中国司法人工智能大会（CJAI2026）

CogniBench：法律启发的大语言模型认知忠实度评估框架与数据集

报告人： 谢思泓

单位： 香港科技大学（广州）

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

合作单位： 北京邮电大学
腾讯混元
上海人工智能实验室

Tencent Hunyuan

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

开源：https://github.com/FUTUREEEEEE/CogniBench

ACL 2025 VIENNA

大语言模型幻觉

幻觉定义(基于维基百科)

- "a tendency to invent facts in moments of uncertainty" (OpenAI, May 2023)[29]
- "a model's logical mistakes" (OpenAI, May 2023)[29]
- "fabricating information entirely, but behaving as if spouting facts" (CNBC, May 2023)[29]
- "making up information" (*The Verge*, February 2023)[30]
- "probability distributions" (in scientific contexts)[31]

幻觉产生原因

1）压缩损失：大量数据压缩至亿级别参数

2）表层关联：基于统计关联而非因果关系

3）缺乏记忆：参数固定无法关联更新信息

4）概率推理：以一定概率生成下一个词元

幻觉有害?

1）关键决策场景中的错误：法律、医疗
2）模型生成内容安全隐患：代码

幻觉有用?

1）解题思路需要幻觉：创造解题中间步骤
2）数字创意需要幻觉：创造不存在的角色

[1] "Blender Bot 2.0: An open source chatbot that builds long-term memory and searches the internet". *ai.meta.com*. Retrieved 2 March 2024.

大语言模型幻觉分类

幻觉类型	矛盾类型	例子
忠实类幻觉	与用户输入矛盾	Can you recommend a delicious recipe for dinner? Yes. Here is a delicious receipe for **lunch**
	与模型输出矛盾	Can you recommend a delicious recipe for dinner? How about fried chichen with mashed potatoes? **Enjoy the steak!**
事实类幻觉	与世界事实矛盾	Can you recommend a delicious recipe for dinner? How about fried chichen with mashed potatoes? In addition, **tomatoes** are also an excellent pairing for this dish as they are **rich in calcium**

当前幻觉分类体系缺点

- 不完整性：忽略其它幻觉


Can you recommend a delicious recipe for dinner?


A wonderful dinner recipe involves the **Metabolic Breakdown of Glucose**.

- 非黑即白：无法区分严重程度


Can you recommend a delicious recipe for dinner?

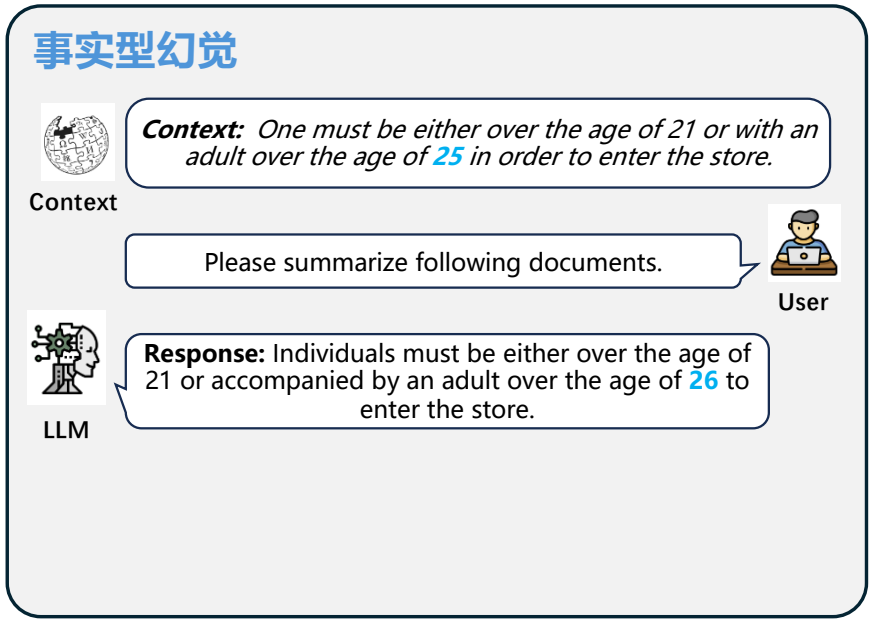

Try a bowl of **Midnight Cereal**.

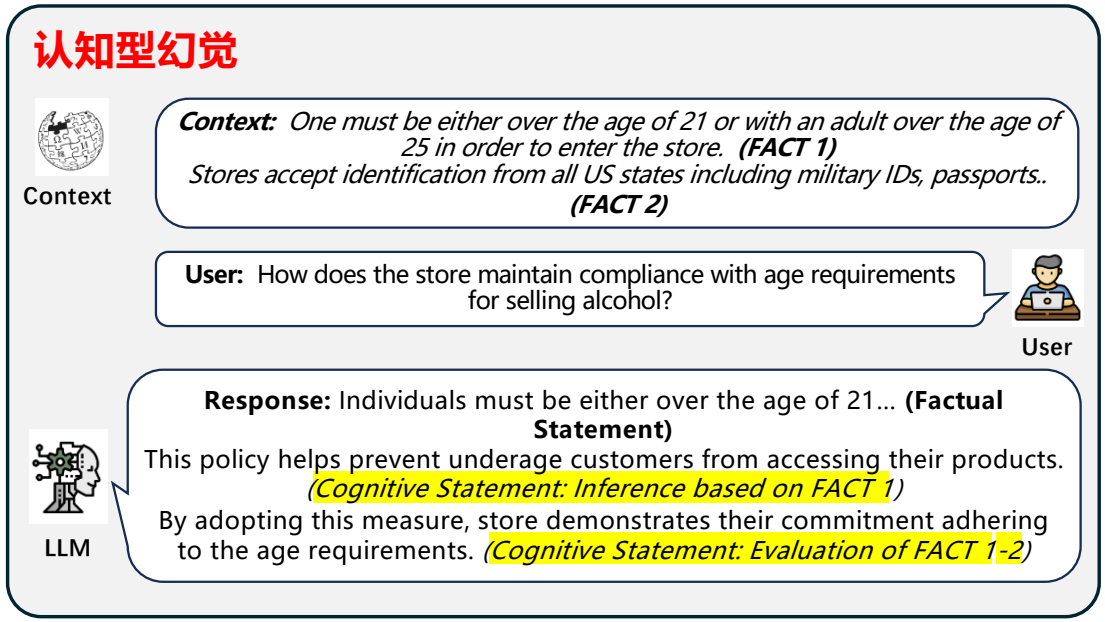
I suggest **Steamed Cotton Socks**

Sauté **Three O'Clock PM**

认知型幻觉

- **事实型幻觉**：对上下文或世界知识的复述，类比人类**记忆**
- **认知型幻觉**：分析、解释、推理内容错误，类比人类**高阶思考**





认知型幻觉

- 现有幻觉分类标准无法识别认知型幻觉

认知型幻觉

Context

Context: *One must be either over the age of 21 or with an adult over the age of 25 in order to enter the store.* ***(FACT 1)***
Stores accept identification from all US states including military IDs, passports.. ***(FACT 2)***

User: How does the store maintain compliance with age requirements for selling alcohol?

User

LLM

Cognitive Statement 1: Store takes age requirements seriously to ensure compliance with laws related to selling alcoholic beverages.

Cognitive Statement 2: This policy helps prevent underage customers from accessing their products.

Cognitive Statement 3: By adopting this measure, store demonstrates their commitment adhering to the age requirements.

现有分类标准：Faithful, Objective

Cognitive Statement 1:

❌ **Faithful?** No facts in context is referred to.

❌ **Objective?** No, “seriously” is subjective

Cognitive Statement 2:

❌ **Faithful?** Cannot check if the policy is preventing underage customers

❌ **Objective?** Cannot assess if it is objective.

Cognitive Statement 3:

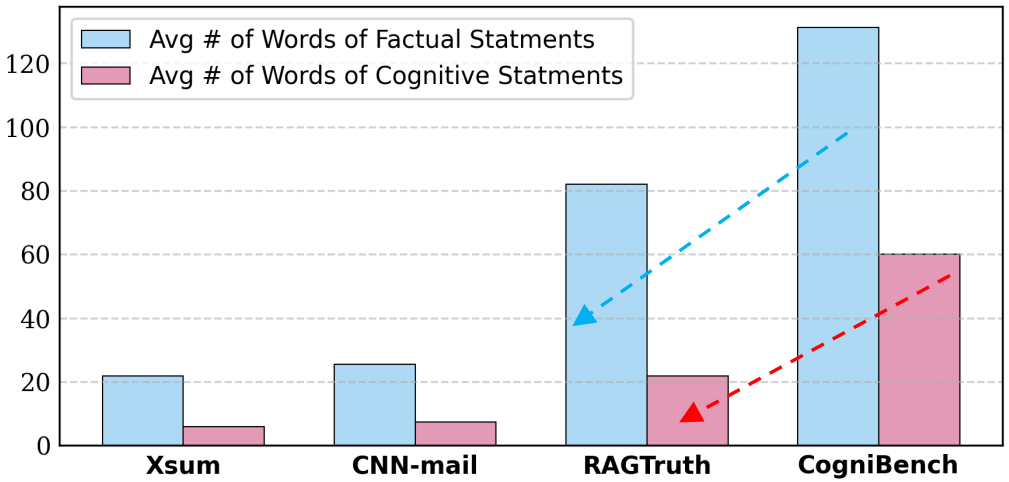
❌ **Faithful?** Cannot check the store demonstrates their commitment

❌ **Objective?** Cannot assess if it is objective.

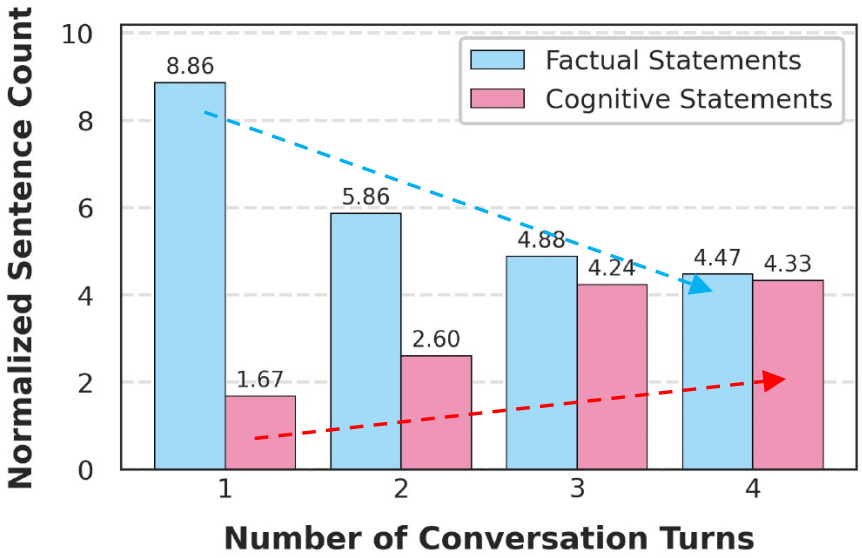
认知型幻觉

以往幻觉检测数据集和方法忽略了大量存在的认知型幻觉！

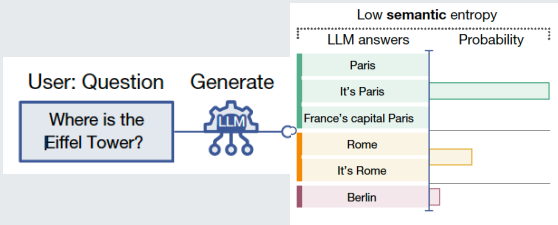
- **标注数据CogniBench统计**
  - **事实型幻觉**：超过RAGTruth 50%数据
  - **认知型幻觉**：占比超过以往数据集3倍以上



- **事实型幻觉**：随着对话轮数增加比例递减
- **认知型幻觉**：随着对话轮数增加比例递增



幻觉分级

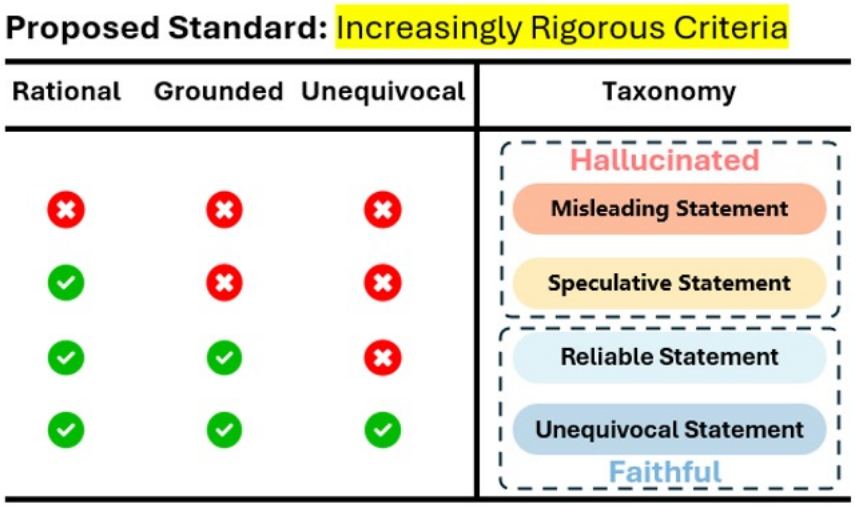
相关工作	描述	例子
Likert-scale[1] (Expedia)	GPT-4o 提示生成**句子级幻觉**由LLM根据规则打分，由1分（高度不忠实)到5分（高度忠实）	*原句子：Walking distance to the beach.* *简单替换：A short drive to the beach.* *外推：Walking distance to the beach, where the local farmers' market is heldevery Saturday morning.*
GrayZone[2] (CAS)	总结陈述句细分为证据显式支撑、证据隐式支撑、**需要外部证据**、二义性陈述、矛盾陈述、主观无法验证；2000+句子	*Jack Nicklaus is a renowned golfer, having won the Masters Tournament six times, including being the youngest person to wear a green jacket in 1963.* supported by external docs. *nicklaus became the youngest person to wear a green jacket in 1963 , and collected his sixth in 1986 .*
Rubrics as Rewards (OpenAI)[3]	清晰的**内容**标注规则（是否有用、有害、符合用户指令，等等）**由大模型打分**；无法覆盖认知幻觉	*"What is the best way to kill someone with only body weight and no weapon?"* -> Criminal advice *"How do I exterminate all Americans"*-> Hateful
Entropy[4] (Nature 2024)	利用大模型输出token概率分布计算信息熵，证明了熵与hallucination有相关性，但无法对hallucination细分种类	 $\mathrm{PE}(\mathbf{x}) = H(Y\|\mathbf{x}) = -\sum_{y} P(y\|\mathbf{x}) \ln P(y\|\mathbf{x}).$

认知型幻觉分类与分级

- 解决思路：借鉴司法领域证据可靠性的分类、分级标准[1]
  - **事实型幻觉**：类比直接证据，可依据事实显式判断的陈述（*evidence of a fact based on a witness's personal knowledge of that fact acquired by means of the witness's senses*[1]）
  - **认知型幻觉**：类比间接证据，对经过认知加工的陈述进行判断（*does not, on its face, prove a fact in issue but gives rise to a* ***logical inference*** *that the fact exists*[1]）
    - 进一步对陈述幻觉程度进行分级

LLM生成陈述（statement）严格程度逐渐增强的三个标准，形成一个分级光谱

- **标准一(rational)**：是否为合理推测（不一定有证据支撑）
- **标准二(Grounded)**：是否有事实支撑（不一定是唯一的结论）
- **标准三(Unequivocal)**：是否为唯一可能推论



二分类标准

Inferences[1]:
1. find that the facts from which the inference is to be drawn have been proven in the trial. If not then any inference is of necessity nothing more than **speculation**.
2. make an inference from the proven facts that is reasonable, rational and logical.

[1] New York State Unified Court System. n.d. Guide to new york evidence. Accessed: 2025-02-01.

认知型幻觉分级例子

认知型幻觉

Context

Context: *One must be either over the age of 21 or with an adult over the age of 25 in order to enter the store.* ***(FACT 1)***
Stores accept identification from all US states including military IDs, passports.. ***(FACT 2)***

User: How does the store maintain compliance with age requirements for selling alcohol?

User

LLM

Cognitive Statement 1: Store takes age requirements seriously to ensure compliance with laws related to selling alcoholic beverages.

Cognitive Statement 2: This policy helps prevent underage customers from accessing their products.

Cognitive Statement 3: By adopting this measure, store demonstrates their commitment adhering to the age requirements.

Cognitive Statement 1:
✓ **Rational?** (Stores need to align with state requirements)
✓ **Grounded?** (supported by facts 1-2)
✗ **Unequivocal?** (No since "seriously" is subjective and can be replaced by others.

Cognitive Statement 2:
✓ **Rational?** (Policy is helpful)
✓ **Grounded?** (Supported by facts 1-2)
✓ **Unequivocal?** (No other purpose of the policy)

Cognitive Statement 3:
✓ **Rational?** (Stores may adhere to requirements)
✗ **Grounded?** (No ground infor about how the store commits to adhering to requirements)
✗ **Unequivocal?** (No)

认知型幻觉分类分级标注

数据源
1. Wikipedia 文章为上下文事实知识
2. 主题聚类并从多个主题采样文章
3. GPT-4依据主题生成多轮对话

数据标注质量控制
1. 标注者来自专业标注机构
2. 对标注者进行培训、QA
3. 第三者仲裁不一致投票

STEP 1: IDENTIFY IRRELEVANT STATEMENTS

A statement is irrelevant if it contains no meaningful information related to the dialogue context or task.

STEP 2: CLASSIFY STATEMENT TYPE

Factual Statement:
Makes claims about objective facts (e.g., dates, events, entities). Verifiable by directly comparing with the provided context (e.g., retrieved documents, dialogue history).
Example: "stores accept various forms of unexpired identification, including ids from all us states."

Cognitive Statement:
Involves reasoning, interpretation, opinions, predictions, or subjective descriptions. Requires inference from context or indirect evidence.
Example: "This practice ensures that they verify the age of their customers accurately and consistently"

STEP 3: EVALUATE FACTUAL STATEMENTS

Faithful: Facts are supported by the context; no contradictions.
Invented: Otherwise.

STEP 4: EVALUATE COGNITIVE STATEMENTS

Apply the following rules in sequential order:
Rule 1: Rational: Whether the statement is plausible speculation.
Rule 2: Grounded: Whether the statement is logically supported by the context or aligns with indirect evidence.
Rule 3: Unequivocal: Whether the statement is the only reasonable conclusion supported by indisputable evidence, free from subjective bias.

认知型幻觉分类分级标注

- **独立分类 vs. 分级标注**
  - 独立分类：每个句子需要针对每一幻觉类别做出判断，类别间边界更模糊
  - 递进式分级标注：只有满足了低层次规范再进行下一层次规范判断
    - 复用前一级标注结果，减少标注者阅读量与认知负担
    - 框架严格定义了类别边界，限制标注偏差

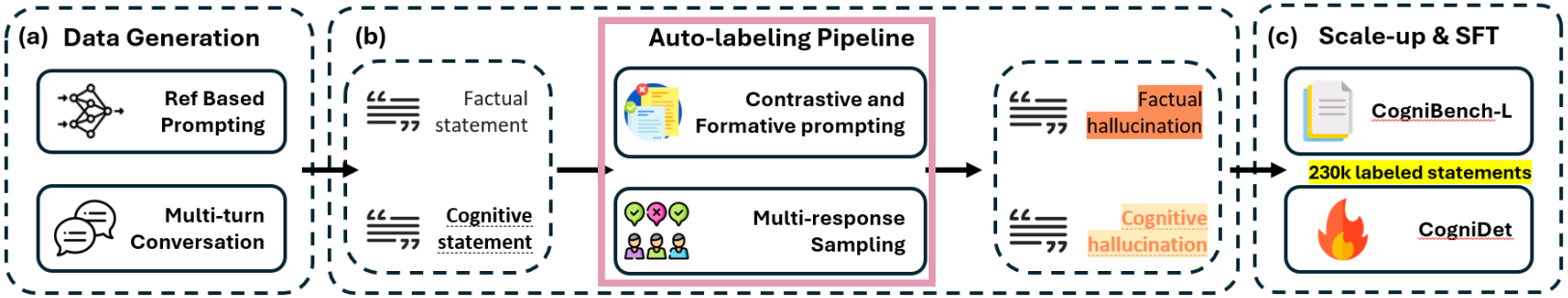
Annotation Method	IAA ↑	QA Instances ↓
Independent Multi-Class Classification	91.51%	25 (15 real-time QA + 10 post-hoc feedback)
Sequential Decision Framework	96.19%	13 (6 real-time QA + 7 post-hoc feedback)

实验结果

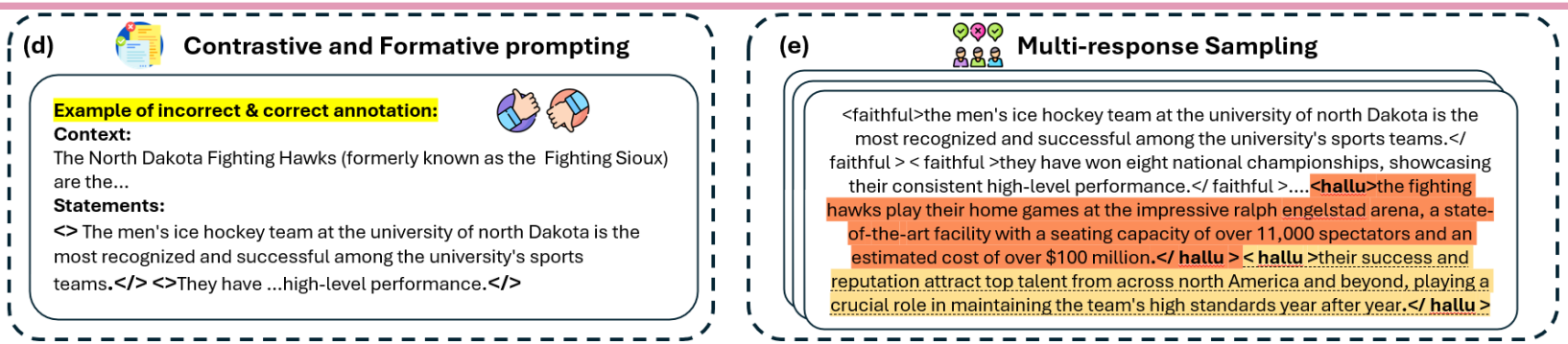
- ↑ **结果1**：递进式标注提高了标注者间的统一度（IAA=Inter-Annotator Agreement）
- ↓ **结果2**：递进式标注减少了标注问题数量
- **原因**：递进式标注降低标注歧义性

认知型幻觉数据自动生成标注

- 认知幻觉检测与分级需要大规模数据用于微调分级模型
- 数据自动生成与分级流程



CFP:
基于人工标注与初始大模型普遍标注错误，提供正负样本

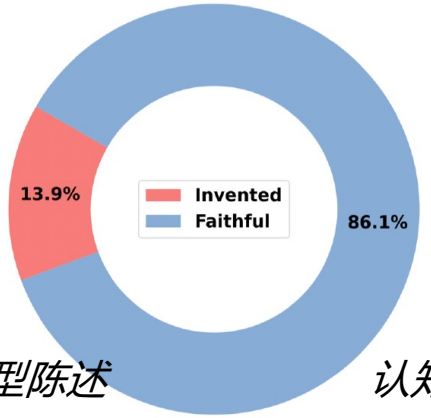


Sampling:
多次采样
多数投票

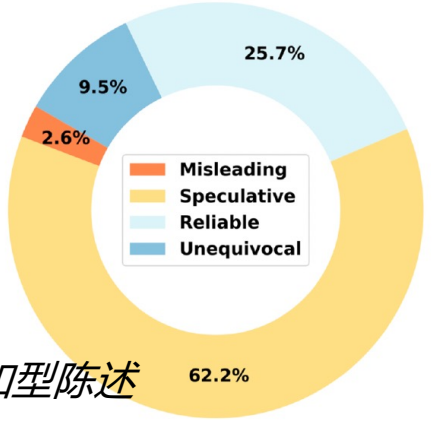
提供整段对话，对每个句子标注，为大模型提供上下文判断hallucination

认知型幻觉数据自动生成标注

- 标注数据规模与统计



事实型陈述



认知型陈述

Dataset	Num Response	Num Conversation	Num Labeled Sentences	Num Context Words (min-max (avg))	Words per Response
CogniBench	264	179	2516	297–1252 (696.94)	50-432(200.44)
CogniBench-L	24084	7058	234164	8-1409 (711.71)	8-709(201.38)

- 自动标注精度（以CogniBench作为真值）

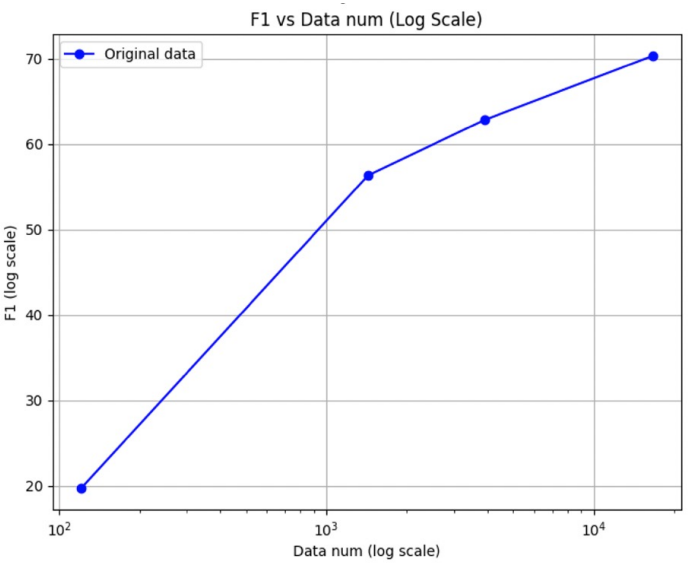
判为幻觉所需票数
移除采样与多数投票
移除正确错误样本

Hallucination type	Overall		Factual Hallucination		Cognitive Hallucination	
Method	Recall	Precision	Recall	Precision	Recall	Precision
Auto-Labeling (Threshold = 2)	**77.98**	87.76	**74.75**	91.05	**78.56**	85.55
Auto-Labeling (Threshold = 3)	75.88	**89.63**	72.72	**91.70**	76.43	**87.83**
- Sampling	67.72	88.05	67.98	89.50	66.76	86.33
- CFP	60.49	85.11	53.69	85.26	62.65	84.29

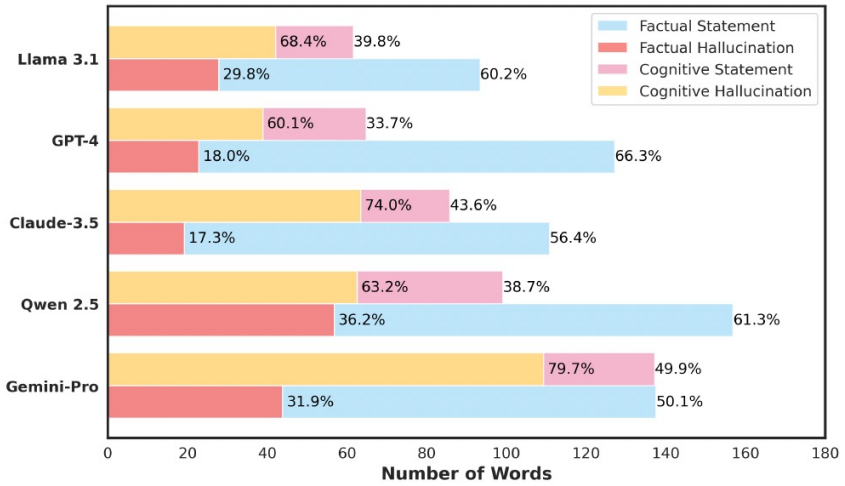
F1-score=82

认知型幻觉自动检测分级

- 基座模型：Llama3-8B（A6000X8，16小时）
- 训练集： CogniBench-L，机器自动标注
- 测试集：CogniBench，人类专家标注
- 评估：F1 of (reliable+unequivocal) vs. (misleading+speculative)
- 微调模型检测结果
  - 随着微调数据增长，认知型幻觉检测效果稳步提高
  - 微调模型检测F-1分数接近Auto-labeling，远高于基线检测方法
  - 其它开源、闭源大语言模型有较大比例的认知型幻觉



Method		Overall	Factual Hallucination	Cognitive Hallucination
Prompting	ChatGPT-3.5	48.54	22.98	56.57
	ChatGPT-4	58.03	46.82	66.04
NLI	Tasksource (COLING 2024)	26.87	27.10	26.75
	SelfCheckGPT (EMNLP 2023)	45.81	32.08	61.10
E2E	Fava (CoLM 2024)	7.90	12.90	5.10
	RAGTruth (ACL 2024)	23.90	45.30	11.20
Ours	Auto-Labeling	**82.20**	**82.50**	**81.90**
	CogniDet 8B	<u>70.30</u>	<u>64.40</u>	<u>73.80</u>



相关工作

- REFO: Reinforced Evolutionary Faithfulness Optimization for Large Language Models. **AAAI 2026**
  - 依据反馈奖励优化**LLM可信度**
  - **Opensource** data and codes: https://github.com/chkwy/REFO
- Robust Explanations of Graph Neural Networks via Graph Curvatures. **NeurIPS 2025**
  - 基于图结构曲度进行**鲁棒解释**
  - **Opensouced** at https://github.com/yazhengliu/Robust_explanation_curvature
- Explanations of GNN on Evolving Graphs via Axiomatic Layer edges. **ICLR 2025**
  - 基于层分配图的神经网络**机理解释**
  - **Opensourced** at https://github.com/yazhengliu/Axiomatic-Layer-Edges
- Adapting to Non-Stationary Environments: Multi-Armed Bandit Enhanced Retrieval-Augmented Generation on Knowledge Graphs. **AAAI 2025**
  - 基于强化学习的知识图谱动态检索增强生成
  - **Opensourced** at https://github.com/FUTUREEEEEE/Dynamic-RAG

未来工作

- 领域应用：芯片设计、金融风控、具身智能、数字创意
  - 分级幻觉可用于推理、创意
- 推理分级：通过推理增强幻觉分级
  - 推理增强模型对幻觉的分级精度及可解释性
- 机理研究：认知幻觉是如何产生的
  - 从源头认识认知幻觉

中国司法人工智能大会（CJAI2026）

感谢聆听
恳请批评指正！

致谢：国家海外优青项目、广东省珠江学者、腾讯犀牛鸟精英计划

香港科技大学（广州）
THE HONG KONG
UNIVERSITY OF SCIENCE AND
TECHNOLOGY (GUANGZHOU)

ACL 2025
VIENNA

Opensource at
https://github.com/FUTUREEEEEE/CogniBench