Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

AI模型计量
图增强的幻觉检测

硕士研究生 刘佳

2026年03月01日

问题回溯

- 总结反思
  - 加深对TIPO的理解
  - 论文实验结果部分需明确区分对比实验、消融实验

- 相关内容
  - 刘佳《AI幻觉陷阱与创造力》——2025.06.08
  - 杨宗源《文本生成中的幻觉》——2023.08.20

内容提要

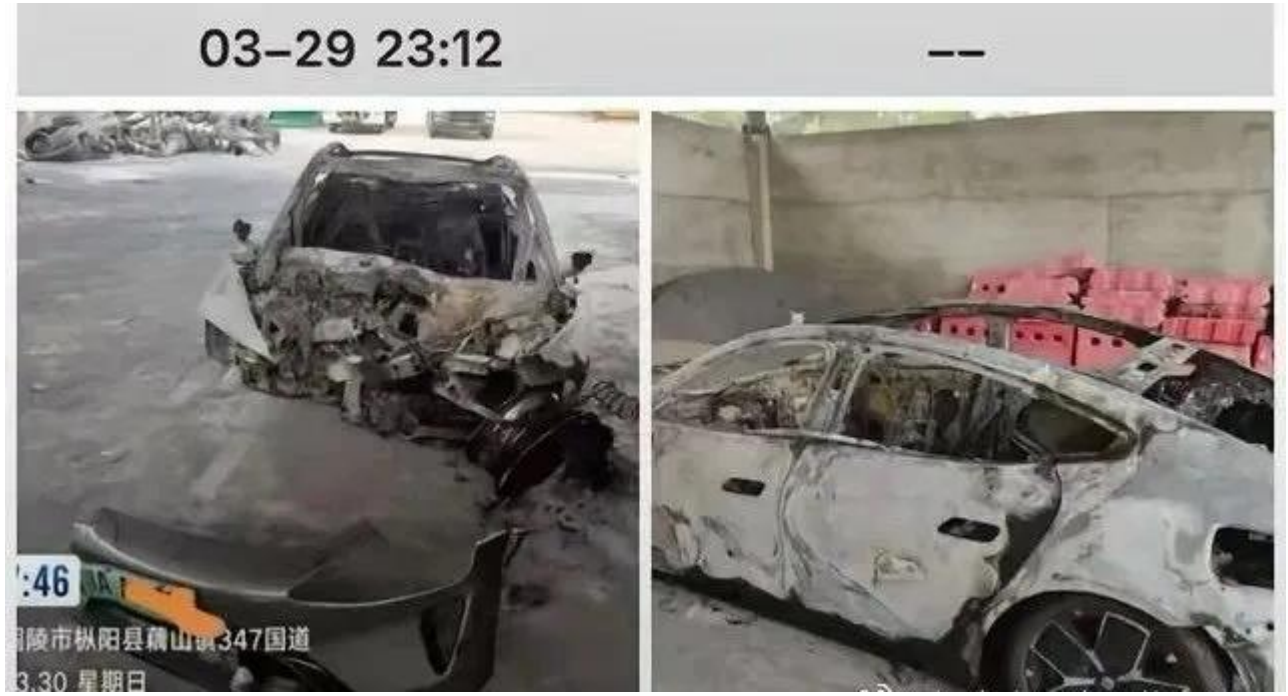
- 预期收获
- 案例/问题引入
- 人工智能模型计量内涵定义
- 研究历史与现状
- 人工智能模型计量体系
- 大模型评测方法
- 幻觉检测算法原理
  - LapEigvals
  - Semantic Graph + Uncertainty
- 特点总结与工作展望
- 参考文献

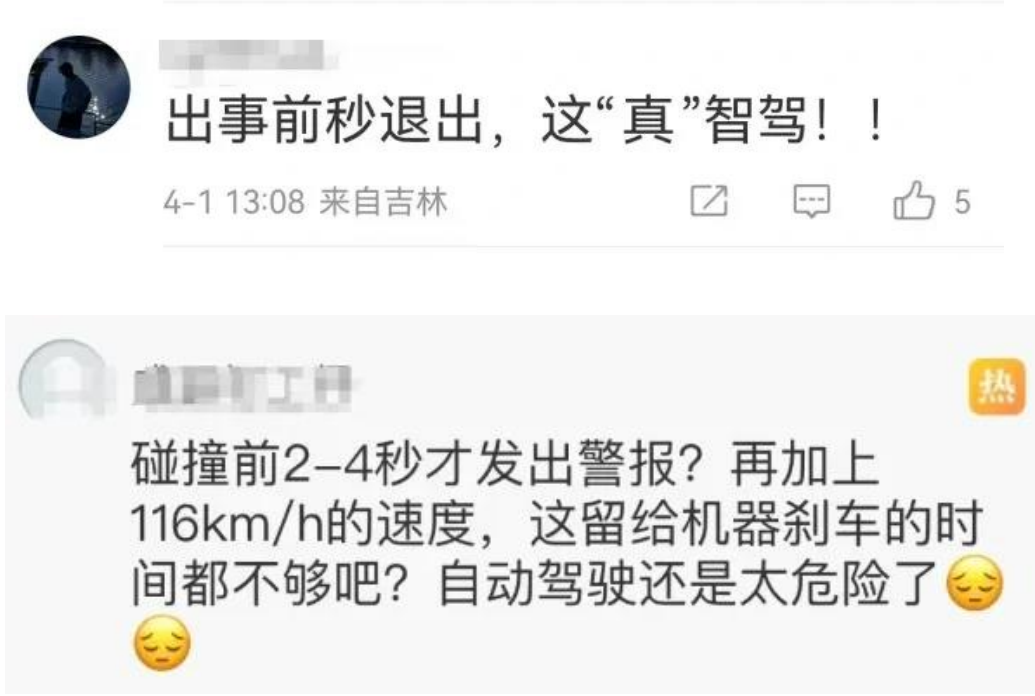
预期收获

- 预期收获
  - 1. 认识人工智能模型的计量的内涵与意义
  - 2. 了解人工智能模型计量体系
  - 3. 以AI幻觉检测为例，理解掌握幻觉检测新思路、新方法

案例引入 小米SU7高速爆燃 2025.03

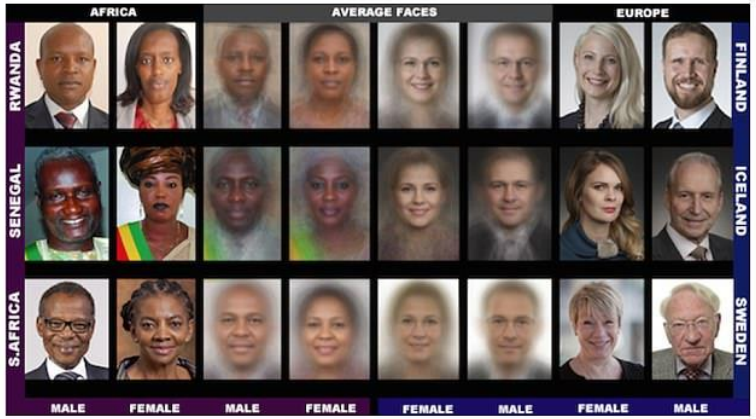
- 2025年3月29日，小米SU7高速爆燃
- 据不完全统计，小鹏汽车、问界以及特斯拉都出现过智能驾驶致人死亡的事件
- 智能驾驶：智驾与安全的天平如何平衡?





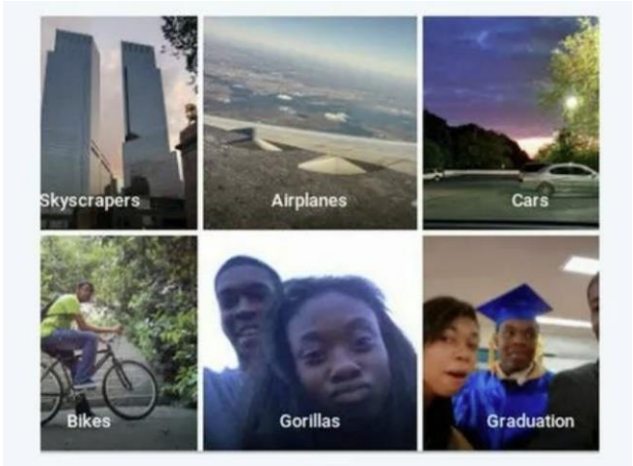
准确性、适当性

实例



某商用人脸识别对深色皮肤识别错误率高

2021年，NHTSA报告807起自动驾驶相关案件

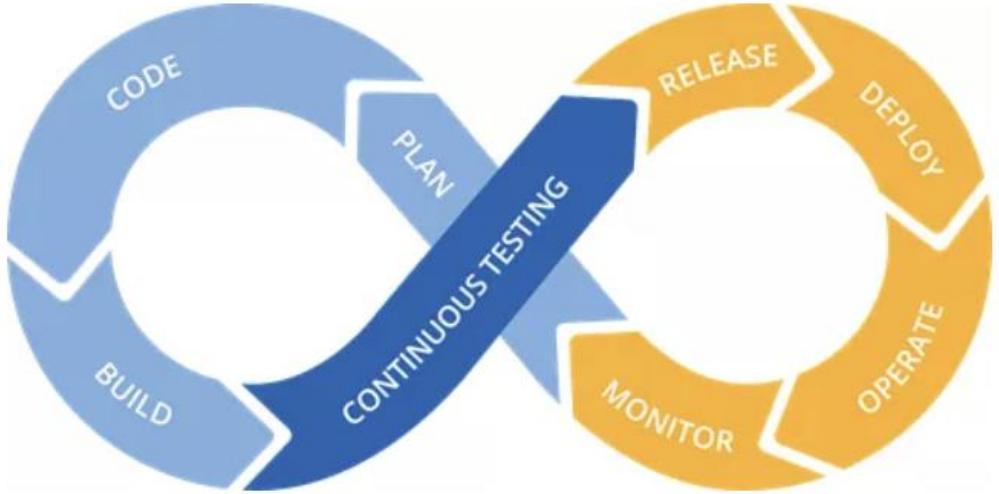


谷歌照片将黑人归类为大猩猩

2000~2013年间，机器人手术致死患者达144人

人工智能模型是否可以计量?

- “计量 (measurement / evaluation)”与传统“软件测试” — 是否适用?
  - 对于传统软件，测试可以通过确定性输入/输出 + 预期 behavior 判断对错
  - 但人工智能模型 (尤其是用学习算法训练出来的模型) 的行为往往不是完全确定
  - 人工智能模型计量：用一系列定量 / 定性指标来衡量模型表现和质量



人工智能模型需要一个更高效、持续的测试方法

人工智能模型计量 内涵定义

- 人工智能模型计量，是指围绕模型在真实应用环境中的表现、风险与运行特性，构建系统化指标体系与规范化评测方法，对模型进行多维度、可量化、可比较的综合评价过程
- 模型计量强调从“算法评价”走向“工程质量度量”
  - 对模型能力与风险的系统化刻画，而非单点性能评估
  - 对模型表现的结构化表达，而非零散指标罗列
  - 对模型状态的可追溯量化，而非主观经验判断
  - 对模型应用的决策支撑能力，而非仅作为研究结果展示

连接“算法性能”与“工程可信应用”的关键枢纽
推动人工智能从“能用”走向“可控、可审计、可监管”的核心基础

人工智能模型计量 重要性

- AI 模型在金融、医疗、工业控制等关键业务中被广泛使用
- 模型训练依赖数据，内部机制不可见，传统软件测试方法无法覆盖
- 不同算法架构、训练方式差异大，模型能力难以横向对比
- 工程部署迫切需要可量化的质量保证：
  - 是否稳定?
  - 是否鲁棒?
  - 是否公平?
  - 是否可解释?
  - ……

构建统一的模型计量体系是提升 AI 可信度与工程化程度的基础

研究历史与现状

2008

Murphy等人对机器学习系统测试做出研究，提出变形测试/蜕变测试来解决测试预言问题，并首次提出了六种变形关系，尽管他们并未进行评估实验，但这项工作确定了变形测试可应用于人工智能系统，也正式揭开了智能化系统测试的序幕

2017

Kexin Pei等人提出DeepXplore，AI 测试开始形成“可操作的白盒框架”：用差分测试把多模型分歧当作弱预言，并提出neuron coverage作为“测试充分性”的近似度量，把传统软件测试里的“覆盖—发现缺陷”路径迁移到深度网络上，标志着深度学习测试从“只看准确率”走向“系统化找角落错误”。

2018

Lei Ma等人提出 DeepGauge，把覆盖从单一神经元扩展到多粒度准则，推动“充分性度量”体系化；同年DeepMutation，把传统软件的突变测试引入深度学习，通过对训练数据/训练程序/模型进行突变来衡量测试集质量与测试策略强度，使 AI 测试从“覆盖指标”进一步走向“覆盖/突变体杀伤能力”这样的有效性评估。

2019

面向真实缺陷发现能力，研究开始把“模糊测试（fuzzing）+ 覆盖引导”系统化地用于 DNN，典型如 Xiaofei Xie等在 ISSTA 提出的 DeepHunter，强调用覆盖反馈驱动输入变异来触发潜在错误行为，把 AI 测试从静态数据集评估进一步推向“自动化生成—覆盖引导—缺陷挖掘”的工程形态。

2020

大模型/基础模型的“能力计量”主线开始明确，Dan Hendrycks 等提出 MMLU，用跨57学科任务的统一测试来衡量模型的广度知识与推理能力，使“通用能力基准”成为模型计量的重要载体，也让评测从单任务指标转向跨任务、跨领域的综合能力刻画。

2022

在能力基准扩张的同时，计量开始强调多维度透明评测，Aarohi Srivastava 等发布 BIG-bench，以大规模众包任务覆盖更广能力空间并研究规模带来的涌现与校准问题；Percy Liang 等提出 HELM，系统化梳理场景与指标，把准确性之外的风险与属性（如校准、鲁棒、公平、毒性、效率等）纳入统一框架，推动 AI 模型测试从“排行榜式能力分数”走向“场景×指标”的多维计量。

2024以后

生成式AI的系统风险成为评测重点，OpenCompass 等通用评测平台生态在这一阶段更强调多模型、多数据集、可复现流水线与统一报告输出，评测对象也从“模型本体”扩展到“模型+检索（RAG）+工具链+代理”的系统级链路。

人工智能模型计量 指标体系

- 一级维度 6 类，二级指标 22 个，三级指标 31 个

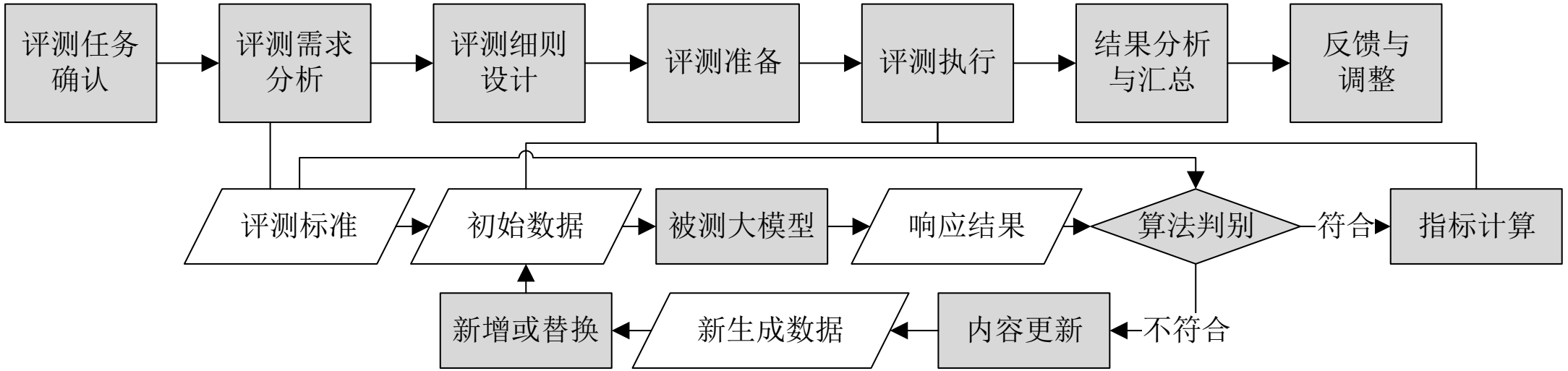
一级指标	二级指标	典型三级度量
功能性	准确性、完整性、适当性	Accuracy、F1、RMSE、MAE、预测覆盖率、输出字段缺失率、输出有效性
行为特性	数据驱动性、不确定性、经验依赖 、演化性	输入分布变化敏感性、不确定性校准误差、版本漂移指标、随机性敏感性
可信性	鲁棒性 、公平性 、可解释性、一致性、可审计性	对抗攻击成功率、噪声下降率、Demographic Parity、EO、SHAP 稳定性、解释覆盖度、多次推理相似度、日志完整度
性能效率	推理性能、资源使用、计算复杂度	Latency、Throughput、Parameter Count、Model Size、FLOPs、GPU peak memory
可靠性	稳定性、错误耐受性、长期一致性	噪声扰动稳定性（ΔAcc）、异常输入失败率
可维护性	可分析性、可修改性、可复用性、可测试性	特征重要度分析难度、微调效率、迁移学习性能、单元测试可覆盖性

大模型评测方法 基于规模验证的数据评测方法

- 结果正确性
  - 故障率、响应率、信息泄露事故率、信息泄露量、测试成功率、数据获取准确率、违规内容检出率、违规内容告警率、基础知识掌握、关系推断准确率、类型划分准确率、推理准确性、反事实率、领域迁移/转化成功率、数据泛化成功率
- 内容相似性
  - 政治反动内容比率、违法违纪内容比率、侵权危害内容比率、语义重合度、实体概念判别、领域实体认知率、平均误差距离
- 功能验证性
  - 攻击域、语言种类、领域性能差异

大模型评测方法 基于算法分析的指标评测方法

- 设计算法与动态测试框架，对模型表现进行量化诊断
- 核心特征是能够与被测目标产生实时交互，并动态调整其测试策略
  - 样本生成算法：幻觉率、攻击检出率、攻击失效率
  - 调度控制算法：最大并发用户数、最大吞吐量、最大接受输入、不确定性、接口覆盖率、功能覆盖率、场景覆盖率、字符容错率、语序容错率、实体覆盖率、问题覆盖率、历史数据重合度



大模型评测方法 基于主观判断的人工评测方法

- 直接评估
  - 组件可分离占比、黑白盒状态分级、实例分析占比、推理层数、推理偏移占比、采纳占比、有效更新比例、错误检测率、错误修正率
- 成对比较
  - 场景解读准确率、资源检索能力、资源整合能力
- 等级排序
  - 信息泄露危害程度、开源社区活跃度、维修难度、可用性评分、可操作性评分、语义相似性、结构相似性、可视化程度
- 组合标准
  - 用户满意度、可视化分级、日志完整性分数、可解释性分数、多样性分数、创新性分数、修正满意度、人类一致性分数

大模型评测方法 基于裁判模型的自动评测方法

- 直接评估
  - 输出一致率、条件复现率、内容定位粒度、实体跳转次数、*推理层数、*推理偏移占比、*错误检测率、*错误修正率
- 成对比较
  - *场景解读准确率、*资源检索能力、*资源整合能力、输入输出一致性分数
- 等级排序
  - 模型开源级别、文档完整程度、领域知识系统性、*语义相似性、*结构相似性
- 组合标准
  - 测试资源复现率、语义一致分数、主题一致分数、*多样性分数、*创新性分数、内容匹配率

以上带“*”指标表示先使用大模型初评，再由人工复核的评测指标

大模型评测方法 基于参数监控的资源评测方法

- 时效类
  - 平均故障间隔时间、平均等待时间、使用可用性、任务前准备时间、再次出动准备时间、更新周期、数据时效性、平均保障延误时间、平均管理延误时间、平均同样故障时间、平均修复时间、平均预防性修复时间、重要零部件平均更换时间
- 能耗类
  - CPU占用、GPU占用、内存占用、自然资源占用、资源占用监控曲线
- 合规项
  - 最小操作流程、接口可达程度、测试接口数量、测试接口类型
- 使用率
  - 保障设备利用率、维修时模型功能占用率、思考过程时间占比

Hallucination Detection in LLMs Using Spectral Features of Attention Maps

算法原理 TIPO

T	目标	利用模型注意力图的谱特征检测LLM输出是否含幻觉
I	输入	用户提示词 + LLM 生成回答；各层各头的attention maps
P	处理	1.输入拼接： prompt 与 response 拼成序列 2.特征提取：把 attention map 视为图的邻接矩阵，构造 Laplacian 3.得分计算：提取 LapEigvals（对角项排序+top-k），跨层/跨头拼接，PCA 降维后用 logistic regression probe 输出幻觉概率 4.幻觉判断：概率超过阈值判为幻觉
O	输出	布尔值 True/False（是否幻觉）或对应的幻觉概率分数

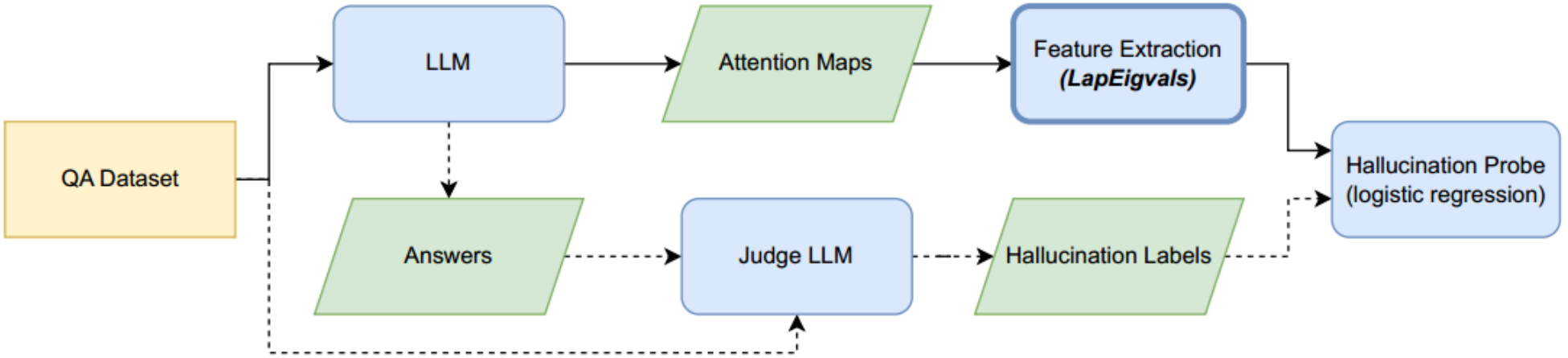
P	问题	仅用attention map 提取更稳健的结构信号，实现低成本幻觉检测
C	条件	能拿到生成过程的 attention maps（跨层/跨头） 模型为 Transformer 自回归推理 训练阶段用带标注的 QA 数据构造监督标签，探针为轻量分类器
D	难点	单响应、谱特征需要具备跨模型泛化、特征提取的稳定性与有效性
L	水平	EMNLP 2025 CCF B

算法原理 创新说明

- **LapEigvals**
  - **attention统计 → 图谱谱特征**
    - **把多层多头的 attention map 视作 token 图的邻接矩阵，引入拉普拉斯结构表征，并提取拉普拉斯谱特征值的 top-k 并跨层跨头拼接作为幻觉检测的核心特征，从根本上区别于常见的 attention 均值/熵/集中度等统计量路线**
  - **先证特征有效再训练探针**
    - **通过统计分析和显著性对比证明谱特征对“幻觉/非幻觉”更可分，再用轻量 probe（PCA+logistic regression）完成检测**
  - **单次、无外部知识、低开销**
    - **推理阶段只需一次 forward 获取 attention maps，不依赖外部知识库/检索，也不需要多次采样多响应**

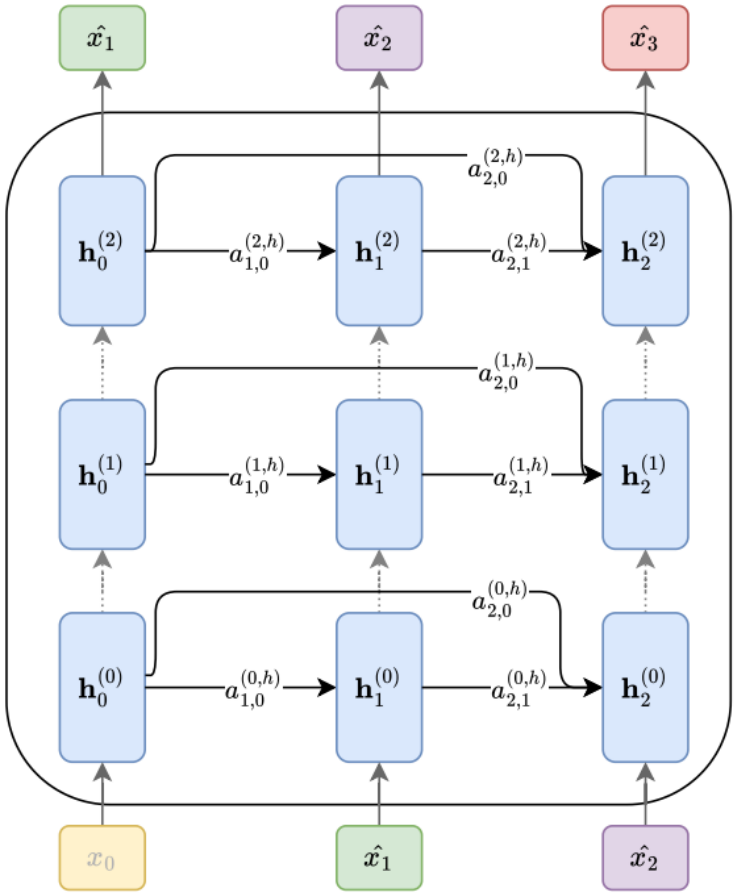
算法原理 LapEigvals

- **LapEigvals**
  - **把 LLM 推理时的注意力矩阵看成一张 token 图的邻接矩阵**
  - **构造拉普拉斯矩阵（Laplacian）**
  - **提取其 top-k 最大“谱特征”（等价于拉普拉斯的对角项并排序得到的特征值序列）跨层跨头拼接成向量**
  - **PCA 降维**
  - **logistic regression 探针输出“是否幻觉”的概率/标签**



算法原理 LapEigvals

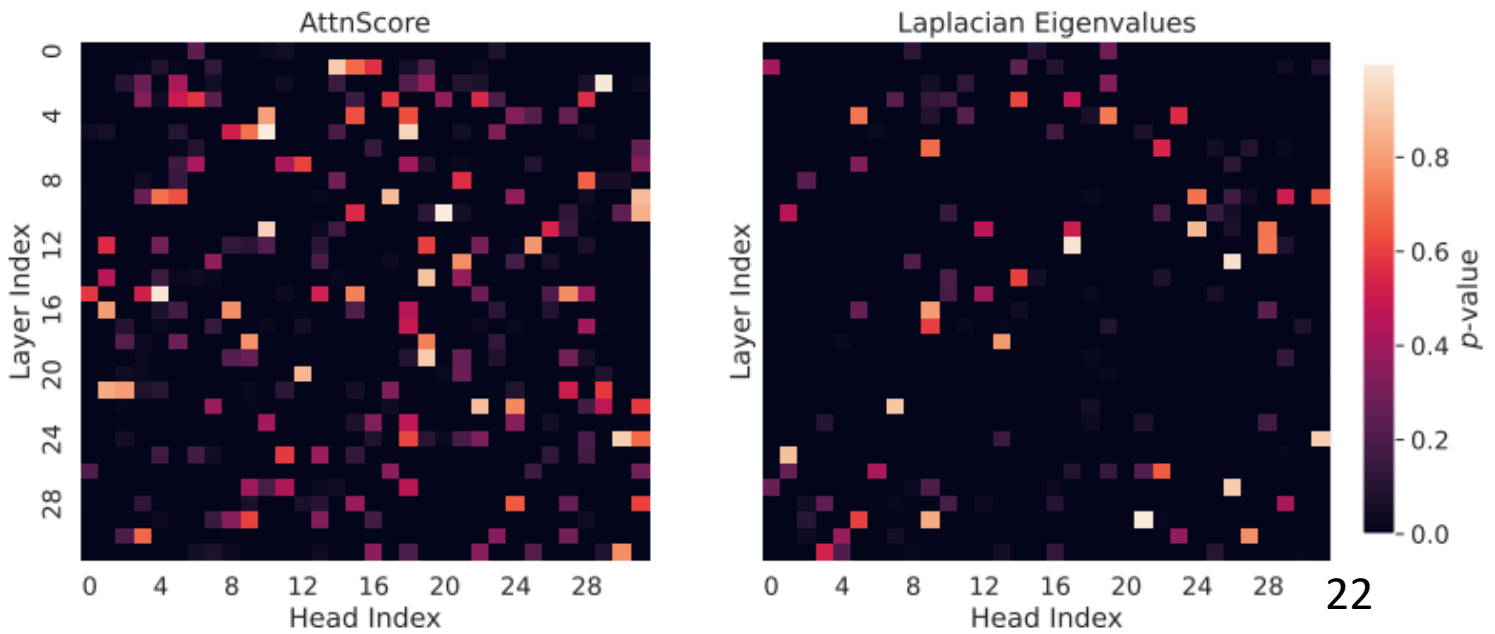
- 获取注意力矩阵（原始观测量）
  - 幻觉与否会反映在模型内部的信息路由模式上，而 Transformer 的注意力矩阵就是信息路由的直接观测
  - 对输入 prompt 生成一次 response，同时读取每层每头的注意力矩阵
- 把注意力矩阵解释为图的邻接矩阵
  - 把 token 看作节点，把注意力权重看作边权，则注意力矩阵 $A(l,h)$ 就是一个有向加权图的邻接矩阵
  - 幻觉不是单点噪声，而是信息流结构异常
    - 例如关键 token 的引用/支撑链断裂、注意力过度集中或漂移
  - 图结构比简单统计量更能刻画这种异常



算法原理 LapEigvals

- 构造拉普拉斯矩阵
  - $L^{(l,h)} = D^{(l,h)} - A^{(l,h)}$，为刻画图结构的基础算子，把结构信息编码进矩阵
  - D 为对角度矩阵（出度矩阵）反映该 token 被关注强度
- 用“谱特征”刻画结构差异
- Top-k 提纯 + 跨层跨头聚合
- PCA 降维
- 轻量探针分类

Llama-3.1-8B的 TriviaQA 实验相比 AttentionScore 拉普拉斯特征值在更多头部上对幻觉/非幻觉差异显著



实验设计　数据资源

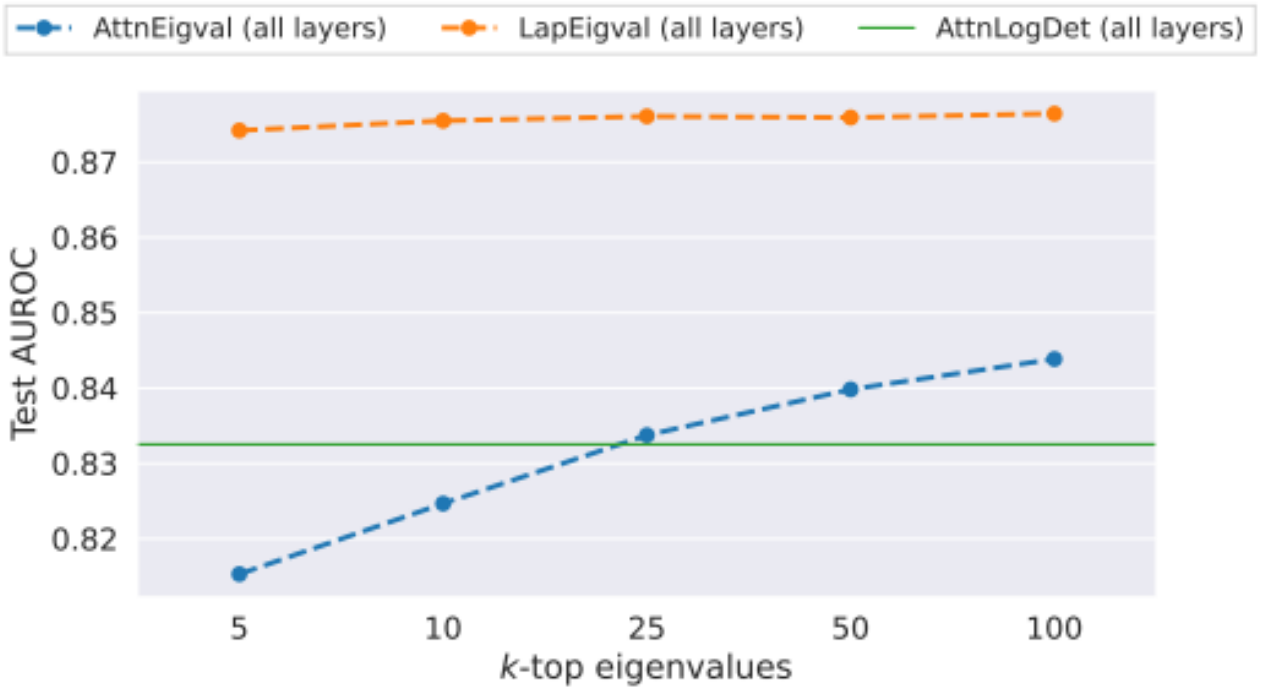
- 数据资源
  - NQ-Open (NQOpen)、SQuAD v2 (SquadV2)、TruthfulQA、HaluEvalQA、CoQA、TriviaQA、GSM8K
- 开源 LLM
  - Llama-3.1-8B-Instruct、Llama-3.2-3B-Instruct、Phi-3.5-mini-instruct、Mistral-Nemo-Instruct-2407、Mistral-Small-24B-Instruct-2501
- 对比方法
  - AttentionScore（ Sriramanan et al., 2024，无监督基线分数）
  - AttnLogDet（把 AttentionScore 的 log-det 作为监督）
  - AttnEigvals（原始注意力矩阵的特征值）

实验设计 对比实验

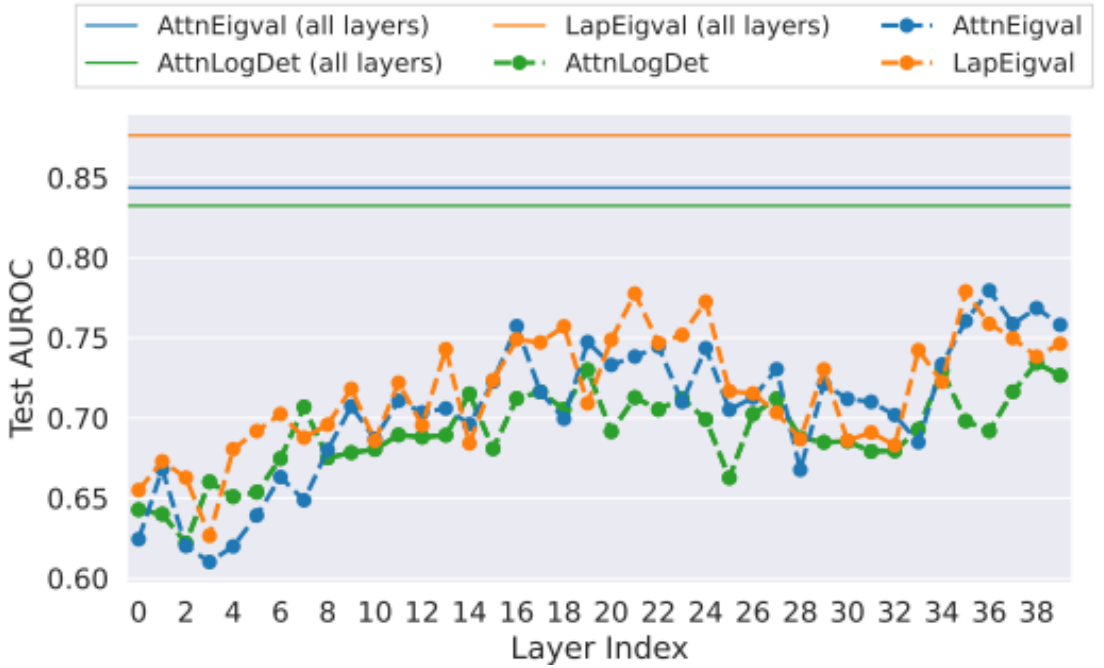
LLM	Feature	Test AUROC (↑)						
		CoQA	GSM8K	HaluevalQA	NQOpen	SQuADv2	TriviaQA	TruthfulQA
Llama3.1-8B	AttentionScore	0.493	0.720	0.589	0.556	0.538	0.532	0.541
Llama3.1-8B	AttnLogDet	0.769	0.826	0.827	0.793	0.748	0.842	0.814
Llama3.1-8B	AttnEigvals	0.782	0.838	0.819	0.790	0.768	0.843	**0.833**
Llama3.1-8B	LapEigvals	**0.830**	**0.872**	**0.874**	**0.827**	**0.791**	**0.889**	0.829
Llama3.2-3B	AttentionScore	0.509	0.717	0.588	0.546	0.530	0.515	0.581
Llama3.2-3B	AttnLogDet	0.700	0.851	0.801	0.690	0.734	0.789	**0.795**
Llama3.2-3B	AttnEigvals	0.724	0.768	0.819	**0.694**	0.749	0.804	0.723
Llama3.2-3B	LapEigvals	**0.812**	**0.870**	**0.828**	0.693	**0.757**	**0.832**	0.787
Phi3.5	AttentionScore	0.520	0.666	0.541	0.594	0.504	0.540	0.554
Phi3.5	AttnLogDet	0.745	0.842	0.818	0.815	0.769	0.848	0.755
Phi3.5	AttnEigvals	0.771	0.794	0.829	0.798	0.782	0.850	**0.802**
Phi3.5	LapEigvals	**0.821**	**0.885**	**0.836**	**0.826**	**0.795**	**0.872**	0.777
Mistral-Nemo	AttentionScore	0.493	0.630	0.531	0.529	0.510	0.532	0.494
Mistral-Nemo	AttnLogDet	0.728	0.856	0.798	0.769	0.772	0.812	**0.852**
Mistral-Nemo	AttnEigvals	0.778	0.842	0.781	0.761	0.758	0.821	0.802
Mistral-Nemo	LapEigvals	**0.835**	**0.890**	**0.833**	**0.795**	**0.812**	**0.865**	0.828
Mistral-Small-24B	AttentionScore	0.516	0.576	0.504	0.462	0.455	0.463	0.451
Mistral-Small-24B	AttnLogDet	0.766	0.853	0.842	0.747	0.753	0.833	0.735
Mistral-Small-24B	AttnEigvals	0.805	0.856	0.848	0.751	0.760	0.844	**0.765**
Mistral-Small-24B	LapEigvals	**0.861**	**0.925**	**0.882**	**0.791**	**0.820**	**0.876**	0.748

TruthfulQA 的结果可能受数据量小/类别极不平衡影响

实验设计　超参数实验



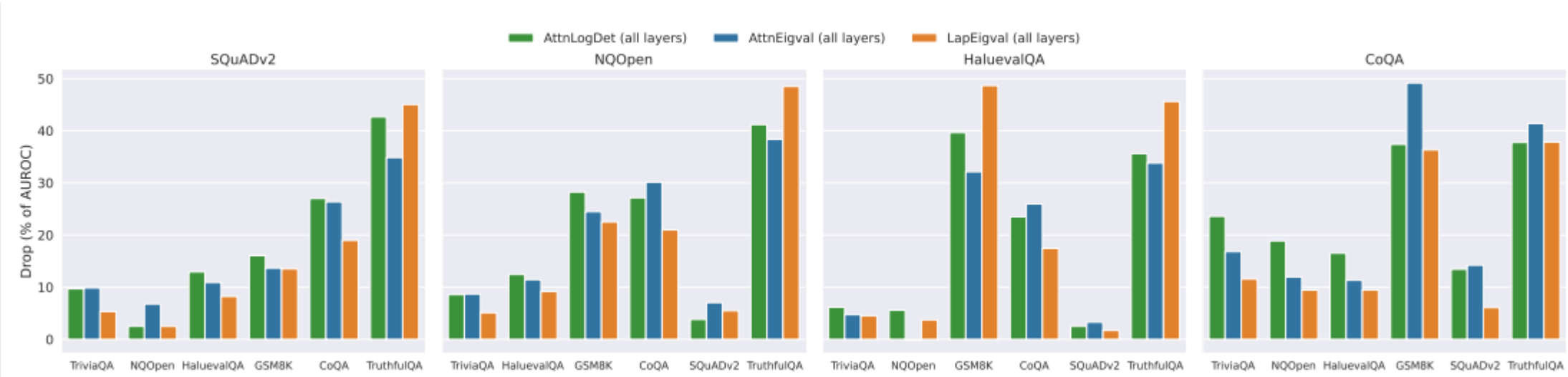
k（取多少个特征值）影响
k 越大越好（信息更多）



单层 vs 全层
all layers 拼接的效果
显著优于只用单层

算法总结

- 算法贡献
  - 提出一种基于注意力谱特征的幻觉检测新路线
  - 跨层×跨头聚合增强稳定性
  - 轻量探针，便于说明特征本身有效，而不是靠重模型拟合
- 算法不足
  - 方法本质是白盒/半白盒
  - 训练阶段依赖 llm-as-judge 生成幻觉标签来监督训练 probe



Enhancing Uncertainty Modeling with Semantic Graph for Hallucination Detection

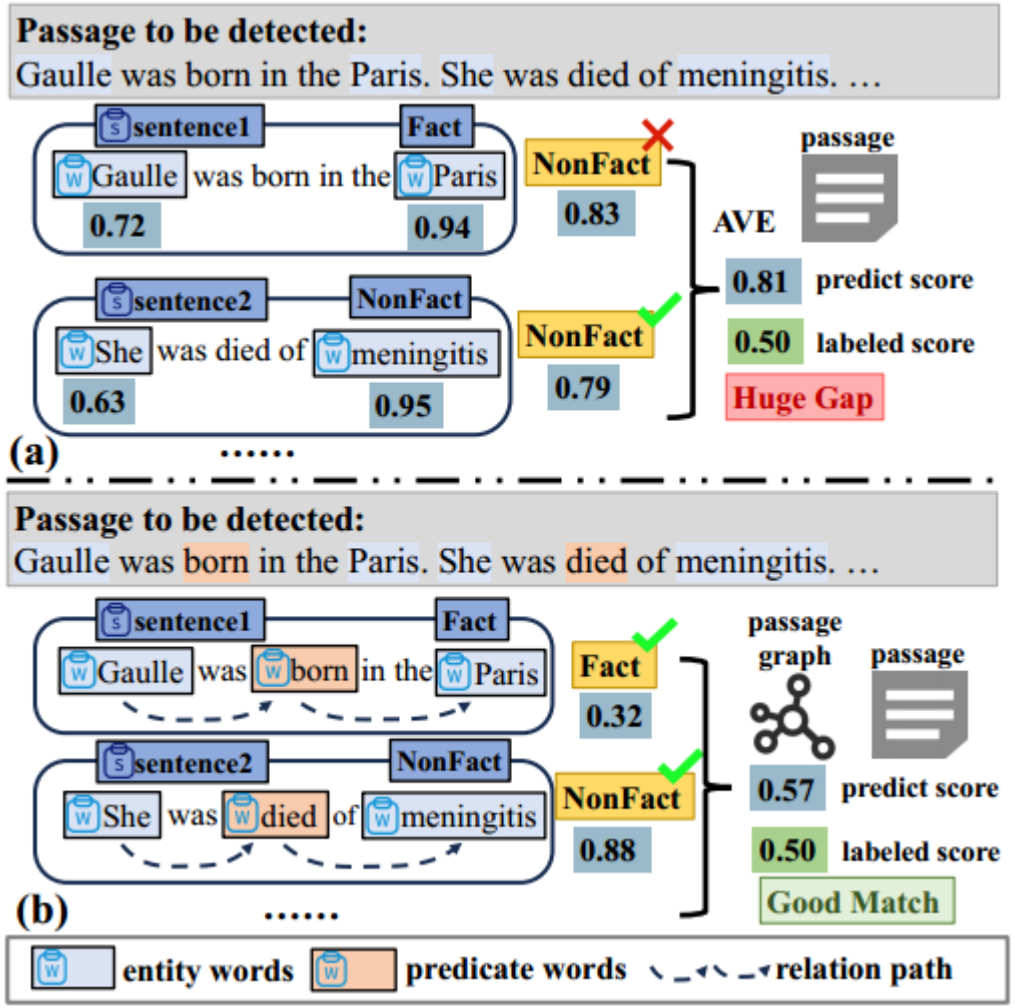
算法原理 TIPO

T	目标	用语义图增强不确定性建模，在句子/篇章层面更稳地检测幻觉
I	输入	用户问题 + LLM 生成回答；代理模型的 token 概率；AMR/指代/实体链接结果；句间矛盾分数
P	处理	1. 由 token 概率分布计算局部不确定性 2. AMR 得句内结构，指代消解/实体链接连成 passage 语义图 3. 实体相关路径传播得 UE，句内分位数聚合得 UG，线性融合成 Us 4. 在语义图邻域内用 NLI 矛盾概率对句子分数做校准/加权汇总 5.对句子或篇章给出幻觉分数并阈值化
O	输出	句子级/篇章级幻觉分数（或 True/False），并可定位高风险句子

P	问题	token 级不确定性难以覆盖跨句依赖与语义冲突，句子/篇章层面易过估计或漏检
C	条件	需要可控代理 LLM； AMR + 指代/实体链接
D	难点	语义图质量误差会传导；传播范围与噪声抑制权衡
L	水平	AAAI 2025 CCF A

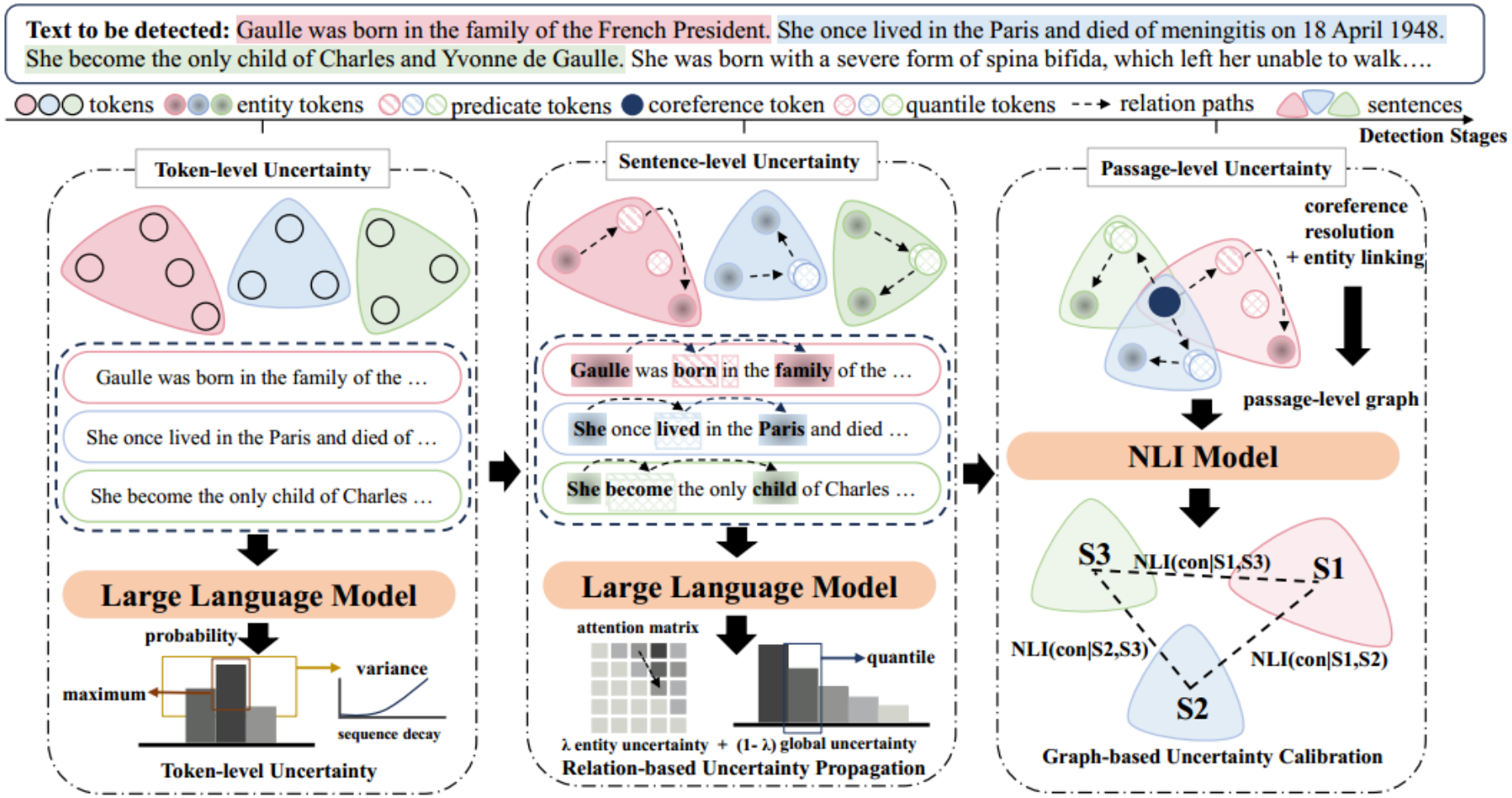
算法原理 创新说明

- 语义图增强不确定性建模的幻觉检测
  - 语义图驱动的不确定性传播：把回答从“token 序列”提升为“跨句语义图”，只沿实体相关的关系路径传播不确定性
  - 实体相关不确定性（UE）+ 句内全局不确定性（UG）
  - 用语义邻域的矛盾信号修正分数：引入基于 NLI 矛盾概率的邻域校准



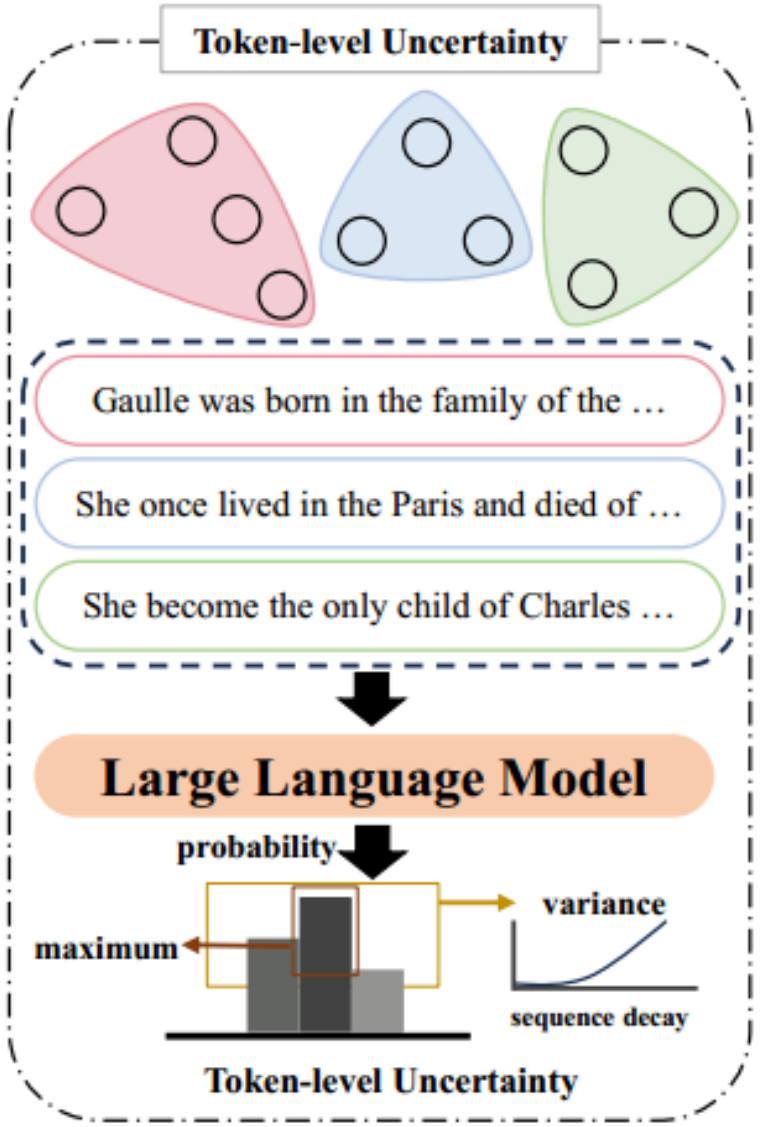
算法原理

幻觉往往不是某一个 token 错误，而是句子里实体关系错误/跨句自相矛盾



算法原理

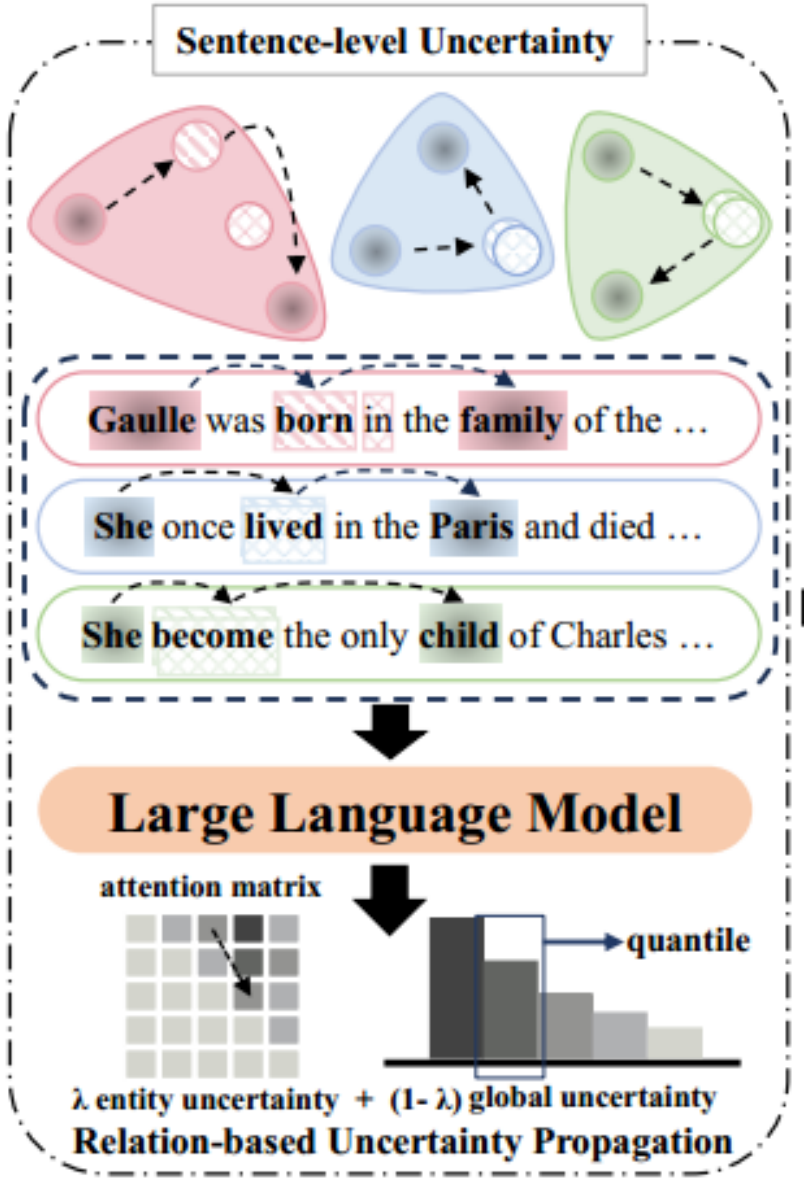
- Token 级不确定性
  - $U(t_j^i) = \frac{1}{max(C_j^i)+\sigma^2(C_j^i)} \cdot (1 + e^{\frac{len(S_{1:i-1})+j}{len(D)-1}})$
    - 分母大（更自信）→ 不确定性小；
    - 越靠后（衰减项越大）→ 不确定性被放大
- 构建语义图
  - 对每一句做 AMR 解析
    - 得到句内“实体—关系—实体”的结构
  - 指代消解 + 实体链接
  - 幻觉常见在实体关系上（谁是谁、时间地点对不对）以及跨句一致性上（前后矛盾），图结构能把这些显式化



算法原理

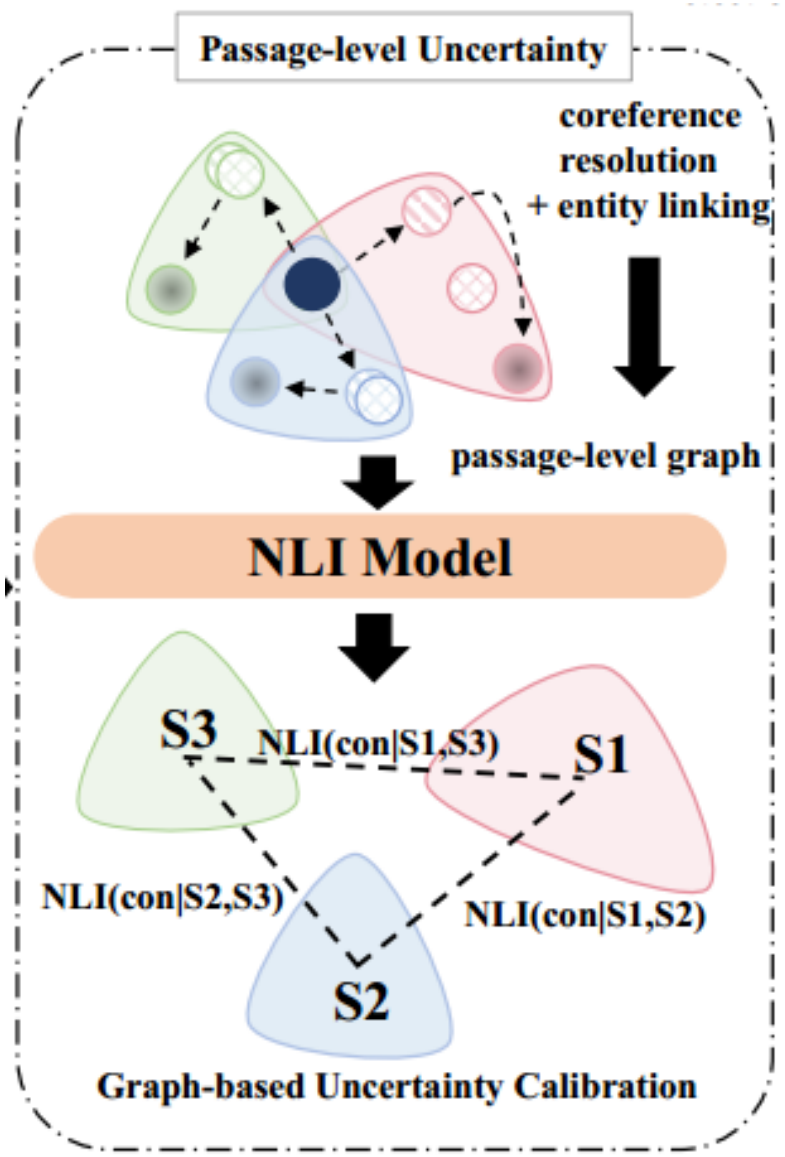
- 关系约束的不确定性传播
  - 不是把前面所有实体都传播到当前实体，而是只沿语义图里“确实有关系的实体”传播，防止过估计
- 句子级不确定性
  - 句子层面要同时覆盖两种风险：
    - 实体错误
    - 整体生成漂移
  - 实体不确定性UE
  - 全局不确定性UG
  - 句子分数Us

实体层面 + 全局层面



算法原理

- 篇章级校准
  - 两句各自都自信，但相互矛盾（典型幻觉）
  - 把每句的不确定性按与邻居句的矛盾概率加权，然后对全篇汇总
    - 如果一句话和它相关的邻居句越矛盾，它更可能是“问题句”，整体风险就上升
    - 只在语义图邻域做 NLI，比全句对全句更低成本、也更聚焦
- 输出（句子级 + 篇章级）
  - 句子级给每句风险分，可做句子定位
  - 篇章级给整段风险分，再用阈值判断是否幻觉



实验设计　数据资源

- 数据资源及指标

数据集	类型	用途
WikiBio（英文）	传记类文本（biography），用于篇章/段落级事实性评估场景	做句子级与篇章级幻觉检测评测
NoteSum（中文）	中文笔记/摘要类数据	同样做句子级与篇章级幻觉检测评测，用于验证跨语言泛化

- 基线实验
  - GPT-3 Uncertainties（Manakul et al., 2023）
  - SelfCheckGPT（Manakul et al., 2023）
  - FOCUS（Zhang et al., 2023c）

实验结果　对比实验

Methods	WikiBio					NoteSum				
	sentence-level			passage-level		sentence-level			passage-level	
	NonFact	NonFact*	Factual	Pearson	Spearman	NonFact	NonFact*	Factual	Pearson	Spearman
GPT-3 Uncertainties										
Avg(-log$p$)	83.21	38.89	53.97	57.04	53.93	80.11	43.69	35.29	39.61	31.55
Avg($\mathcal{H}$)	80.73	37.09	52.07	55.52	50.87	80.08	43.95	38.04	40.36	33.25
Max(-log$p$)	87.51	35.88	50.46	57.83	55.69	79.86	40.17	36.70	38.13	34.75
Max($\mathcal{H}$)	85.75	32.43	50.27	52.48	49.55	81.02	47.33	39.03	42.88	37.24
SelfCheckGPT (gpt-3.5-turbo)										
BertScore	81.96	45.96	44.23	58.18	55.90	76.44	39.69	36.89	25.91	21.24
QA	84.26	40.06	48.14	61.07	59.29	79.69	45.30	39.32	41.07	36.54
Unigram (max)	85.63	41.04	58.47	64.71	64.91	79.48	43.88	36.15	38.80	33.35
Combi	87.33	44.37	61.83	69.05	67.77	82.38	53.19	40.17	47.79	41.27
FOCUS										
LLaMA-13B	87.90	43.84	62.46	70.62	63.03	81.11	49.98	38.88	38.17	38.31
LLaMA-30B	89.79	48.80	65.69	77.15	73.24	82.17	43.12	49.85	37.37	40.09
OURS										
LLaMA-13B	90.14	**61.65**	64.82	72.11	64.35	85.06	50.70	53.03	**55.62**	60.81
LLaMA-30B	**90.93**	61.16	**65.70**	**77.60**	**74.44**	**87.95**	**54.42**	**61.51**	54.77	**61.05**
Δ	+1.14	+12.85	+0.01	+0.45	+1.20	+5.57	+1.23	+11.66	+7.83	+19.78

相比 FOCUS 那种“把所有前文关键词不确定性都传播到后面”的做法，本文在 NonFact* 与 Factual（中度/无幻觉）上提升更明显，说明“关系约束传播”能缓解过估计

实验结果　消融实验

	sentence-level			passage-level	
	NonFact	NonFact*	Fact	Pear.	Spear.
Ours	90.93	61.16	65.70	77.60	74.44
- max	86.48	64.86	63.52	23.32	38.57
- var	90.17	50.94	64.82	75.60	72.36
- decay	89.01	43.57	63.48	70.19	66.49
- entity	88.31	43.06	63.10	65.81	60.34
- global	88.75	43.88	65.19	70.36	65.49
- graph	-	-	-	75.89	72.20

Token层 max/var/decay 三项都有效，max 对篇章级最关键，var/decay 主要帮助细化不确定性

Sentence层 UE（实体相关传播）与 UG（全局分位数）都重要；

Passage层 实体传播对篇章级更关键

去掉图邻域矛盾校准（-graph）：篇章级 Pearson 75.89；Spearman 72.20
基于语义图邻域 + NLI 矛盾概率的校准有效

算法总结

- 算法贡献
  - 提出“语义图增强的不确定性建模”框架
    - 把幻觉检测从 token 级扩展到句子/篇章级，显式处理跨句依赖与一致性问题
  - 关系约束的不确定性传播
    - 仅沿语义图中实体相关路径传播不确定性，缓解以往方法“全前文粗传播”导致的过估计与误报
  - UE（实体传播）+ UG（全局分位数）
  - 篇章级图邻域校准
- 算法不足
  - 依赖多个外部组件，链路长、工程复杂度与误差传播风险高
  - 语义图质量高度敏感
  - 超参较多且跨域可能需调参，迁移成本存在

特点总结与未来展望

特点总结与未来展望

- 特点总结
  - 基于注意力谱特征的幻觉检测
    - 单次响应即可
    - 特征提取算子简单、成本较低（利用自回归结构做高效谱特征提取）
    - 强调“特征可分性”与跨层跨头聚合带来的稳定性
  - 语义图增强不确定性建模
    - 通过“只沿相关关系传播”缓解传统不确定性方法的过估计
    - 支持句子级定位 + 篇章级评分，更贴近应用需求
- 未来展望
  - 从单一信号到多证据融合
  - 长上下文/多轮对话的场景适用性
  - AI 模型领域的相关方法的迁移

参考文献

[1] Chen, Kedi, et al. "Enhancing uncertainty modeling with semantic graph for hallucination detection." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 39. No. 22. 2025.

[2] Binkowski, Jakub, et al. "Hallucination detection in llms using spectral features of attention maps." Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025.

道德经

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！