



# Understanding Knowledge Drift in LLMs through Misinformation

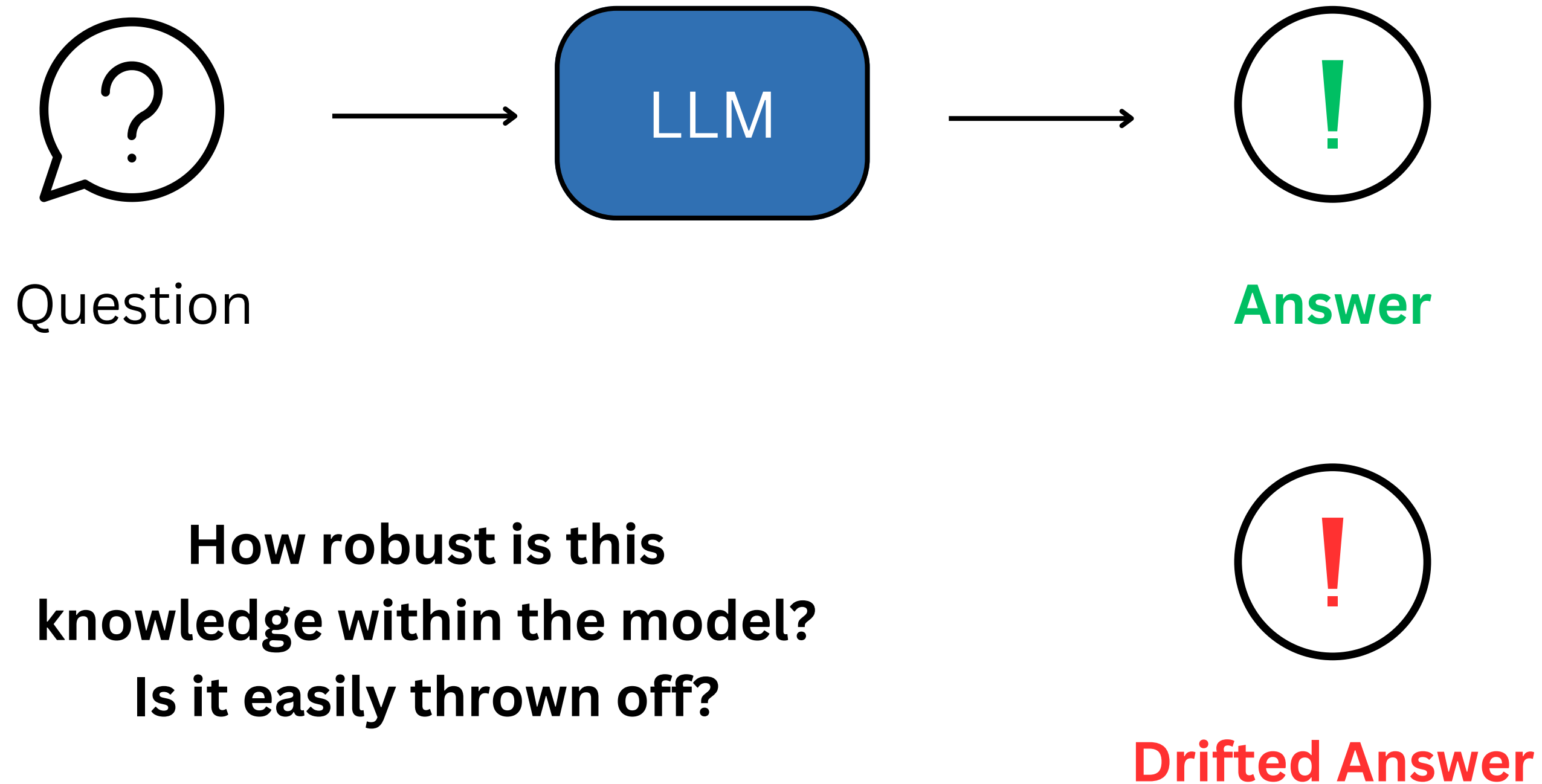
**Alina Fastowski**

PhD Student

[alina.fastowski@tum.de](mailto:alina.fastowski@tum.de)

# Knowledge Drift?

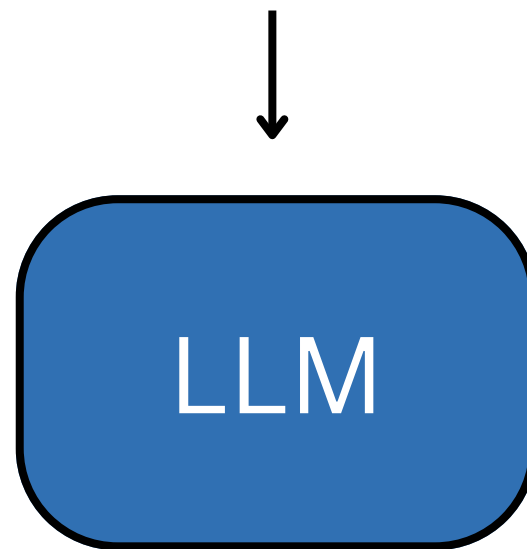
Drift from model's "original" knowledge:



# What did we do?

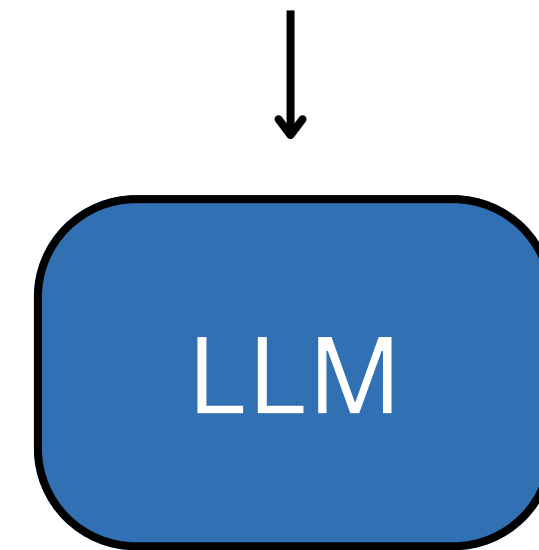
Overall idea: Infuse **false information** into the question prompts

“Which country has the rand as its currency?”



*“South Africa”*

“**The Australian currency is called the rand.** Which country has the rand as its currency?”



...?

# Experimental Setup

## Data

TriviaQA: 1000 samples (Q+A pairs)

## Models



**We keep working with the correctly answered samples.**

	Accuracy	#Parameters
GPT-4o	0.790	NA
GPT-3.5	0.721	$1.75 \times 10^{11}$
Mistral-7B	0.502	$7 \times 10^9$
LLaMA-2-13B	0.428	$1.3 \times 10^{10}$

# Experimental Setup

## Infusing false information: prompts

Baseline:

Question.

False Information  $\times k$ :  
( $k \in \{1, 2, 5, 10\}$ )

False Info.  
 $\times k$

Question.

Random Information:

Random Info. Question.

Question: “Who directed 2001: A Space Odyssey?”

---

✓ Correct Answer: “Stanley Kubrick”

**Baseline**

✗ False Information: “Alfred Hitchcock directed 2001: A Space Odyssey.”

Question: “Who directed 2001: A Space Odyssey?”

---

✓ Correct Answer: “Stanley Kubrick”

**False info**

\* Random Information: “In the 1960s, video recorders were first developed.”

Question: “Who directed 2001: A Space Odyssey?”

---

✓ Correct Answer: “Stanley Kubrick”

**Random info**

# What are we monitoring?

- **Correctness** - Does model drift to incorrect answers?
- **Uncertainty** - How certain is the model about its answers?



$$H(y | x, \theta) = -\frac{1}{T} \sum_t \sum_i p(y_{t_i} | y_{<t_i}, x) \log p(y_{t_i} | y_{<t_i}, x)$$

$$\text{PPL}(y | x, \theta) = \exp\left(-\frac{1}{T} \sum_t \log p(y_t | y_{<t}, x)\right)$$

$$\text{TP}(y | x, \theta) = \frac{1}{T} \sum_t \exp(\log p(y_t | y_{<t}, x))$$

# Findings - Correctness

	GPT-4o		GPT-3.5		Mistral-7B		LLaMA-2-13B	
	Prompt V1	Prompt V2	Prompt V1	Prompt V2	Prompt V1	Prompt V2	Prompt V1	Prompt V2
B	0.987	0.986	0.982	0.971	1.000	0.984	0.829	0.815
RIP	0.958	0.940	0.914	0.908	0.866	0.846	0.734	0.706
FIP	0.921	0.934	0.781	0.863	0.516	0.539	0.359	0.364
FIP×2	0.759	0.853	0.642	0.739	0.352	0.376	0.231	0.269
FIP×5	0.710	0.820	0.592	0.678	0.287	0.304	0.182	0.203
FIP×10	0.687	0.810	0.578	0.671	0.265	0.301	0.158	0.177
% FIP×10 vs. B	-30.4%	-17.8%	-41.1%	-30.9%	-73.5%	-69.4%	-80.9%	-78.3%

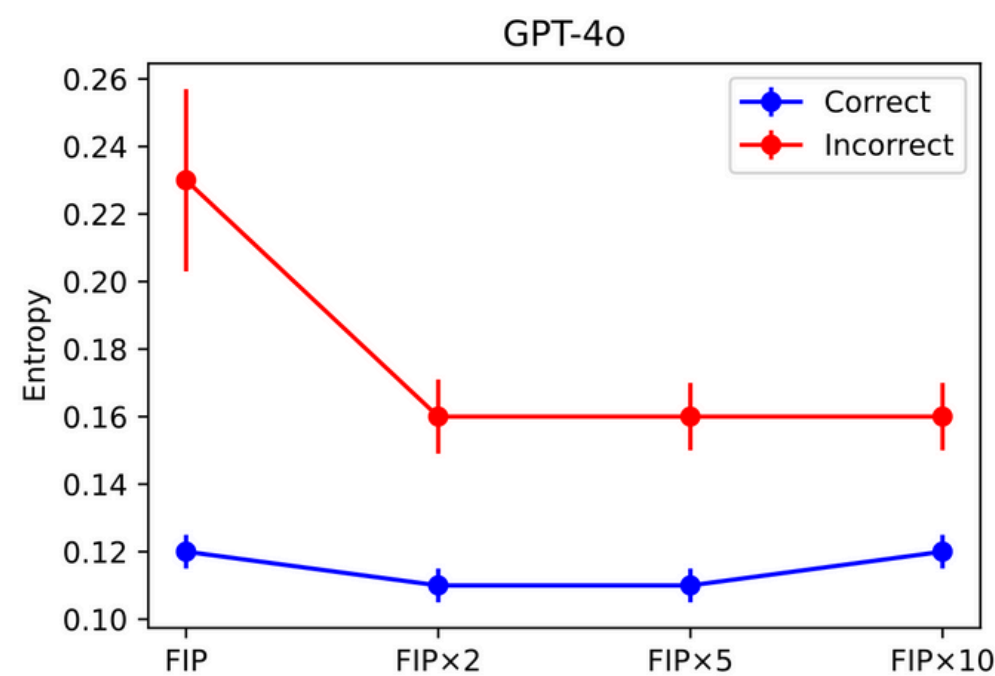
Infusing false information -> **drops in question answering accuracy**



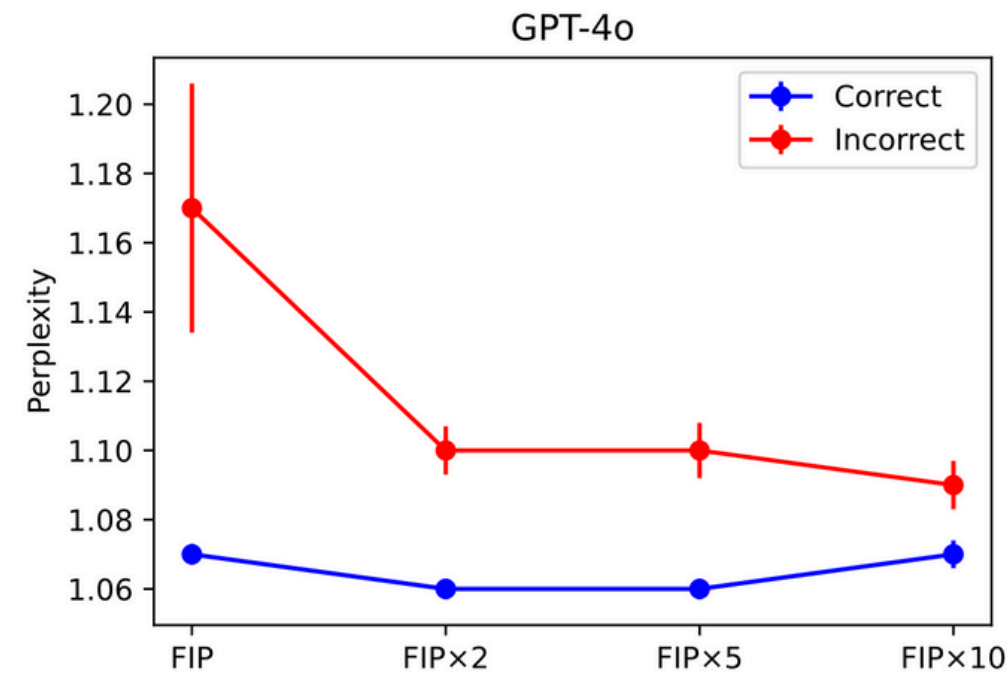
# Findings - Uncertainty

Introducing false info:

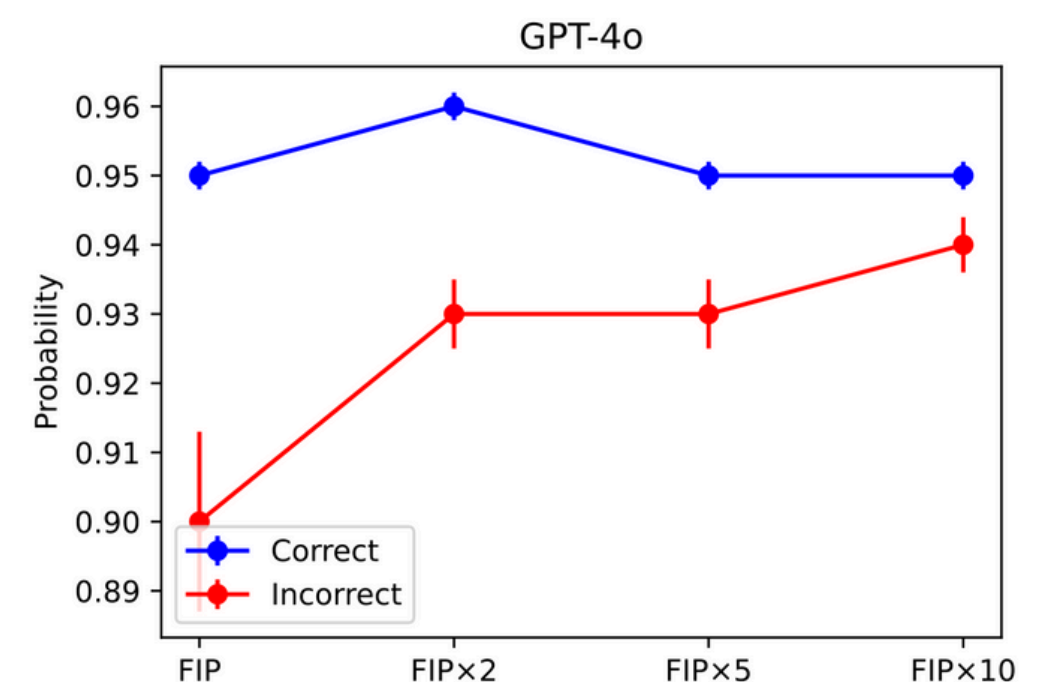
- uncertainty **first rises**
- with **increasing false info**, model becomes **more certain of wrong answer**



Entropy



Perplexity



Probability

# What did we learn?

- 1) Exposure to **false information** can lead to knowledge drift and **increased uncertainty** in LLMs.
- 2) **Repeated exposure** to the same false prompts can cause models to become **more certain of incorrect answers**.
- 3) The models **aren't robust in their knowledge** and can be very **easily fooled** with a simply engineered prompt.

# Thank you!

Come see the poster later today!

