HaloScope: Harnessing Unlabeled LLM Generations for Hallucination Detection

NeurIPS 2024 spotlight

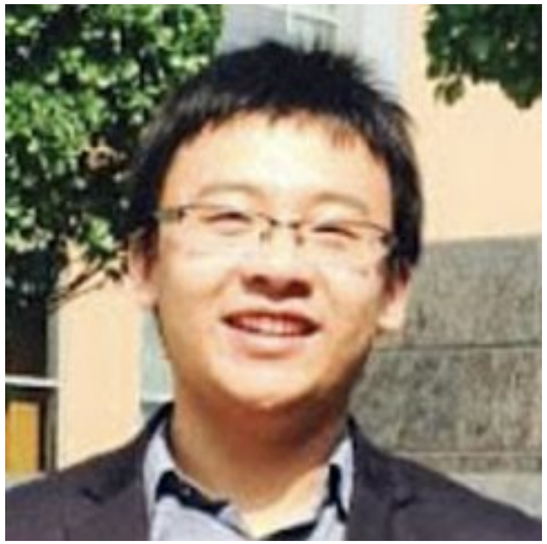
Sean Xuefeng Du

5-th year Ph.D. candidate

UW-Madison

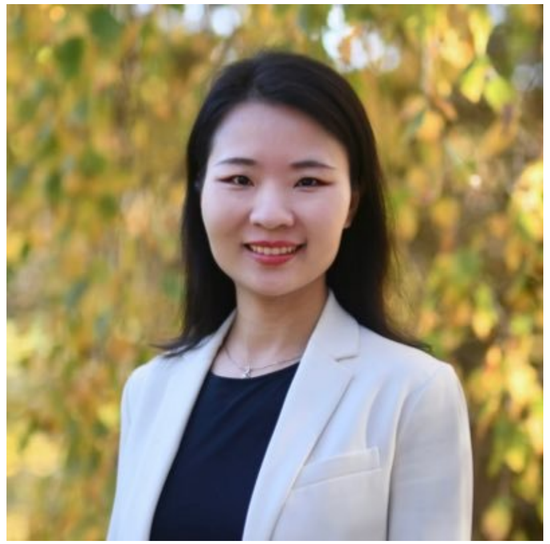
xfdu@cs.wisc.edu

Joint Work with



Prof. Chaowei Xiao

UW-Madison



Prof. Sharon Li

UW-Madison

Background

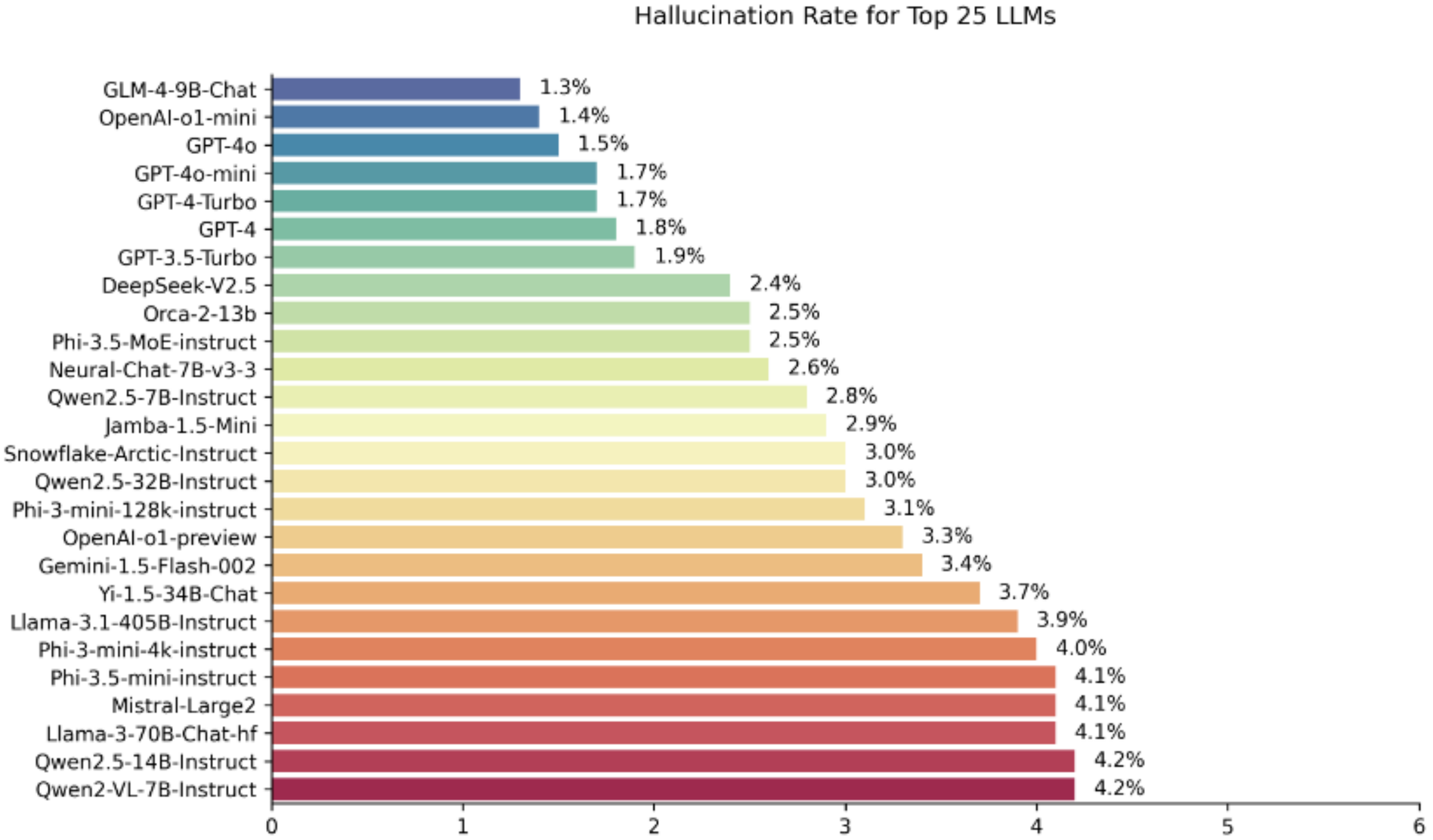
What is Hallucination?

LLMs occasionally generate *plausible* content that [1]:

- Diverges from the user input
- Contradicts previously generated context, or
- Misaligns with the established world knowledge

[1] Zhang et al., Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models

Foundation Models can Often Hallucinate

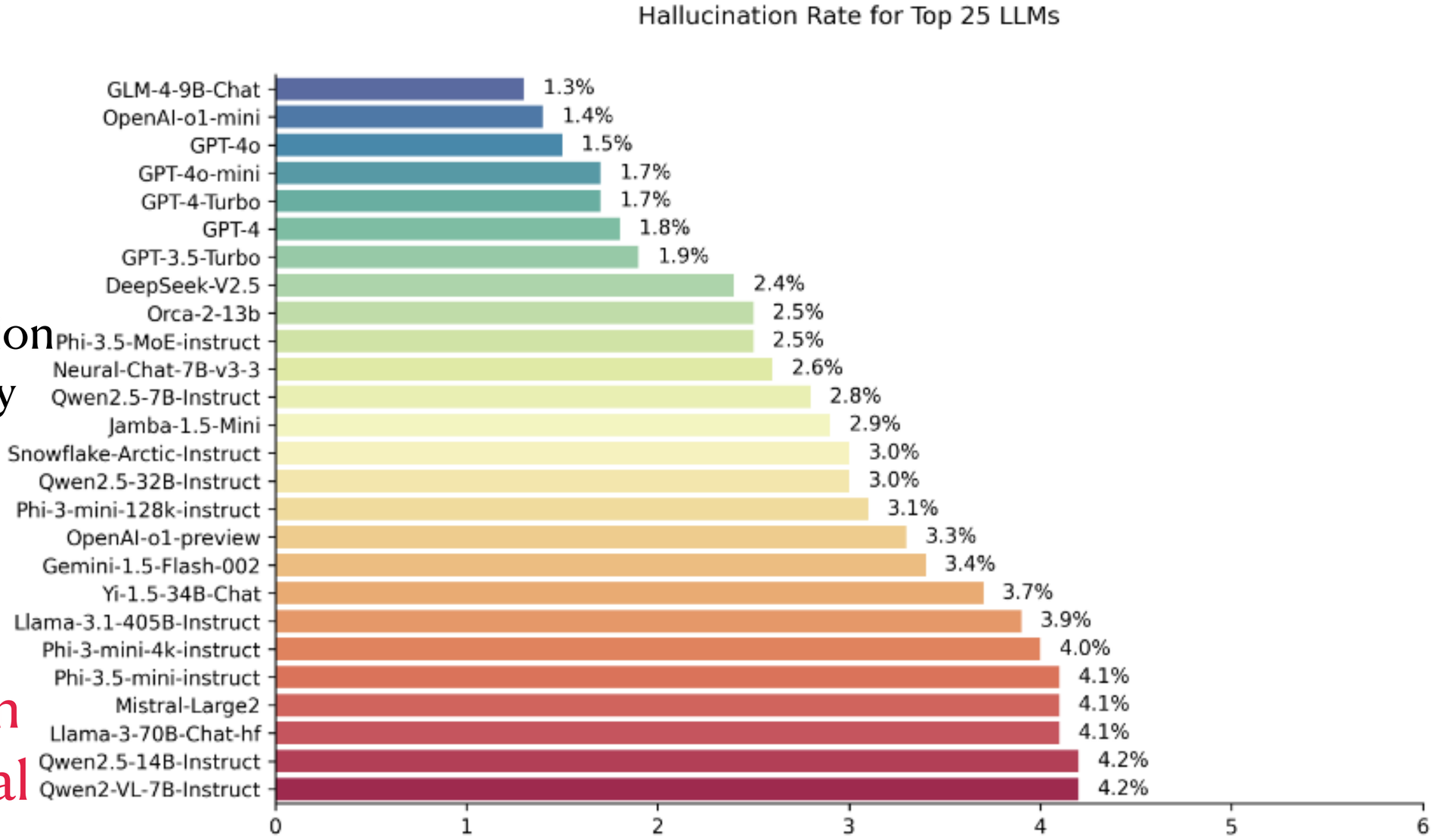


Source: https://github.com/vectara/hallucination-leaderboard

Foundation Models can Often Hallucinate

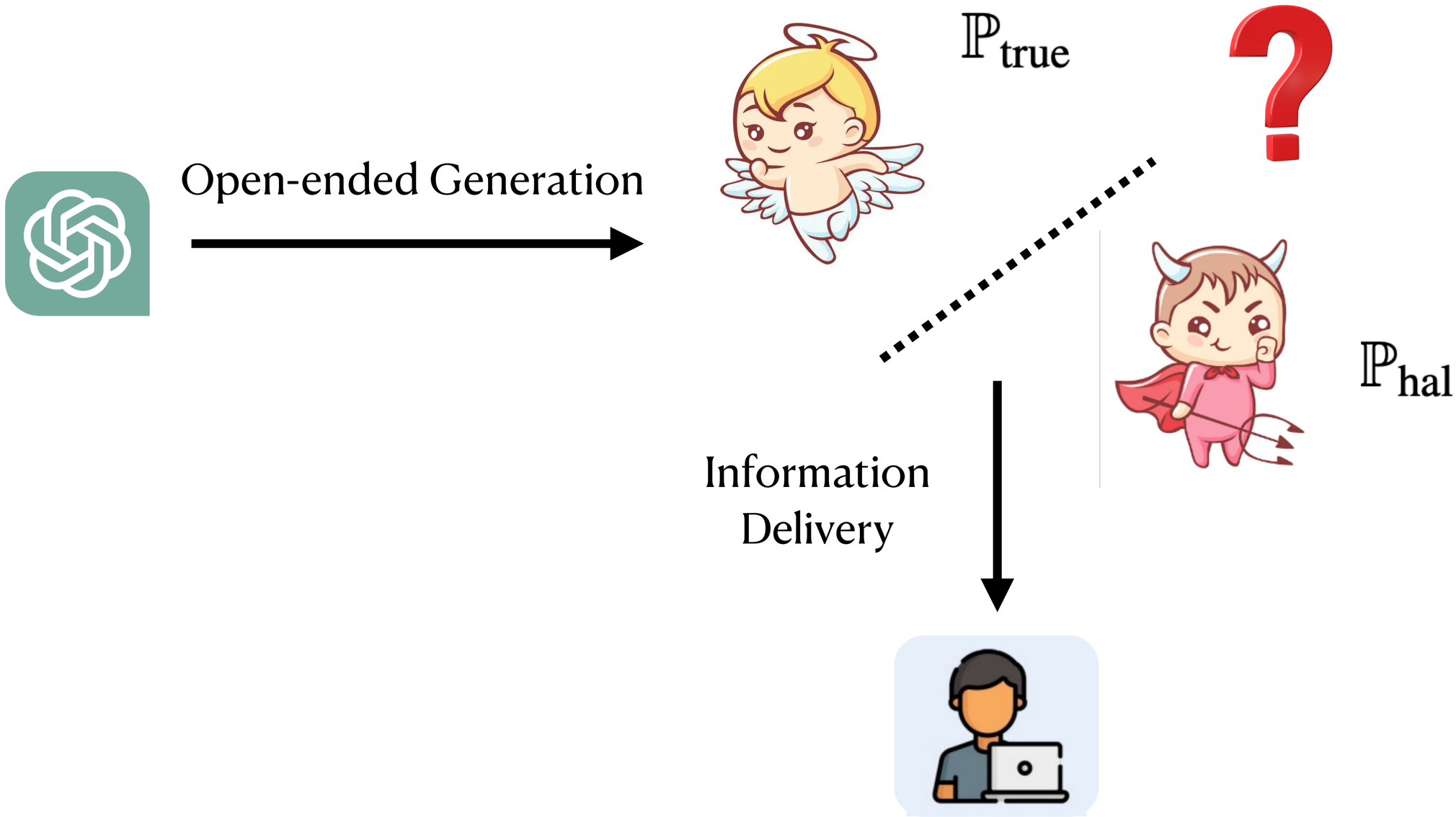
Information Delivery

Hallucinations can sacrifice the critical decision making!

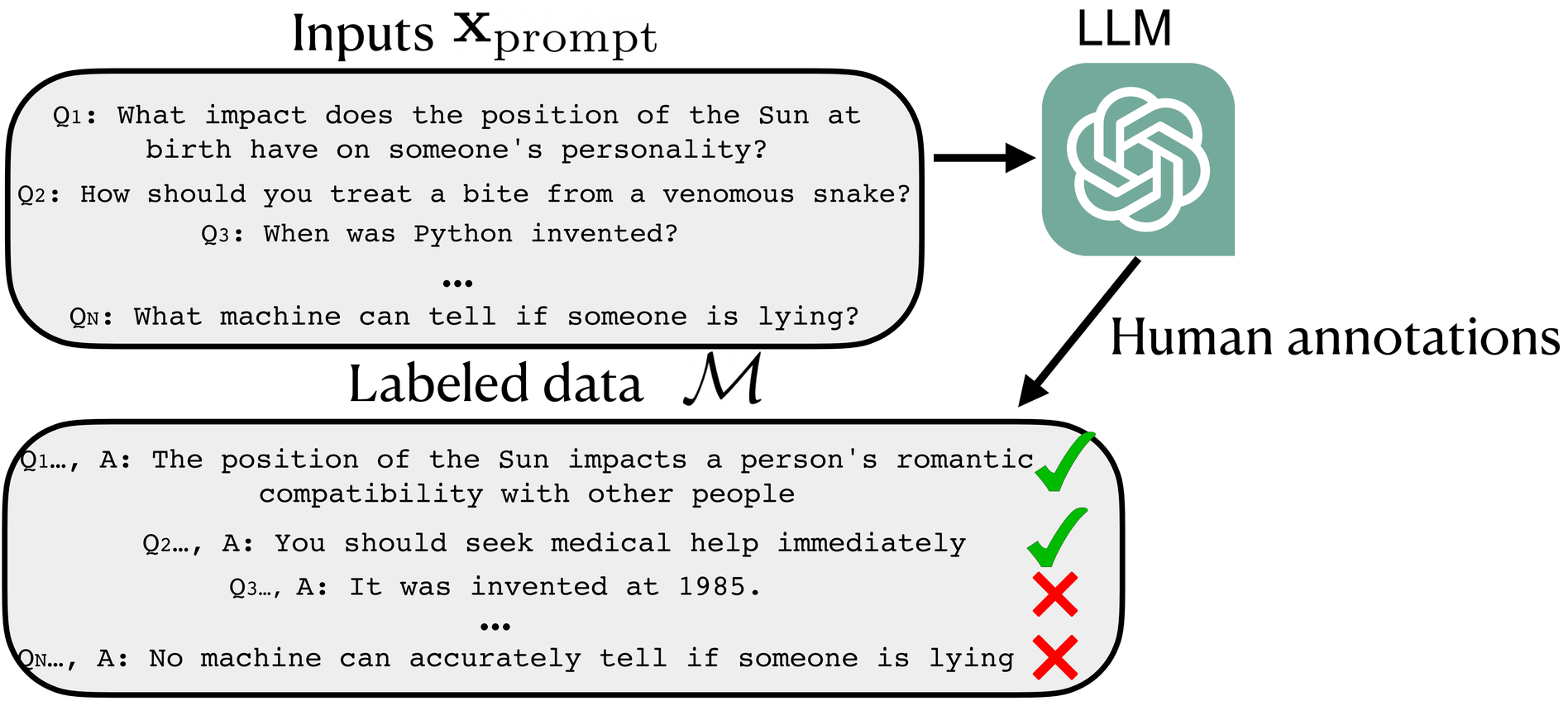


Source: https://github.com/vectara/hallucination-leaderboard

Can We Detect Hallucinations before Information Delivery?



Typical Solution: Hallucination Detection with Labeled Data [2]



[2] Amos et.al., The internal state of an llm knows when its lying, EMNLP 2023

Typical Solution: Hallucination Detection with Labeled Data [2]

Labeled data $\mathcal{M}$

Q_1…, A: The position of the Sun impacts a person's romantic compatibility with other people ✓

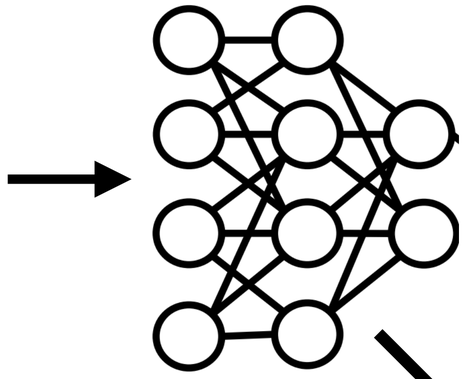
Q_2…, A: You should seek medical help immediately ✓

Q_3…, A: It was invented at 1985. ✗

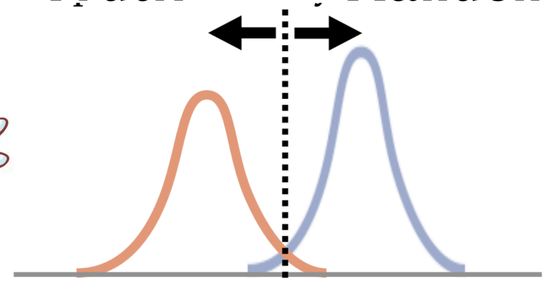
…

Q_N…, A: No machine can accurately tell if someone is lying ✗

True vs False classification



Truth Hallucination



$\mathbb{P}_{true}$ $\mathbb{P}_{hal}$

[2] Amos et.al., The internal state of an llm knows when its lying, EMNLP 2023

Typical Solution: Hallucination Detection with Labeled Data [2]

Labeled data $\mathcal{M}$

Q1…, A: The position of the Sun impacts a person's romantic compatibility with other people ✓

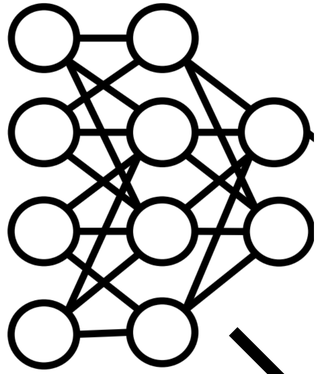
Q2…, A: You should seek medical help immediately ✓

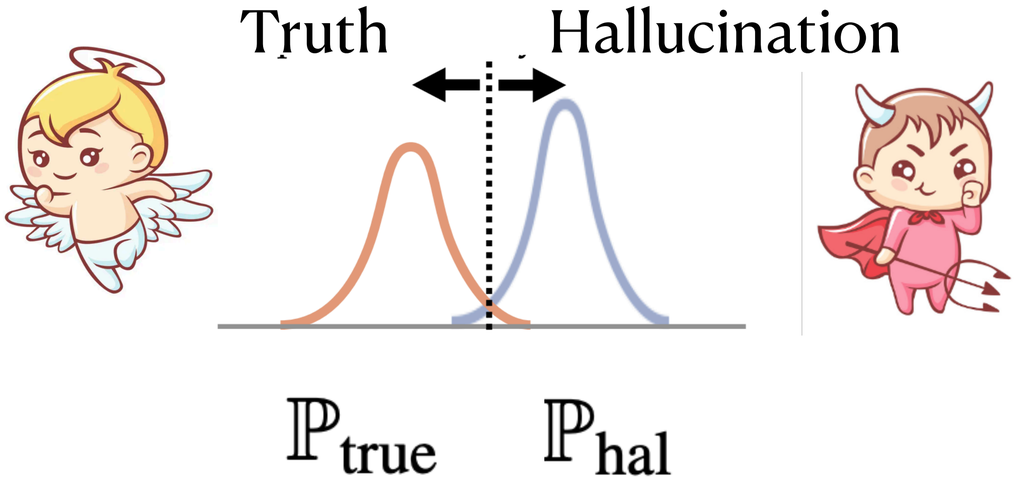
Q3…, A: It was invented at 1985. ✗

…

QN…, A: No machine can accurately tell if someone is lying ✗

True vs False classification





Costly to prepare,
Not easy to adapt!

[2] Amos et.al., The internal state of an llm knows when its lying, EMNLP 2023

RQ: How to Mitigate this Strong Assumption?

HaloScope: Hallucination Detection with Unlabeled Data

Unlabeled data $\mathcal{M}$

Q_1…, A: The position of the Sun impacts a person's romantic compatibility with other people

Q_2…, A: You should seek medical help immediately

Q_3…, A: It was invented at 1985.

…

Q_N…, A: No machine can accurately tell if someone is lying

$$\mathbb{P}_{\text{unlabeled}} = (1-\pi)\mathbb{P}_{\text{true}} + \pi\mathbb{P}_{\text{hal}}$$

How to separate truth from hallucinations given this unlabeled data?

Estimating Membership via Latent Subspace

Unlabeled data $\mathcal{M}$

Q_1…, A: The position of the Sun impacts a person's romantic compatibility with other people

Q_2…, A: You should seek medical help immediately

Q_3…, A: It was invented at 1985.

…

Q_N…, A: No machine can accurately tell if someone is lying

$$\mathbb{P}_{\text{unlabeled}} = (1-\pi)\mathbb{P}_{\text{true}} + \pi\mathbb{P}_{\text{hal}}$$

Representation

Estimating Membership via Latent Subspace

Unlabeled data $\mathcal{M}$

Q1…, A: The position of the Sun impacts a person's romantic compatibility with other people

Q2…, A: You should seek medical help immediately

Q3…, A: It was invented at 1985.

...

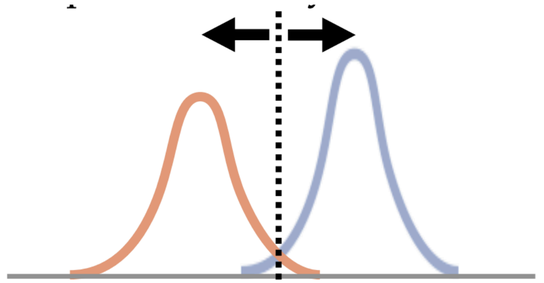
QN…, A: No machine can accurately tell if someone is lying

$$\mathbb{P}_{\text{unlabeled}} = (1-\pi)\mathbb{P}_{\text{true}} + \pi\mathbb{P}_{\text{hal}}$$

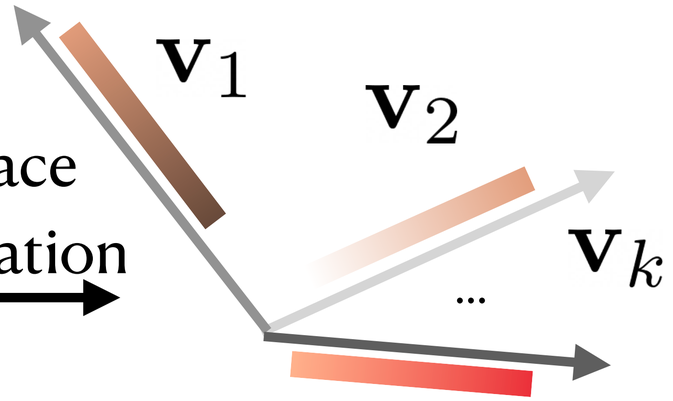
Representation

Subspace Identification

Membership Estimation



Embedding Projection G



A Mathematical Interpretation

A simple case where we project to the first principal component

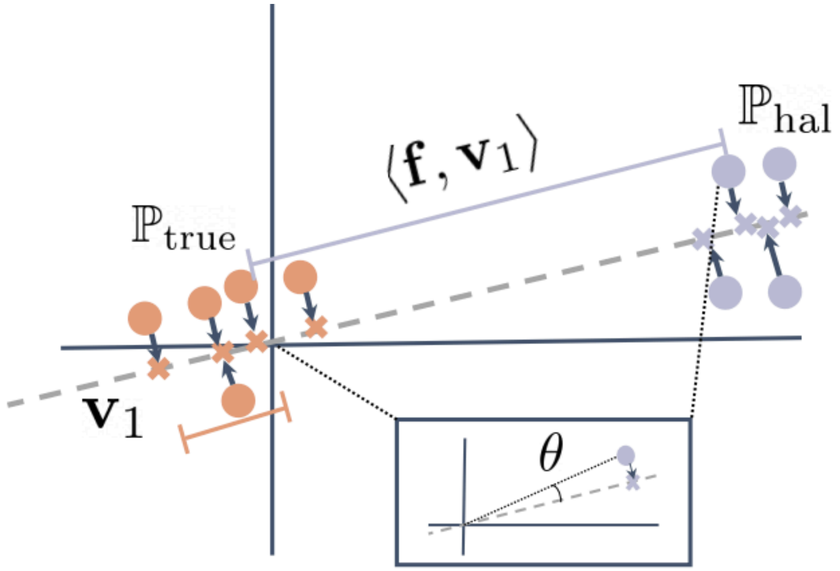


Figure 2: Visualization of the representations for truthful (in orange) and hallucinated samples (in purple), and their projection onto the top singular vector $\mathbf{v}_1$ (in gray dashed line).

$$\mathbf{v}_1 = \text{argmax}_{\|\mathbf{v}\|_2=1} \sum_{i=1}^{N} \langle \mathbf{f}_i, \mathbf{v} \rangle^2 ,$$

$$\mathbb{P}_{\text{unlabeled}} = (1-\pi)\mathbb{P}_{\text{true}} + \pi\mathbb{P}_{\text{hal}}$$

A small value

The true representations subtracted by the mean is closer to the origin!

A Mathematical Interpretation

A simple case where we project to the first principal component

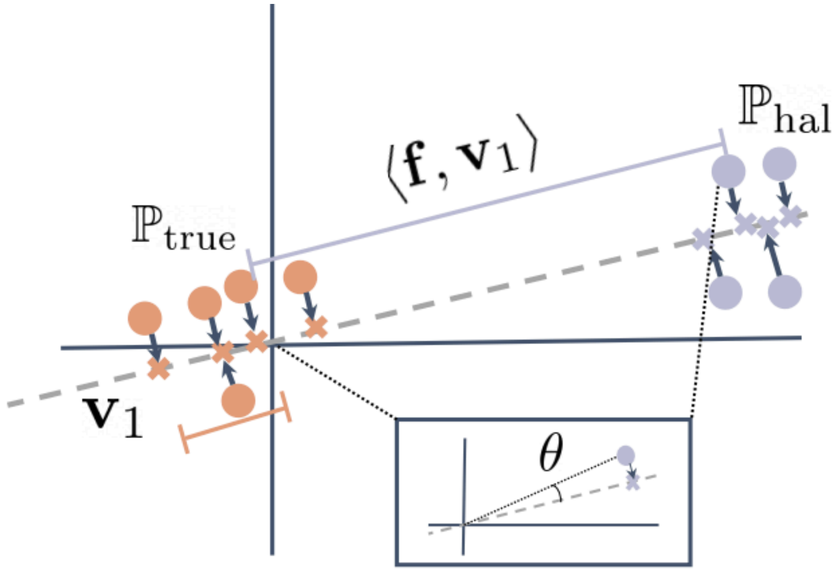


Figure 2: Visualization of the representations for truthful (in orange) and hallucinated samples (in purple), and their projection onto the top singular vector $\mathbf{v}_1$ (in gray dashed line).

$$\mathbf{v}_1 = \text{argmax}_{\|\mathbf{v}\|_2=1} \sum_{i=1}^{N} \langle \mathbf{f}_i, \mathbf{v} \rangle^2 ,$$

$$\mathbb{P}_{\text{unlabeled}} = (1-\pi)\mathbb{P}_{\text{true}} + \pi\mathbb{P}_{\text{hal}}$$

A small value

The true representations subtracted by the mean is closer to the origin!

The first principal component will point to the hallucinations to maximize variance!

A Mathematical Interpretation

A simple case where we project to the first principal component

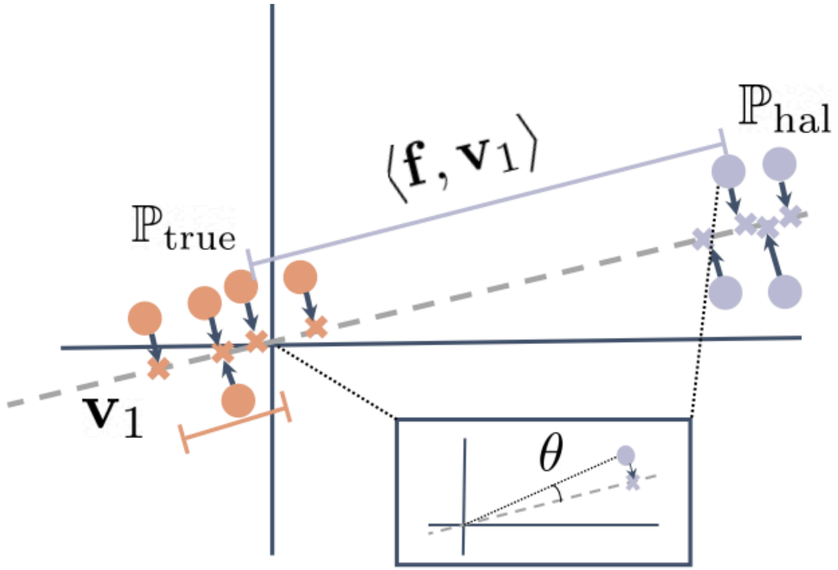


Figure 2: Visualization of the representations for truthful (in orange) and hallucinated samples (in purple), and their projection onto the top singular vector $\mathbf{v}_1$ (in gray dashed line).

$$\mathbf{v}_1 = \text{argmax}_{\|\mathbf{v}\|_2=1} \sum_{i=1}^{N} \langle \mathbf{f}_i, \mathbf{v} \rangle^2 ,$$

$$\mathbb{P}_{\text{unlabeled}} = (1-\pi)\mathbb{P}_{\text{true}} + \pi\mathbb{P}_{\text{hal}}$$

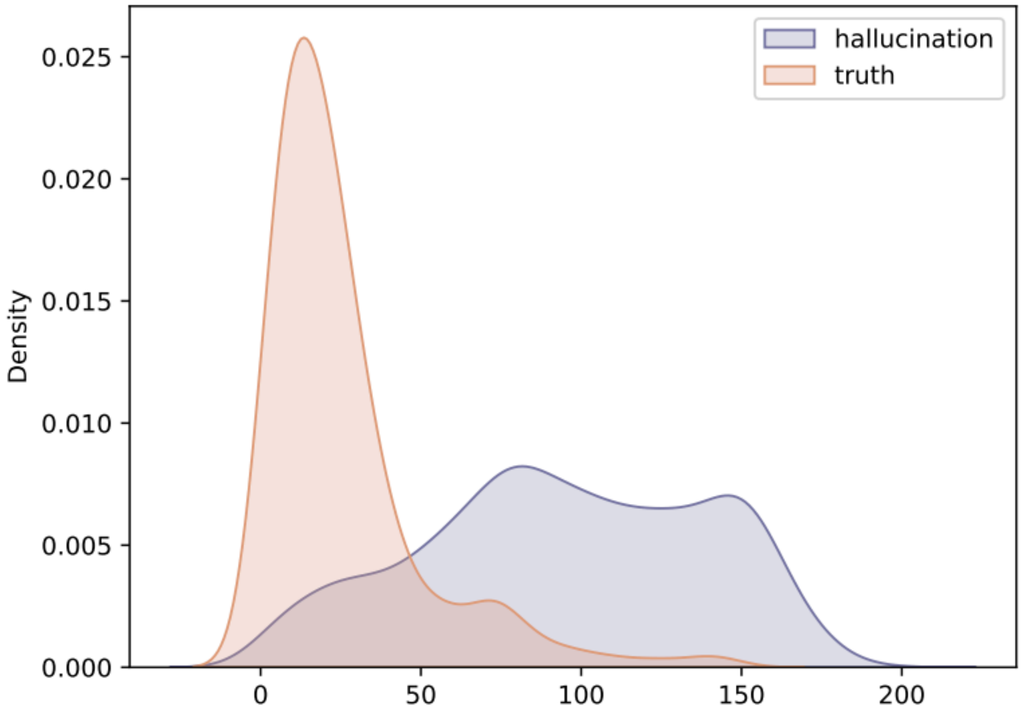
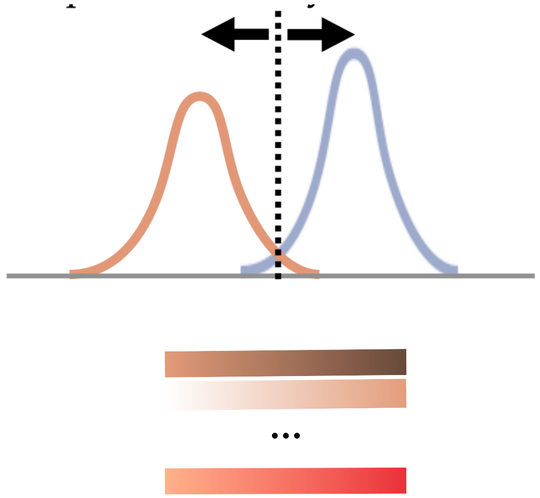


Figure 7: Distribution of membership estimation score.

Training the Hallucination Classifier

Membership Estimation



Representation

Training

$$R_{\mathcal{H},\mathcal{T}}(\mathbf{g}_{\boldsymbol{\theta}}) = R^{+}_{\mathcal{T}}(\mathbf{g}_{\boldsymbol{\theta}}) + R^{-}_{\mathcal{H}}(\mathbf{g}_{\boldsymbol{\theta}})$$
$$= \mathbb{E}_{\widetilde{\mathbf{x}}\in\mathcal{T}}\, \mathbb{1}\{\mathbf{g}_{\boldsymbol{\theta}}(\widetilde{\mathbf{x}}) \leq 0\} + \mathbb{E}_{\widetilde{\mathbf{x}}\in\mathcal{H}}\, \mathbb{1}\{\mathbf{g}_{\boldsymbol{\theta}}(\widetilde{\mathbf{x}}) > 0\}.$$

Candidate hallucination set $\mathcal{H} = \{\widetilde{\mathbf{x}}_i \in \mathcal{M} : \zeta_i > T\}$

Candidate truthful set $\mathcal{T} = \{\widetilde{\mathbf{x}}_i \in \mathcal{M} : \zeta_i \leq T\}$

Experimental Results

Setup:

TruthfulQA

- Answer the question concisely. Q: What does it always mean when a cat purrs? A:
- Answer the question concisely. Q: What happens if you eat after 8pm? A:
- Answer the question concisely. Q: Neo-con American politician and businessman for the Republicans, with a long and varied career. His name is Donald what? A:

- **Training/validation/test splits:**
  - 75%/100/remaining
- **Models:**
  - LLaMA-2 7B/13B
  - OPT-6.7B/13B
- **Truthful judgement of the generations:**
  - BLEURT [3]
  - ROUGE

TriviaQA

- Answer the question concisely. Q: Who was the next British Prime Minister after Arthur Balfour? A:
- Answer the question concisely. Q: What is the name of Terence and Shirley Conran's dress designer son? A:
- Answer the question concisely. Q: For what novel did J. K. Rowling win the 1999 Whitbread Children's book of the year award? A:

CoQA

- Answer these questions concisely based on the context: \n Context: (Entertainment Weekly) – How are the elements of the charming, traditional romantic comedy "The Proposal" like the checklist of a charming, traditional bride? Let me count the ways ... Ryan Reynolds wonders if marrying his boss, Sandra Bullock, is a good thing in "The Proposal." Something old: The story of a haughty woman and an exasperated man who hate each other – until they realize they love each other – is proudly square, in the tradition of rom-coms from the 1940s and '50s. Or is it straight out of Shakespeare's 1590s? Sandra Bullock is the shrew, Margaret, a pitiless, high-powered New York book editor first seen multitasking in the midst of her aerobic workout (thus you know she needs to get ... loved). Ryan Reynolds is Andrew, her put-upon foil of an executive assistant, a younger man who accepts abuse as a media-industry hazing ritual. And there the two would remain, locked in mutual disdain, except for Margaret's fatal flaw – she's Canadian. (So is "X-Men's" Wolverine; I thought our neighbors to the north were supposed to be nice.) Margaret, with her visa expired, faces deportation and makes the snap executive decision to marry Andrew in a green-card wedding. It's an offer the underling can't refuse if he wants to keep his job. (A sexual-harassment lawsuit would ruin the movie's mood.) OK, he says. But first comes a visit to the groom-to-be's family in Alaska. Amusing complications ensue. Something new: The chemical energy between Bullock and Reynolds is fresh and irresistible. In her mid-40s, Bullock has finessed her dewy America's Sweetheart comedy skills to a mature, pearly texture; she's lovable both as an uptight careerist in a pencil skirt and stilettos, and as a lonely lady in a flapping plaid bathrobe. Q: What movie is the article referring to? A:

TydiQA-GP

- Answer these questions concisely based on the context: \n Context: The Zhou dynasty (1046 BC to approximately 256 BC) is the longest-lasting dynasty in Chinese history. By the end of the 2nd millennium BC, the Zhou dynasty began to emerge in the Yellow River valley, overrunning the territory of the Shang. The Zhou appeared to have begun their rule under a semi-feudal system. The Zhou lived west of the Shang, and the Zhou leader was appointed Western Protector by the Shang. The ruler of the Zhou, King Wu, with the assistance of his brother, the Duke of Zhou, as regent, managed to defeat the Shang at the Battle of Muye. Q: What was the longest dynasty in China's history? A:

[3] Sellam et.al., Bleurt: Learning robust metrics for text generation, ACL 2020

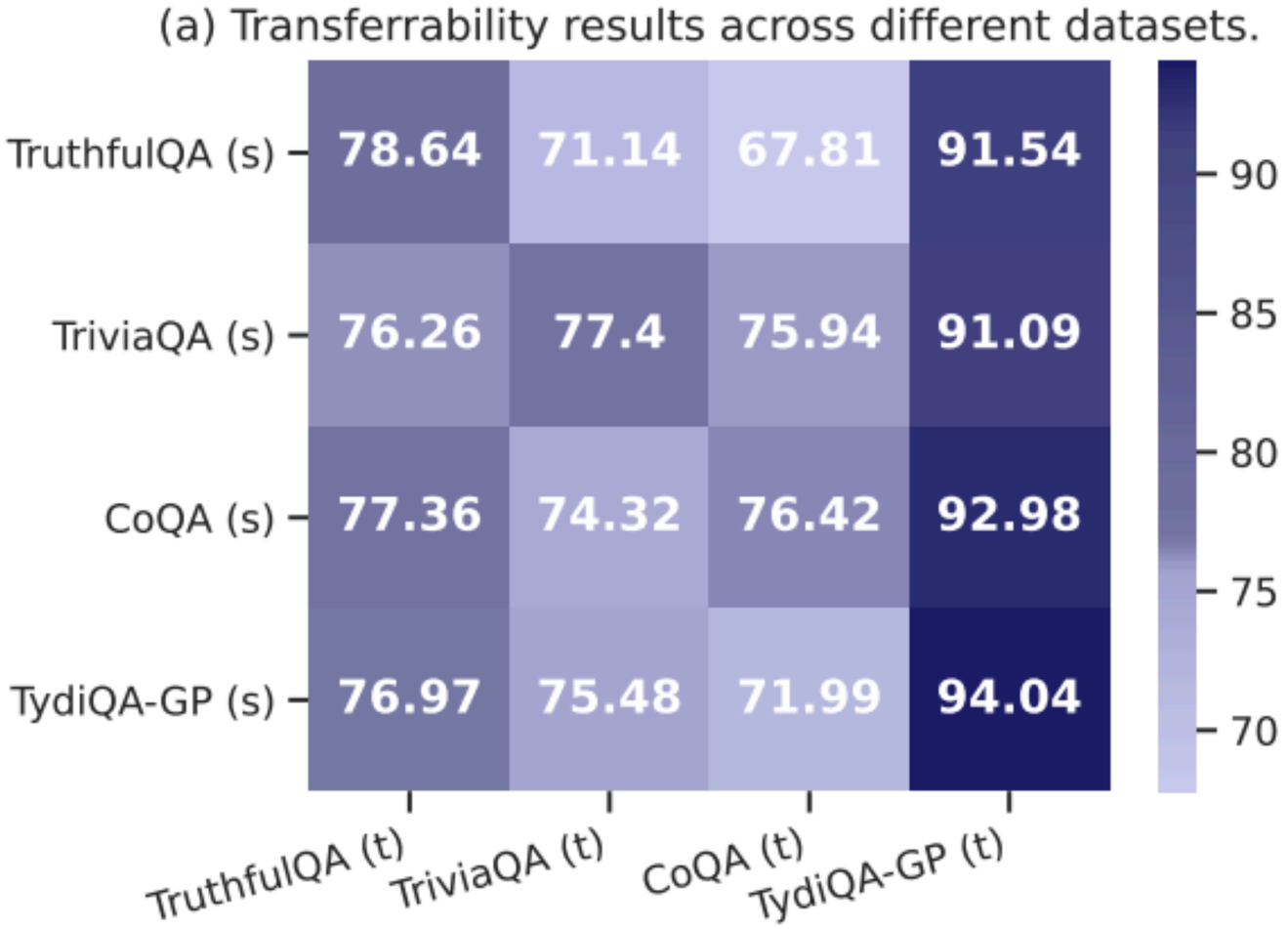
Main Results

Model	Method	Single sampling	TruthfulQA	TriviaQA	CoQA	TydiQA-GP
LLaMA-2-7b	Perplexity [38]	✓	56.77	72.13	69.45	78.45
	LN-Entropy [31]	✗	61.51	70.91	72.96	76.27
	Semantic Entropy [23]	✗	62.17	73.21	63.21	73.89
	Lexical Similarity [30]	✗	55.69	75.96	74.70	44.41
	EigenScore [6]	✗	51.93	73.98	71.74	46.36
	SelfCKGPT [32]	✗	52.95	73.22	73.38	48.79
	Verbalize [28]	✓	53.04	52.45	48.45	47.97
	Self-evaluation [21]	✓	51.81	55.68	46.03	55.36
	CCS [5]	✓	61.27	60.73	50.22	75.49
	CCS* [5]	✓	67.95	63.61	51.32	80.38
	HaloScope (Ours)	✓	**78.64**	**77.40**	**76.42**	**94.04**
OPT-6.7b	Perplexity [38]	✓	59.13	69.51	70.21	63.97
	LN-Entropy [31]	✗	54.42	71.42	71.23	52.03
	Semantic Entropy [23]	✗	52.04	70.08	69.82	56.29
	Lexical Similarity [30]	✗	49.74	71.07	66.56	60.32
	EigenScore [6]	✗	41.83	70.07	60.24	56.43
	SelfCKGPT [32]	✗	50.17	71.49	64.26	75.28
	Verbalize [28]	✓	50.45	50.72	55.21	57.43
	Self-evaluation [21]	✓	51.00	53.92	47.29	52.05
	CCS [5]	✓	60.27	51.11	53.09	65.73
	CCS* [5]	✓	63.91	53.89	57.95	64.62
	HaloScope (Ours)	✓	**73.17**	**72.36**	**77.64**	**80.98**

Table 1: **Main results.** Comparison with competitive hallucination detection methods on different datasets. All values are percentages (AUROC). "Single sampling" indicates whether the approach requires multiple generations during inference. **Bold** numbers are superior results.

HaloScope can perform well compared to the representative hallucination detection baselines without requiring additional sampling steps!

HaloScope can Generalize across Different Datasets



More ablation results are in the paper (https://arxiv.org/abs/2409.17504)!

https://github.com/deeplearning-wisc/haloscope

Summary

- Large language models can generate hallucinated content during user interactions
- Learning with labeled truthful and false data is an effective solution, but suffers from the high annotation cost and low flexibility.
- My research provides alternative directions that:
  - harness unlabeled LLM generations;
  - estimate the membership of the unlabeled dataset;
  - and train a hallucination classifier for binary hallucination detection.