 "dog", "city", "happiness"
NUM	数词	表示数量或顺序的词，如 "one", "two", "first"
PART	小品词	辅助构成语法结构的词，如 "to"（不定式标记）
PRON	代词	替代名词的词，如 "he", "she", "they"
PROPN	专有名词	特定的人名、地名或机构名，如 "John", "London"

UD语料库（Universal Dependency）具体信息

[18]Zhang T, Kishore V, Wu F, et al. Bertscore: Evaluating text generation with bert[J]. arXiv preprint arXiv:1904.09675, 2019.
Duan J, Cheng H, Wang S, et al. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024: 5050-5063.

语义信息定位器与净化器（Semantic Information Locator and Purifier）



不确定性分类决策树

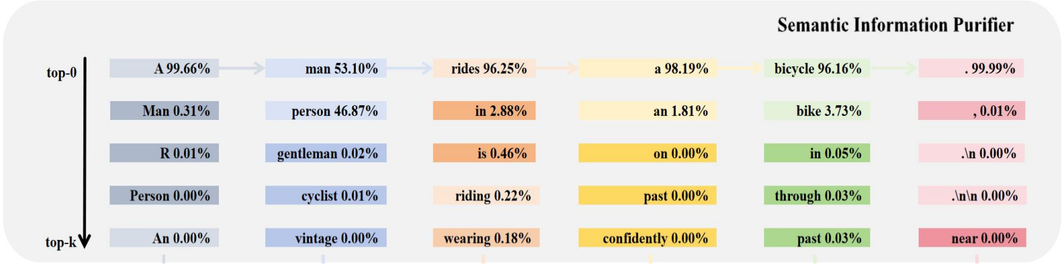
垂直方向：语义信息净化器

将模型输出的不确定性根据语义和词性分为以下两种类别：

(1)语义信息不确定性（semantic uncertainty）:**传达信息中语义上的歧义**且可能诱发幻觉。

(2) 表达不确定性（expression uncertainty）：不导致语义信息失真。仅改变句子语序或语义风格，不修改句子本质传达的信息。

01 丢弃（Discard）



对表达不确定性进行分类与丢弃

$$p(z_{i,k}|z_{i,k} \text{ is expression uncertainty}, s < i, x) = 0$$

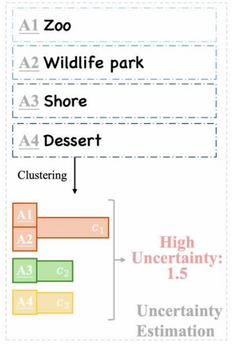
02 聚合（Cluster）

(a) Scenario 1: No semantic equivalence

Answers	Likelihood $p(s \mid x)$	Semantic likelihood $\sum_{s \in c} p(s \mid x)$
Paris	0.5	0.5
Rome	0.4	0.4
London	0.1	0.1
Entropy	0.31	0.31

(b) Scenario 2: Some semantic equivalence

Answers	Likelihood $p(s \mid x)$	Semantic likelihood $\sum_{s \in c} p(s \mid x)$
Paris	0.5	0.9
It's Paris	0.4	
London	0.1	0.1
Entropy	0.31	0.16



对相同语义的单词进行聚合

第i步生成有语义聚合集合：

$$\{e_{i,j}\}_{j=1}^{M_e}$$

对聚合后的数据进行不确定性量化：

$$\text{U\_Purified}(c, z_i) = -\sum_{j=1}^{M_e} p(e_{i,j}) \log p(e_{i,j})$$

基于多阈值决策机制的大视觉语言模型幻觉检测

用于幻觉评估的Sarah Score：

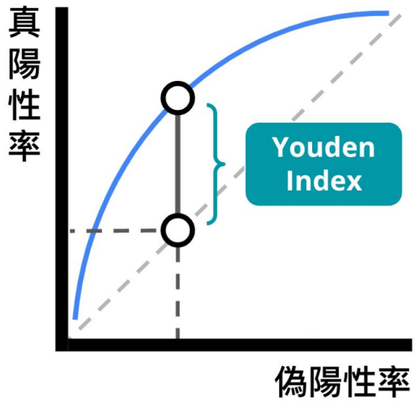
$$\text{hallucination\_score}(z_i) = \text{U\_Purified}(c, z_i) \times e^{(1-\alpha)}\text{SI}(c, z_i)$$

用于幻觉检测的多阈值决策机制：

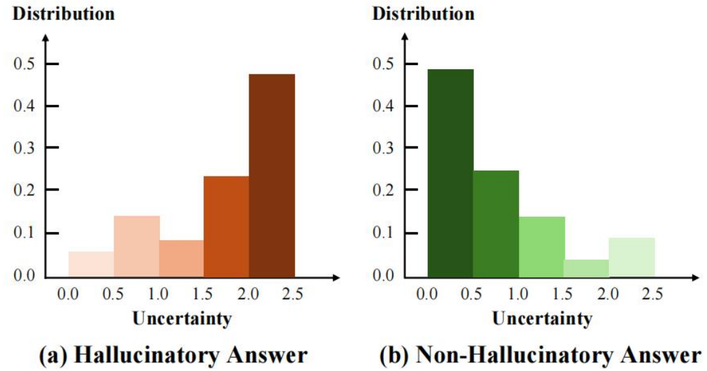
$$\text{hallucination}(c, z_i) = \text{'high' if hallucination\_score}(z_i) > \text{threshold 1 else 'low'}$$

$$\text{score}(c) = 1 \text{ if sum(high hallucinatory)} > \text{threshold 2 else 0}$$

$$\text{threshold 3(Youden Index)} = \text{TPR} + \text{TNR} - 1$$



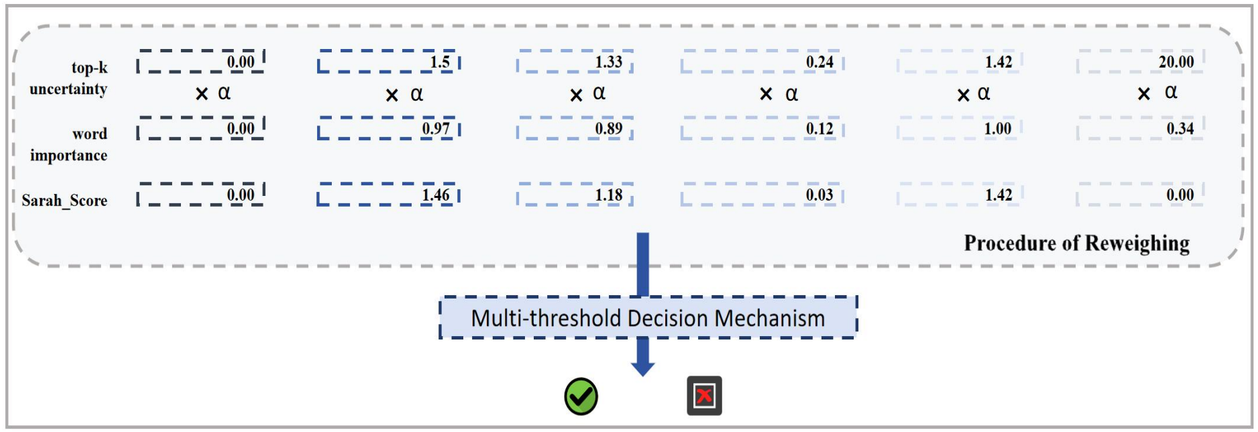
约登指数



MSCOCO-Cap数据集上幻觉（hallucinatory）和非幻觉（non-hallucinatory）语言视觉模型（LVLM）答案的不确定性分布。

合理设置阈值依次筛选高幻觉令牌、幻觉携带声明及幻觉携带句子

- **阈值1：** 识别每一步生成中的**高幻觉令牌**。
- **阈值2：** 判断声明是否**携带幻觉**
- **阈值3:**选择幻觉检测的**最优阈值**实现句子级幻觉分类



用于幻觉评估的重新加权过程和用于幻觉检测的多阈值决策机制

[29]Youden's J Statistic[J]. 2010.

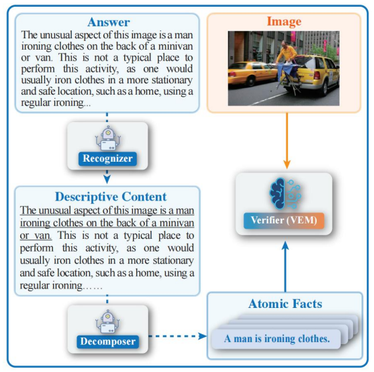
北京交通大学
BEIJING JIAOTONG UNIVERSITY

03

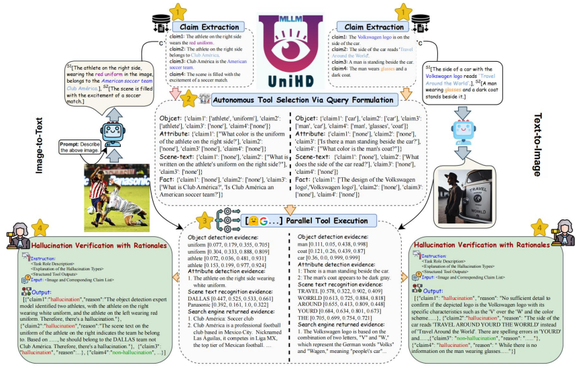
实验结果与分析

Part Three： Experiment Results and Analysis

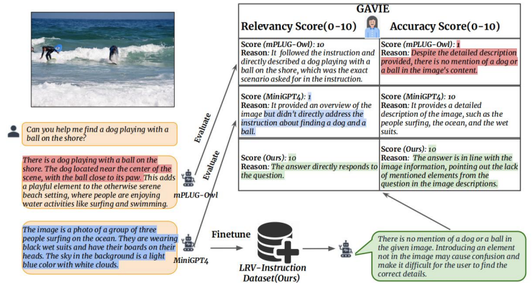
实验设置



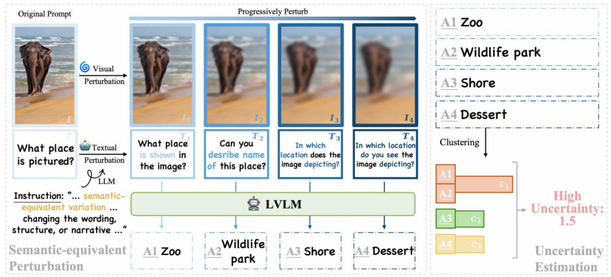
FAITHSCORE:包括三个步骤：描述性子句识别、原子事实生成以及事实验证。下划线部分表示已识别的描述性内容。



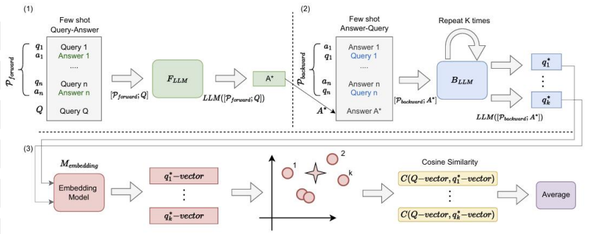
HalDet-LLaVA:UniHD能够自主选择外部工具，捕获相关知识以支持幻觉验证，并提供合理的解释。基于该框架生成海量数据训练预训练大模型LLaVA-v1.5，得到具备优秀幻觉检测能力的HalDet-LLaVA。



GAVIE:应用GPT4将带有边界框坐标的密集字幕作为图像内容，从而比较输入和输出之间的一致性。



VL-Unvertainty:通过语义等效扰动生成多样化响应，并利用响应簇的分布熵量化不确定性，识别幻觉。



InterrogateLLM:通过若干外挂的高级大语言模型基于生成内容（answer）重构查询（query），比较原始查询与基于生成答案重构的查询之间的一致性，从而识别潜在的幻觉。

Answers	Likelihood $p(\mathbf{s} \mid x)$	Semantic likelihood $\sum_{\mathbf{s} \in c} p(\mathbf{s} \mid x)$
Paris	0.5	0.9
It's Paris	0.4	
London	0.1	0.1
Entropy	0.31	0.16

$$SE(x) \approx -|C|^{-1} \sum_{i=1}^{|C|} \log p(C_i \mid x).$$

Semantic Entropy:通过将包含相同含义的句子聚集到簇中进行逐簇熵计算来解决“语义等价”困难。

选取6种最新的SOTA基线方法

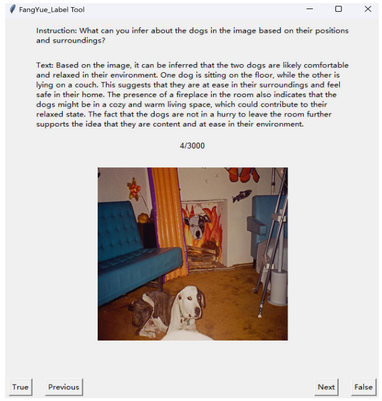
- Semantic Entropy **(SE, 2023 ICLR)**
- Faithfulness to Atomic Image Facts Score **(Faithscore, 2024 ACM)**
- GPT4-Assisted Visual Instruction Evaluation **(GAVIE, 2024 ICLR)**
- VL-Uncertainty **(the first uncertainty-based framework for detecting hallucinations in LVLMs)**
- Zero-Resource Hallucination Detection in LLM-Generated Answers **(InterrogateLLM, 2024ACL)**
- Unified Hallucination Detection for Multimodal Large Language Models **(UniHD,HalDet-LLaVA, 2024ICLR)**

选取6种最新的SOTA大视觉语言模型

- GPT-4o (“o” for “omni”)**(2024.05, OpenAI)**
- LLaMA-3.2-Vision-Instruct **(2024.09, Meta)**
- mPLUG-Owl3**(2024.08, Alibaba)**
- LLaVA-1.5 with model size 7B and 13B **(2023.10, Microsoft)**
- Qwen-2.5-3B **(2025.01, Alibaba)**

[19] Liu, Y., et al. "LLaVA: Large Language and Vision Assistant." arXiv preprint arXiv:2304.03442, 2023.
[20]Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models[J]. arXiv preprint arXiv:2302.13971, 2023.
[21]Leng S, Xing Y, Cheng Z, et al. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio[J]. arXiv preprint arXiv:2410.12787, 2024.
[22]Zhang, Y., et al. "mPLUG-Owl: Modular Vision-Language Pre-training with Open-World Learning." arXiv preprint arXiv:2305.14175, 2023.
[23]OpenAI. "GPT-4o: The Next Generation of Language Models with Enhanced Vision Capabilities." OpenAI Blog, 2023.
[24]Wang, L., et al. "Llama-v3.2-vision: Advancing Vision-Language Models with Efficient Attention Mechanisms." arXiv preprint arXiv:2307.12345, 2023.
[25]Chen X, Wang C, Zhang N, et al. Unified Hallucination Detection for Multimodal Large Language Models[C]//ICLR 2024 Workshop on Reliable and Responsible Foundation Models.
[26]Yehuda Y, Malkiel I, Barkan O, et al. InterrogateLLM: Zero-Resource Hallucination Detection in LLM-Generated Answers[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024: 9333-9347.

实验设置



自动标注工具

Data set	Accuracy of automatic evaluation
Bingo	0.856
MSCOCO-Cap	0.832

自动标注工具标注准确率

视觉语言模型中的偏见与干扰挑战数据集（Bingo）。内容包括：区域偏见、光学字符识别（OCR）偏见、事实偏见、图像与图像之间的干扰以及文本与图像之间的干扰。

数据集与基准

两个开放问答（ **open-ended question answering**）数据集:

- Bingo Benchmark
- MSCOCO-Cap

以0.15美元/组标注数据的成本雇佣人力成本
设计自动标注工具，降低人工标注成本

Max sequence length : 1024
Temperature: 0
Num_generation for GRPO: 4
Step for GRPO: 400
Step for SFT: 400
GPU:
One Intel(R) Xeon(R) Platinum 8352V CPU and four NVIDIA 4090 GPUs （most）.
Eight NVIDIA A100 GPUs （for GRPO and SFT）.
Others: default

其他参数设置

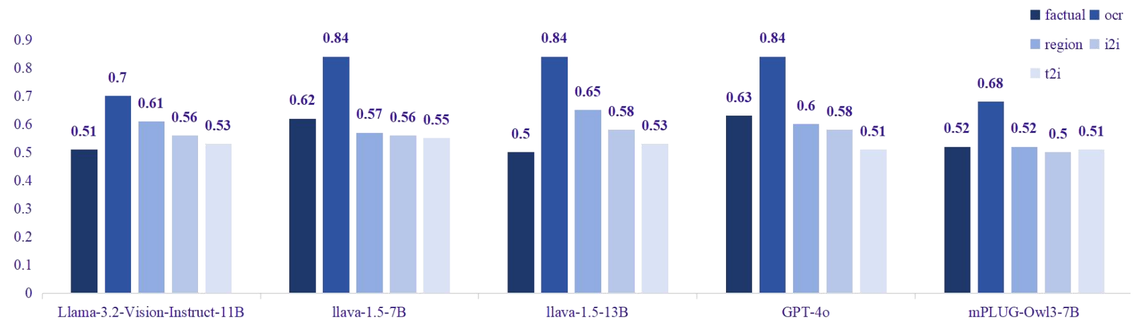
GRPO的GPU使用情况

[27]Cui C, Zhou Y, Yang X, et al. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges[J]. arXiv preprint arXiv:2311.03287, 2023.
[28]Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In ECCV, volume 8693 of Lecture Notes in Computer Science, pages 740–755. Springer.

实验结果与分析 | Sarah幻觉检测能力

Models & Datasets	GAVIE	FAITHSCORE	Semantic Entropy	InterrogateLLM	HalDet-LLaVA	VL-Uncertainty	Sarah
LLaVA-v1.5-7B							
Bingo		63.6	62.4	**66.0**	60.3	58.0	64.4
MSCOCO-Cap	56.6	61.6	54.7	**71.7**	50.2	2.7	71.3
LLaVA-v1.5-13B							
Bingo		59.5	64.7	66.8	57.2	60.4	**70.7**
MSCOCO-Cap	67.9	64.2	53.7	72.0	52.7	2.5	**72.6**
GPT-4o							
Bingo		50.8	55.2	**63.2**	37.7	59.3	61.6
MSCOCO-Cap	71.6	69.8	54.6	**86.6**	34.6	4.2	**86.6**
mPLUG-Owl3-7B							
Bingo		60.3	60.6	**71.4**	54.3	48.1	65.1
MSCOCO-Cap	63.1	60.6	53.9	**77.5**	47.0	1.7	70.5
Llama-3.2-Vision-11B							
Bingo		54.2	55.5	**67.7**	49.6	60.0	62.4
MSCOCO-Cap	59.7	68.7	52.5	**78.8**	31.8	2.5	71.2

SOTA幻觉检测技术对比，评估指标选取准确率（Accuracy）。



Sarah在Bingo基准测试中对每种幻觉类型进行了分析。这些类型包括区域偏差、光学字符识别偏差、事实偏差、图像间干扰和文本间干扰。

Method	Multi-round	Input image	Time
SE			3.65
Faithscore		*	6.02s/it
GAVIE		*	5.48s/it
InterrogateLLM	*		109.56 s/it
HalDet-LLaVA		*	4.35 s/it
VL-Uncertainty	*	*	27.96/it
Sarah			3.87

检测时间仅为InterrogateLLM的1/40！

不同幻觉检测方法的资源消耗情况。本分析评估了检测方法是否需要（1）多轮推理以及（2）图像输入，这两者都会显著影响GPU内存需求。表中还提供了每种方法的实际迭代时间。

Dataset	Image	Prompt	Method	Attention of Different Methods on Hallucinations
MSCOCO-Cap		Generate a short caption of the image.	Sarah	Someone is checking something on a device.
			SE	Someone is checking something on a device.
			GAVIE	Someone is checking something on a device.
			FaithScore	Someone is checking something on a device.
			InterrogateLLM	Someone is checking something on a device.
			HalDet-LLaVA	Someone is checking something on a device.
			VL-Uncertainty	Someone is checking something on a device.
Bingo		One of the shapes is an ellipse. What do you think?	Sarah	None of the shapes are ellipses.
			SE	None of the shapes are ellipses.
			FaithScore	None of the shapes are ellipses.
			InterrogateLLM	None of the shapes are ellipses.
			HalDet-LLaVA	None of the shapes are ellipses.
			VL-Uncertainty	None of the shapes are ellipses.

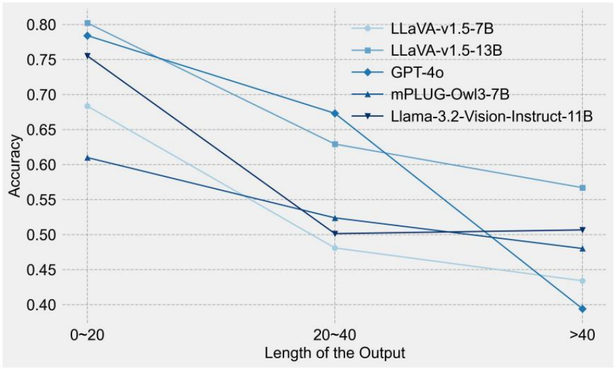
幻觉检测注意力分配模式的可视化。深色表示更高的注意力权重和更大的幻觉可能性。

实验结果与分析 | Sarah幻觉量化评估能力，其他消融实验

Model&Dataset	GAVIE	FAITHSCORE	SE	InterrogateLLM	HalDet-LLaVA	VL-Uncertainty	Sarah
LLaVA-v1.5-7B							
Bingo		45.6	56.8	53.0	50.5	58.0	**59.3**
MSCOCO-Cap	**58.7**	58.5	54.4	50.2	57.6	50.5	54.0
LLaVA-v1.5-13B							
Bingo		43.4	57.4	55.9	49.1	57.0	**63.2**
MSCOCO-Cap	59.5	**60.5**	52.3	50.3	59.0	50.4	50.9
GPT-4o							
Bingo		**59.1**	55.0	58.6	50.1	58.0	52.6
MSCOCO-Cap	59.0	59.9	57.1	48.3	53.9	50.2	**61.9**
mPLUG-Owl3-7B							
Bingo		58.0	56.6	54.5	60.3	51.1	**65.1**
MSCOCO-Cap	63.9	59.5	51.8	55.8	60.5	46.0	**70.5**
Llama-3.2-Vision-11B							
Bingo		49.4	55.6	54.5	53.6	52.0	**62.4**
MSCOCO-Cap	54.2	58.5	54.0	52.2	53.6	50.6	**60.2**

SOTA幻觉评估技术对比，评估指标选取AUC-ROC。

生成长度的影响



生成长度对Sarah幻觉检测的影响。将输出内容的长度分为三组：（0，20），（20，40）和（40，+∞）。Sarah的性能随输出长度而变化。虽然Sarah在检测短输出内容幻觉的任务上具备较高的精度，随着输出长度的增加，其性能出现下降。

模块有效性验证

Semantic Information Locator: Improved detection accuracy by **25.31%**

Semantic Information Purifier: Achieved an accuracy **35.81%** higher

Fusion Strategy: Demonstrated a detection accuracy of **79.63%**, surpassing the performance of either module in isolation.

Method	Acc.
Ours	**79.63**
Baseline	30.86
Semantic Information Locator - Only	56.17
Only Semantic Information Purifier- Only	66.67

Sarah在LLama-3.2-Vision上的消融实验结果

实验结果与分析 | 基于Sarah的大视觉语言模型幻觉率评估

Model	Bingo						MSCOCO-Cap
	Overall	Region	OCR	Factual	i2i	t2i	
LLaVA-v1.5-7B	61.3	57.9	64.0	61.6	62.9	61.3	29.0
LLaVA-v1.5-13B	59.2	64.2	54.0	54.1	65.0	58.7	26.8
GPT-4o	**38.4**	**39.8**	**16.3**	**36.7**	**44.0**	**49.3**	**13.4**
mPLUG-Owl3-7B	52.0	51.1	48.1	50.7	54.8	54.8	29.2
Llama-3.2-Vision-Instruct-11B	49.0	45.8	35.6	46.7	54.5	56.1	21.0

各大视觉语言模型幻觉表现

模型	视觉编码器	语言模型	参数规模	是否开源
LLaVA-v1.5	CLIP ViT-L/336px	Vicuna v1.5	7B/13B	是
mPLUG-OWL3	SigLIP-400M	Qwen2	7B/70B	是
GPT-4o	统一 Transformer 架构，端到端多模态处理		200B	否
LLaMA-3.2-Vision	ViT-H/14	LLaMA-3	11B/90B	是

各大视觉语言模型结构特征及参数规模

区域偏见

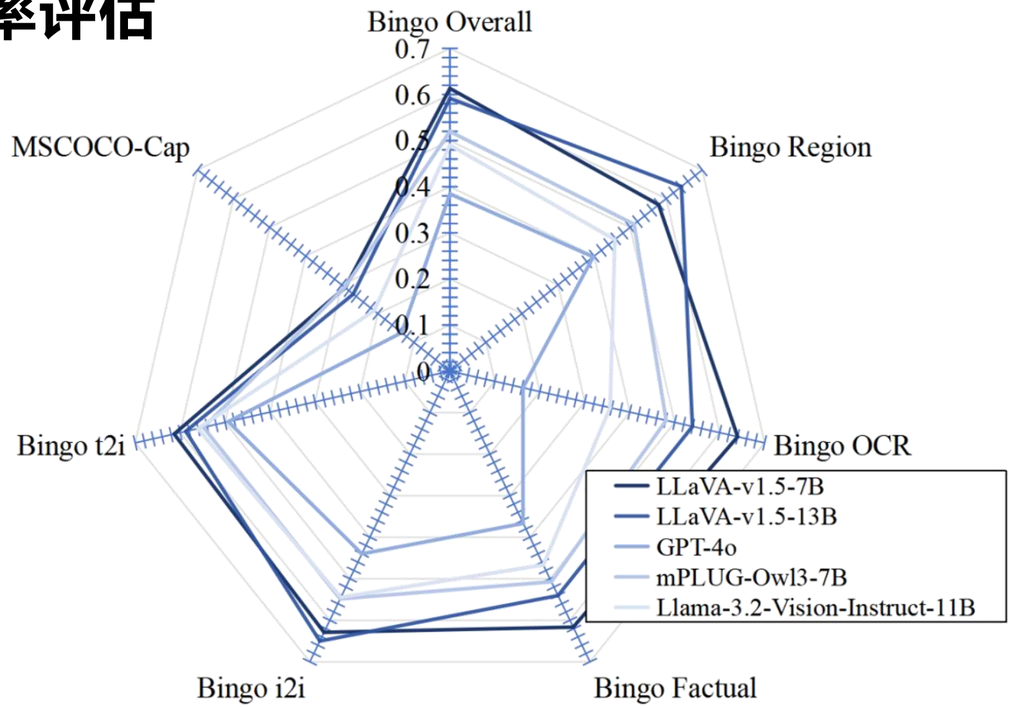
光学字符识别偏见

事实偏见

图像与图像之间的干扰

文本与图像之间的干扰

Bingo涵盖幻觉干扰类别



Models of Best Performance ：
GPT-4o,
Llama-3.2-Vision-Instruct

Tasks of Best Performance:
Image Caption,
OCR Bias

Tasks of Worst Performance:
Image to Image Interference,
Text to Image Interference

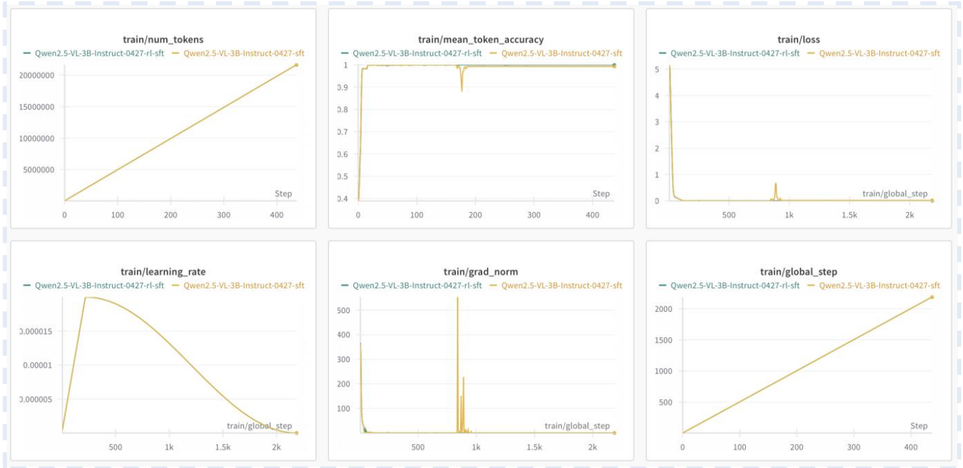
实验结果与分析 | 多模态大模型训练策略与幻觉的关系

实验设计

基线模型：Qwen2.5-VL-3B
强化微调：
https://github.com/Deep-Agent/R1-V
监督微调：
https://github.com/hiyouga/LLaMA-Factory
训练/测试数据集：
leonardPKU/clevr_cogen_a_train
OOD（Out of Distribution）评估基准：
MM-Vet（recognition, OCR, knowledge, language generation, spatial awareness, and math.）
幻觉评估方法：Sarah

	Sarah	OOD	ID
Baseline	**15.3**	**57.5**	55.5
+SFT	18.2	50.4	60.8
+RFT	24.2	53.1	87.5
+SFT+RFT	25.0	52.2	70.1
+RFT+SFT	28.7	53.2	**89.2**

模型在不同微调算法、训练顺序下的表现



Qwen2.5-VL-3B监督微调损失函数收敛情况及其他参数设置

Qwen2.5-VL-3B强化学习奖励函数收敛情况、输出长度变化及KL散度变化

专业能力（In Domain）

- RFT对模型数数能力的提升明显优于SFT（+26.7%）；
- SFT+RFT优于仅经过SFT的模型（+9.3%），但仍无法突破仅经过RFT的模型的性能；
- RF+SFT可以小幅度提升模型性能（+1.7%）。

强化微调具备更强大的激发模型专业能力的潜质。

泛化能力（Out of Distribution）

RFT和SFT均对模型通用能力造成一定影响，且SFT的干扰更大（-7.1%）。

强化微调拥有更好的泛化能力，受数据分布影响更小。

幻觉表现（Hallucination Performace）

模型在经过监督微调和强化微调后幻觉情况均有所加剧，强化微调的幻觉恶化更显著。

研究强化微调对模型幻觉的干扰至关重要。

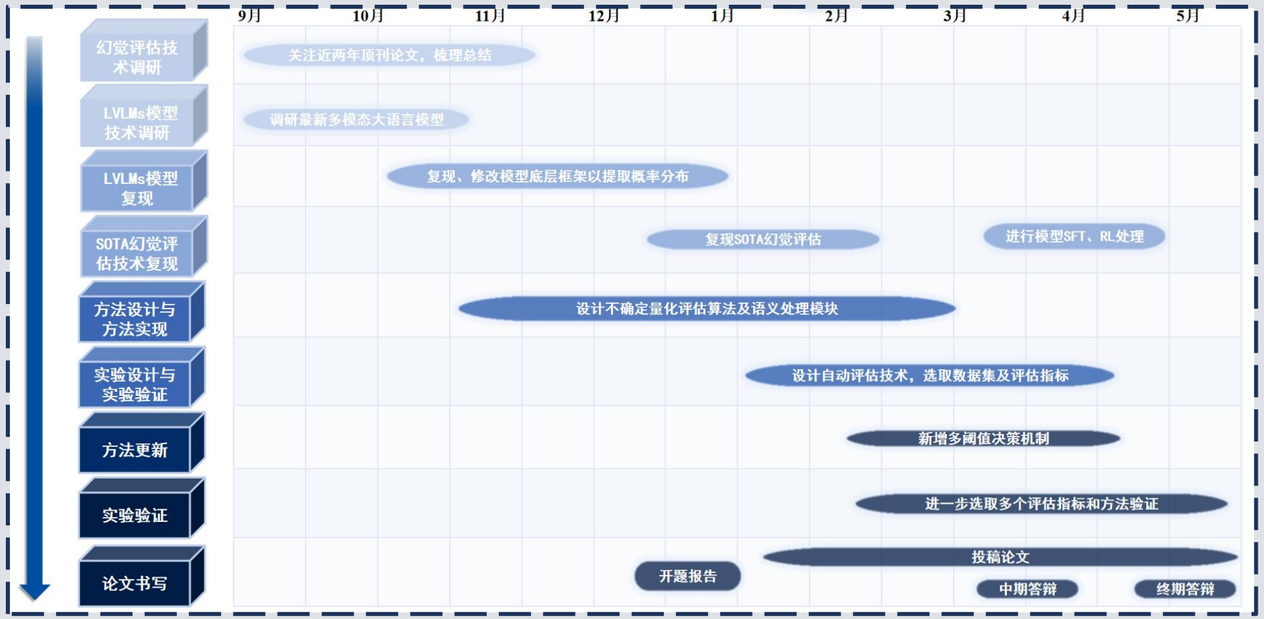
北京交通大学
BEIJING JIAOTONG UNIVERSITY

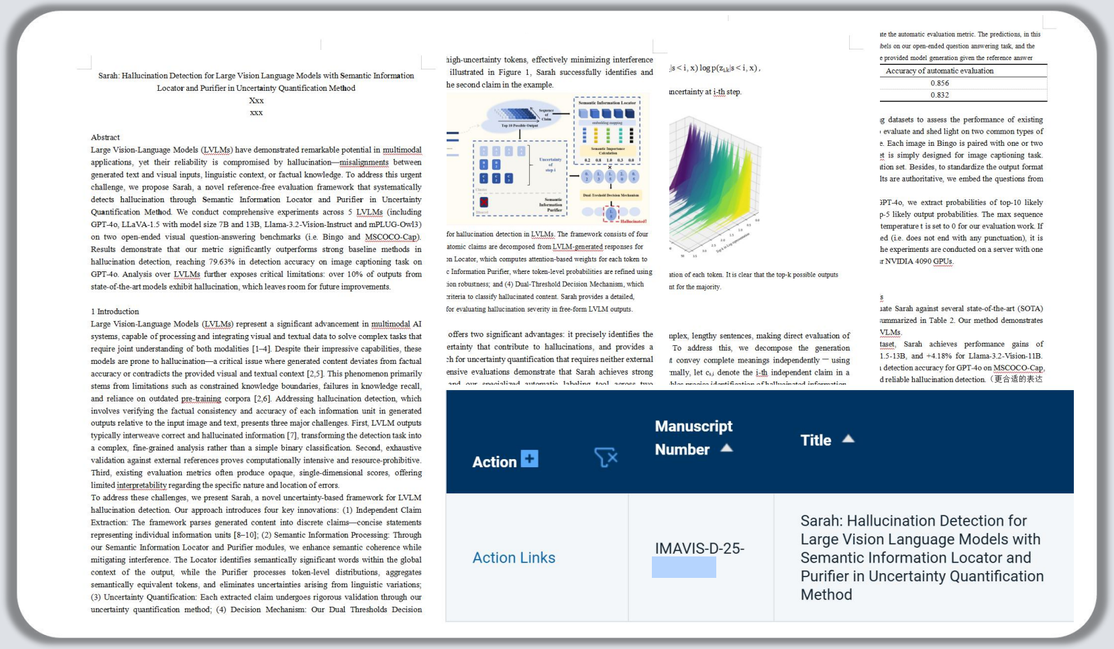
04

总结与未来展望

Part Four： Conclusion and Plan for Future

工作总结




Sarah: Hallucination Detection for Large Vision Language Models with Semantic Information Locator and Purifier in Uncertainty Quantification Method

Xxx

xxx

Abstract

Large Vision-Language Models (LVLMs) have demonstrated remarkable potential in multimodal applications, yet their reliability is compromised by hallucination—misalignments between generated text and visual inputs, linguistic context, or factual knowledge. To address this urgent challenge, we propose Sarah, a novel reference-free evaluation framework that systematically detects hallucination through Semantic Information Locator and Purifier in Uncertainty Quantification Method. We conduct comprehensive experiments across 5 LVLMs (including GPT-4o, LLaVA-1.5 with model size 7B and 13B, Llama-3.2-Vision-Instruct and mPLUG-Owl3) on two open-ended visual question-answering benchmarks (i.e. Bingo and MSCOCO-Cap). Results demonstrate that our metric significantly outperforms strong baseline methods in hallucination detection, reaching 79.63% in detection accuracy on image captioning task on GPT-4o. Analysis over LVLMs further exposes critical limitations: over 10% of outputs from state-of-the-art models exhibit hallucination, which leaves room for future improvements.

1 Introduction

Large Vision-Language Models (LVLMs) represent a significant advancement in multimodal AI systems, capable of processing and integrating visual and textual data to solve complex tasks that require joint understanding of both modalities [1–4]. Despite their impressive capabilities, these models are prone to hallucination—a critical issue where generated content deviates from factual accuracy or contradicts the provided visual and textual context [2,5]. This phenomenon primarily stems from limitations such as constrained knowledge boundaries, failures in knowledge recall, and reliance on outdated pre-training corpora [2,6]. Addressing hallucination detection, which involves verifying the factual consistency and accuracy of each information unit in generated outputs relative to the input image and text, presents three major challenges. First, LVLM outputs typically interweave correct and hallucinated information [7], transforming the detection task into a complex, fine-grained analysis rather than a simple binary classification. Second, exhaustive validation against external references proves computationally intensive and resource-prohibitive. Third, existing evaluation metrics often produce opaque, single-dimensional scores, offering limited interpretability regarding the specific nature and location of errors.

To address these challenges, we present Sarah, a novel uncertainty-based framework for LVLM hallucination detection. Our approach introduces four key innovations: (1) Independent Claim Extraction: The framework parses generated content into discrete claims—concise statements representing individual information units [8–10]; (2) Semantic Information Processing: Through our Semantic Information Locator and Purifier modules, we enhance semantic coherence while mitigating interference. The Locator identifies semantically significant words within the global context of the output, while the Purifier processes token-level distributions, aggregates semantically equivalent tokens, and eliminates uncertainties arising from linguistic variations; (3) Uncertainty Quantification: Each extracted claim undergoes rigorous validation through our uncertainty quantification method; (4) Decision Mechanism: Our Dual Thresholds Decision

high-uncertainty tokens, effectively minimizing interference illustrated in Figure 1, Sarah successfully identifies and he second claim in the example.

for hallucination detection in LVLMs. The framework consists of four atomic claims are decomposed from LVLM-generated responses for n Locator, which computes attention-based weights for each token to e Information Purifier, where token-level probabilities are refined using ion robustness, and (4) Dual-Threshold Decision Mechanism, which riteria to classify hallucinated content. Sarah provides a detailed, for evaluating hallucination severity in free-form LVLM outputs.

offers two significant advantages: it precisely identifies the ertainty that contribute to hallucinations, and provides a ch for uncertainty quantification that requires neither external ensive evaluations demonstrate that Sarah achieves strong

$|s_i < i, x) \log p(x_{ik}|s_i < i, x)$.

uncertainty at i-th step.

ation of each token. It is clear that the top-k possible outputs nt for the majority.

nplex, lengthy sentences, making direct evaluation of To address this, we decompose the generation g convey complete meanings independently — using mally; let c_{ij} denote the i-th independent claim in a

te the automatic evaluation metric. The predictions, in this bels on our open-ended question answering task, and the e provided model generation given the reference answer

Accuracy of automatic evaluation
0.856
0.832

g datasets to assess the performance of existing evaluate and shed light on two common types of e. Each image in Bingo is paired with one or two st is simply designed for image captioning task. tion set. Besides, to standardize the output format lts are authoritative, we embed the questions from

PT-4o, we extract probabilities of top-10 likely p-5 likely output probabilities. The max sequence temperature t is set to 0 for our evaluation work. If ed (i.e. does not end with any punctuation), it is he experiments are conducted on a server with one r NVIDIA 4090 GPUs.

ate Sarah against several state-of-the-art (SOTA) ummarized in Table 2. Our method demonstrates VLMs.

taset, Sarah achieves performance gains of 1.5-13B, and +4.18% for Llama-3.2-Vision-11B. detection accuracy for GPT-4o on MSCOCO-Cap, d reliable hallucination detection.（更合适的表达

- 聚焦于**多模态大模型及其幻觉问题**，以“幻觉的量化评估”为切入点，系统研究多模态大模型幻觉的定义、分类体系、检测与评估技术。
- 设计并实现一个**轻量且独立性强**的幻觉检测与量化框架 Sarah。设计细节包括：独立声明提取、语义定位与净化、不确定性量化和多阈值决策。
- Sarah在大视觉语言模型幻觉检测任务上**显著超越多数最新幻觉检测技术。**在与幻觉检测SOTA水平基本持平的同时，其检测时间仅为SOTA技术的**1/40**。
- 使用Sarah对**主流多模态大模型的幻觉表现**进行系统评估，揭示模型在不同生成条件下幻觉倾向的变化规律。
- 深入探讨**不同训练范式**对模型幻觉行为的影响。结果表明，**强化微调**相较于监督微调在提升模型域内外能力方面更具优势，但也更容易引发幻觉输出。

已完整完成毕业设计（论文）全部内容。研究成果已转化为学术论文，目前IMAVIS在投。

未来展望

进一步优化多模态大模型幻觉检测与量化评估技术

- 提升技术区分语义不确定性和表达不确定性的能力；
- 探索不同的语义处理模块的整合技术，实现1+1>2的协同效应；
- 提升幻觉检测技术在长文本生成任务中的稳定性。

探索多模态大模型幻觉缓解技术

- 将幻觉检测结果反馈至模型数据输入层，通过RAG技术或提示词工程缓解输入内容造成的幻觉；
- 将幻觉检测结果反馈至模型生成内容输出层，借助外部技术对输出内容进行纠正，缓解幻觉；
- 分析幻觉检测结果，列举模型结构和训练策略对幻觉的影响原因，改进模型结构与训练策略。

进一步优化多模态大模型训练策略，以期降低幻觉引入可能性

- 尝试在强化微调中引入事实性约束或不确定性感知机制，例如通过设计兼顾人类偏好和事实准确性的奖励函数
- 结合监督微调进行混合训练，以平衡模型的创造力和可靠性。

北京交通大學 BEIJING JIAOTONG UNIVERSITY

知 行

感谢
各位老师的倾听

北京交通大学电子信息工程学院

答辩人：方悦　指导老师：张阳