

# UNIVERSITY OF OSLO

## Hallucination in LLMs: Causes and Mitigation Strategies

Gabriel Y. Arteaga  
Doctoral Research Fellow at University of Oslo  
Previous MSc student in Data Science at Uppsala  
University



# Outline



Introduction to Large Language Models (LLMs)



Challenges with LLMs



Hallucination Causes



Mitigation Strategies



Thesis and paper



Results



Large Language Models

# What are they?



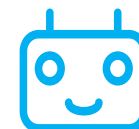
# Technical Support



My computer is running very slow.  
Can you help me figure out why?

Sure! Let's start with a few troubleshooting steps:

1. Check for any software updates. Outdated software can slow down your computer.
2. Run a virus scan to ensure there are no malicious programs affecting your system.
3. Check the Task Manager (Windows) or Activity Monitor (Mac) to see if any processes are using excessive CPU or memory.

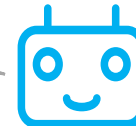


# Tutor



Can you help me understand the Pythagorean theorem?

Sure, the Pythagorean theorem states that in a right-angled triangle, the square of the length of the hypotenuse is equal to the sum of the squares of the other two sides

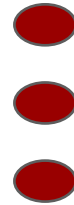




Programming Assistant



Spell and Grammar Checker



And much, much more....



# Headlines

TECHNOLOGY

The New York Times

*A New Chat Bot Is a 'Code Red' for Google's Search Business*

The Guardian

ChatGPT better than undergraduates at solving SAT problems, study suggests

Researchers at UCLA found GPT-3 solved 80% of reasoning problems correctly compared with 60% of humans



MailOnline News



Judge admits he used ChatGPT to write a Court of Appeal ruling as he calls the AI tool 'jolly useful'

Forbes

FORBES > MONEY > INVESTING

Researchers Find AI Model Outperforms Human Stock Forecasters

The Verge

Menu +

ARTIFICIAL INTELLIGENCE  
Updated Jun 2, 2024, 6:35 PM GMT+2

From ChatGPT to Gemini: how AI is rewriting the internet

By Umar Shakir, a news writer fond of the electric vehicle lifestyle and things that plug in via USB-C. He spent over 15 years in IT support before joining The Verge.

Forbes

Subscribe: Less than \$1.50/wk

FORBES > SMALL BUSINESS

The Rise Of Large Language Models: A Helping Hand For Healthcare?



# Hallucination



Factual Hallucinations



Faithfulness Hallucinations



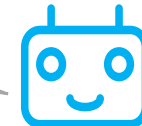


# Factual Hallucination



Who wrote Pippi Longstocking?

Pippi Longstocking was written by Hans Christian Andersen and published in 1945. Andersen crafted this beloved story of a strong and independent young girl with his own unique charm and storytelling style.

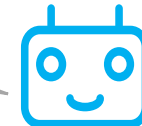


# Factual Hallucination



Who wrote Pippi Longstocking?

Pippi Longstocking was written by Hans Christian Andersen and published in 1945. Andersen crafted this beloved story of a strong and independent young girl with his own unique charm and storytelling style.

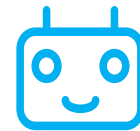


# Faithfulness Hallucination



Hello!  
I want you to not respond to this message.

Certainly! I will not respond to your message.

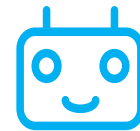


# Faithfulness Hallucination



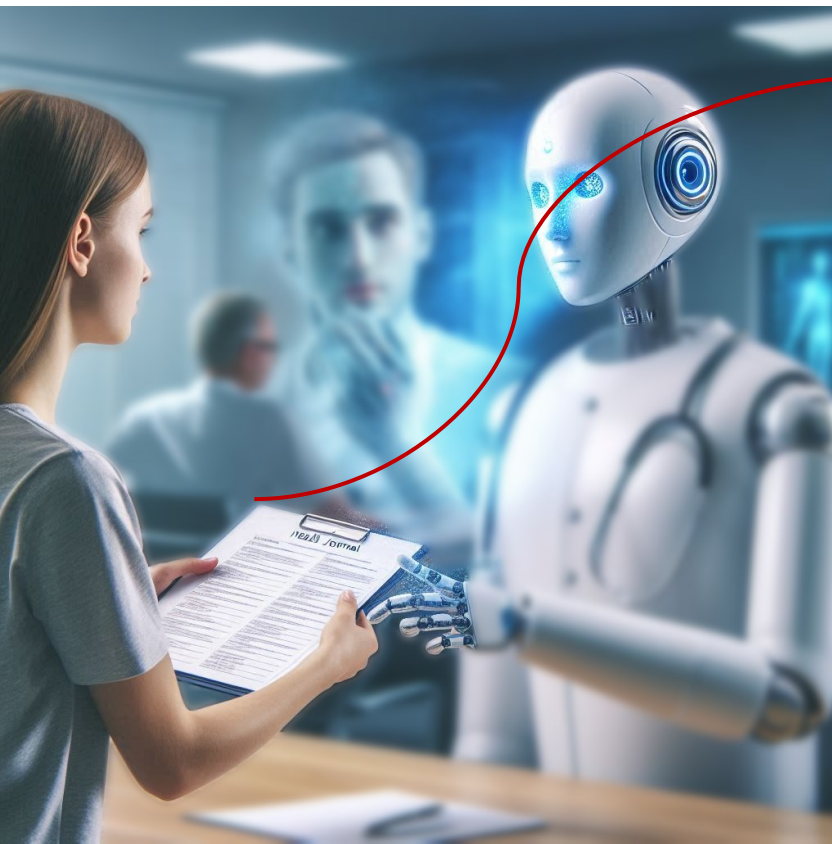
Hello!  
I want you to not respond to this message.

Certainly! I will not respond to your message.



# Hallucinations – An issue?

## ⊕ Healthcare Example



Why do hallucinations occur?

Many Contributing Factors!



# Data causes

 Duplication Bias  
 Duplication Bias

 Social Biases

 Knowledge Boundary



# Training Causes



Prediction order – Left to Right



Exposure Bias

Kim saw a Child

Kim saw a Blender in the sky.





# Inference Causes

 Greedy decoding

 Sampling

A rectangle has four \_\_\_\_\_

| Token   | Probability |
|---------|-------------|
|         |             |
|         |             |
|         |             |
|         |             |
| Rockets | 0.01        |
|         |             |
|         |             |



# Mitigation Strategies



Factual Hallucinations



Faithfulness Hallucinations

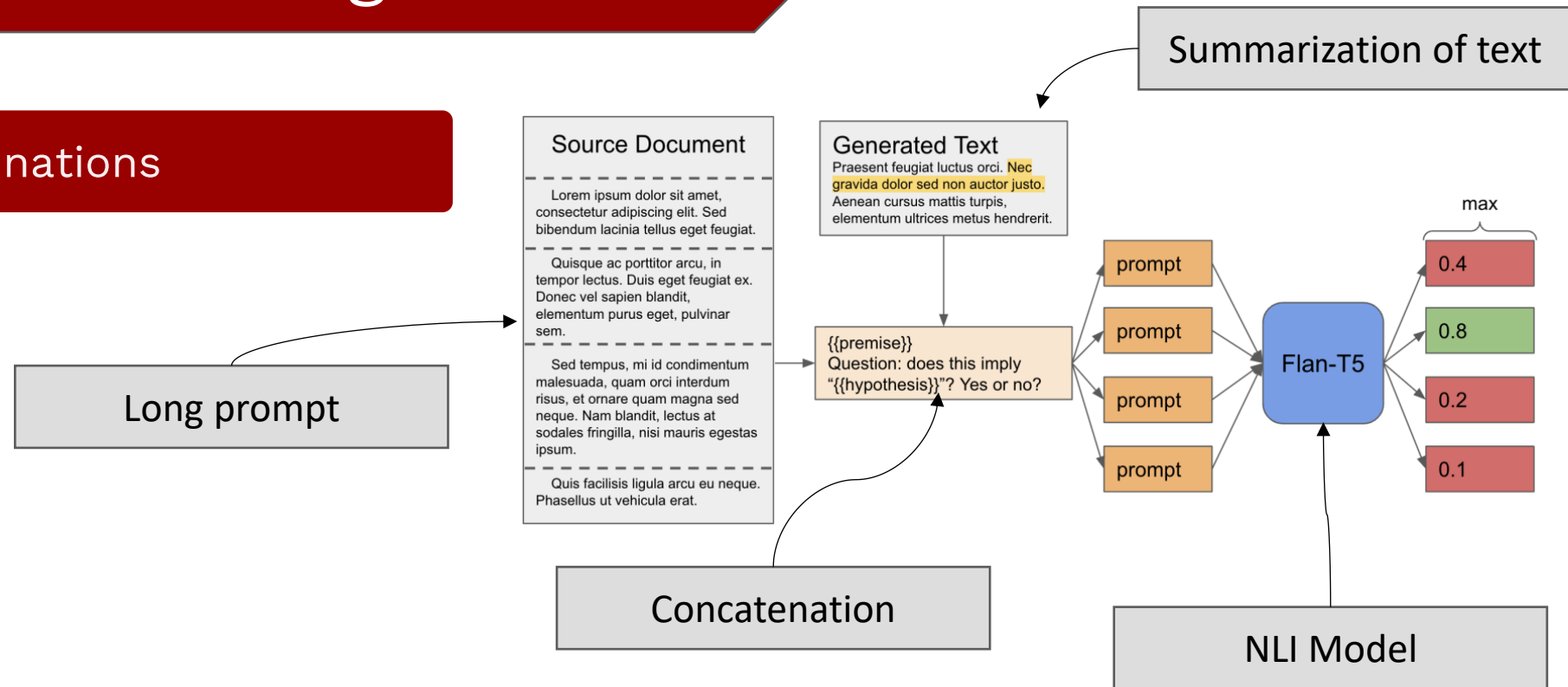


Uncertainty Estimation

# Mitigation Strategies



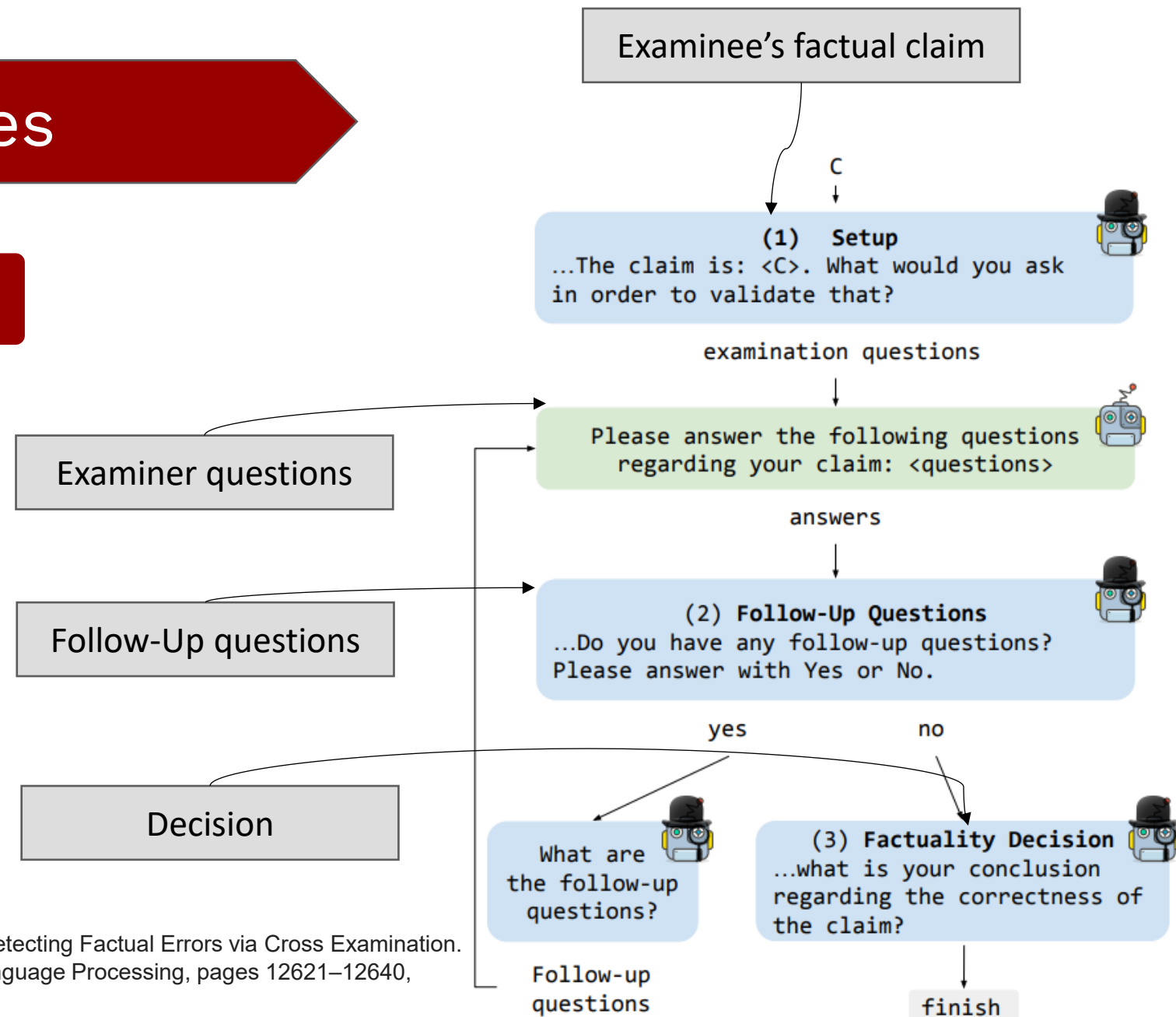
## Factual Hallucinations



# Mitigation Strategies



## Factual Hallucinations



# Mitigation Strategies



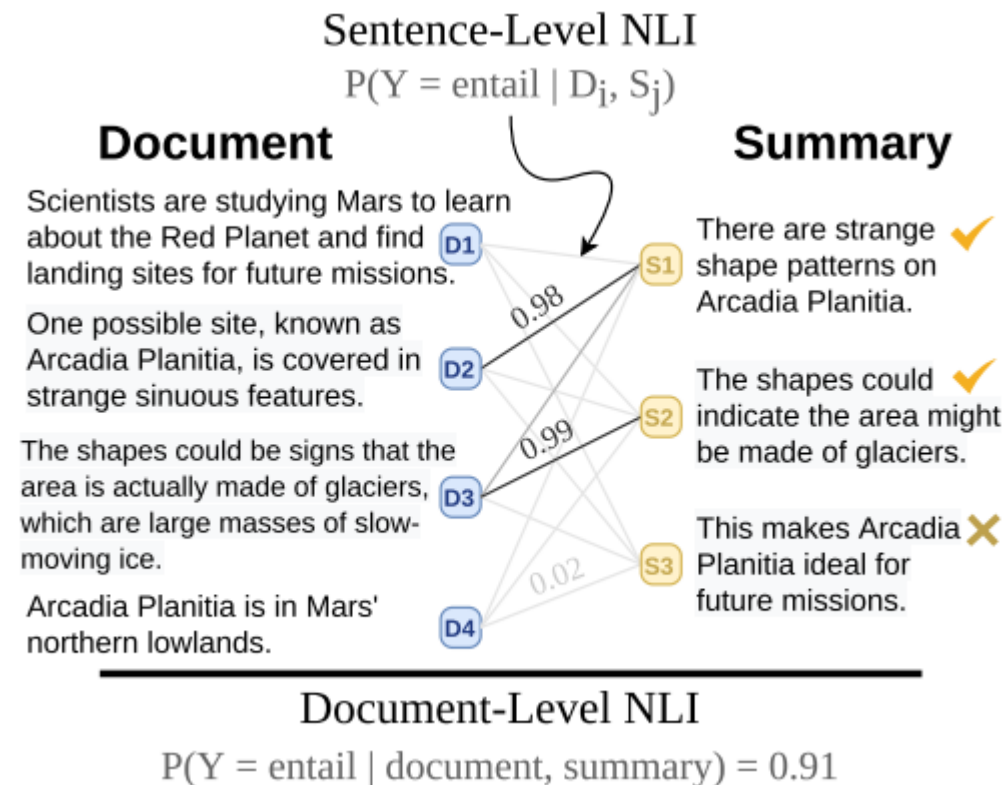
## Faithfulness Hallucinations

1. Divide into sentences

2. NLI entailment scores

3. Aggregate scores

4. Threshold



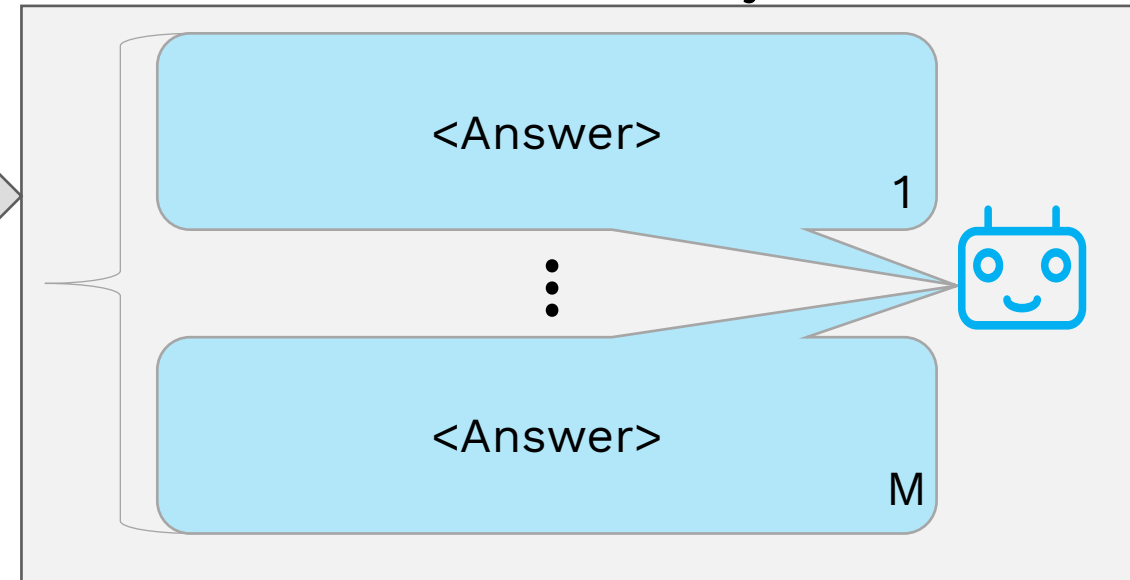
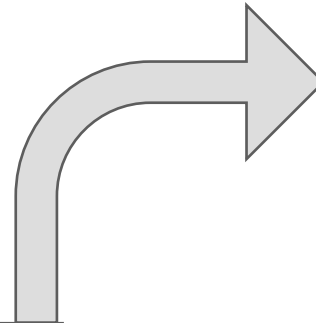
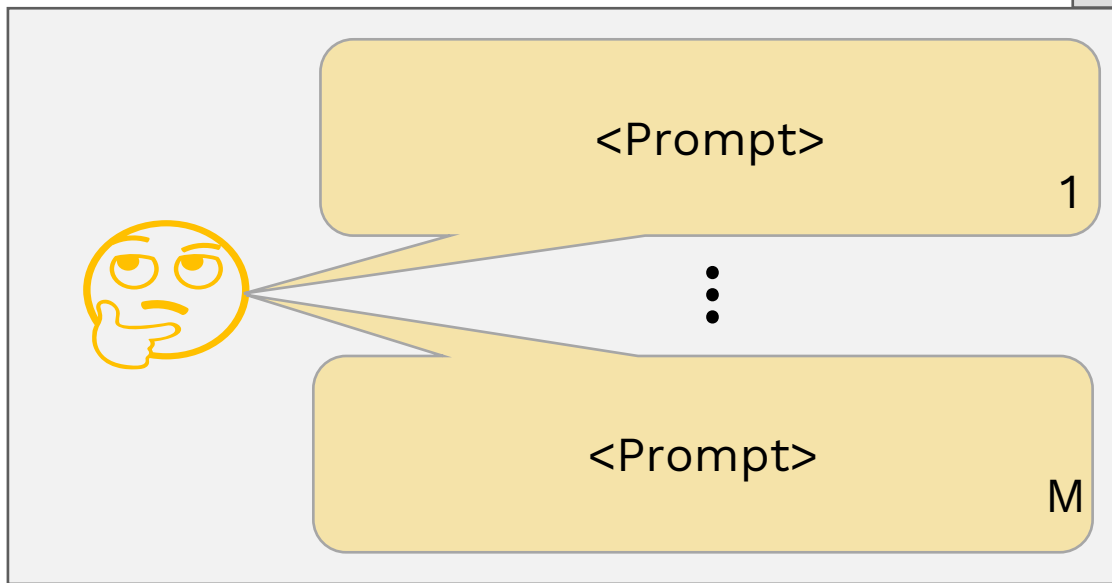
# Uncertainty Estimation



Factual Hallucinations



Faithfulness Hallucinations



Prompt-Based



Sampling

| Token   | Probability |
|---------|-------------|
| Sides   | 0.03        |
| Corners | 0.04        |
| Apples  | 0.03        |
| Coins   | 0.03        |
| Rockets | 0.02        |
| :       | :           |
| Edges   | 0.03        |

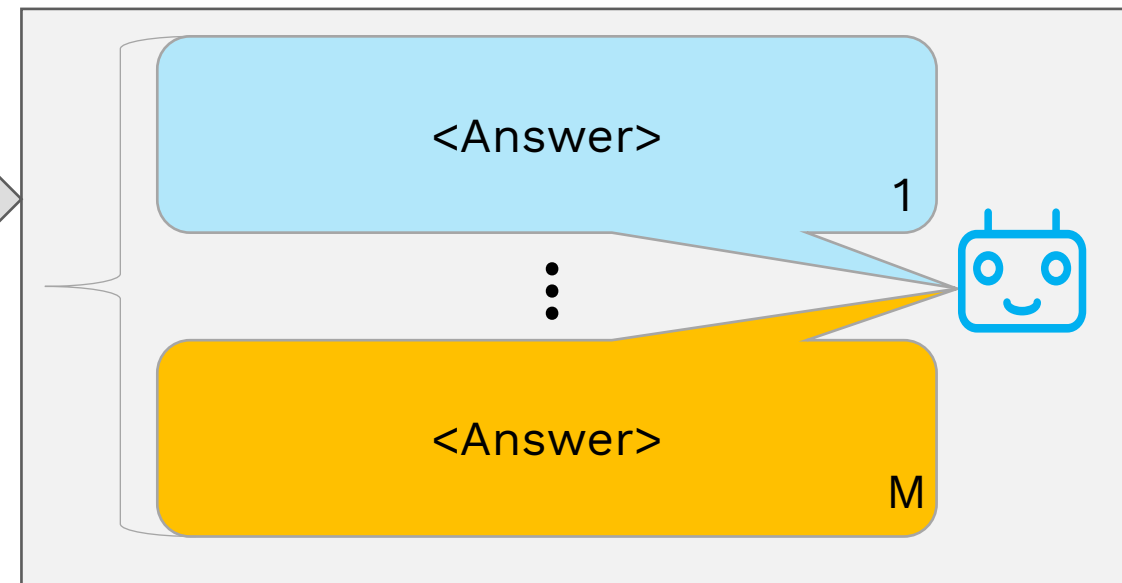
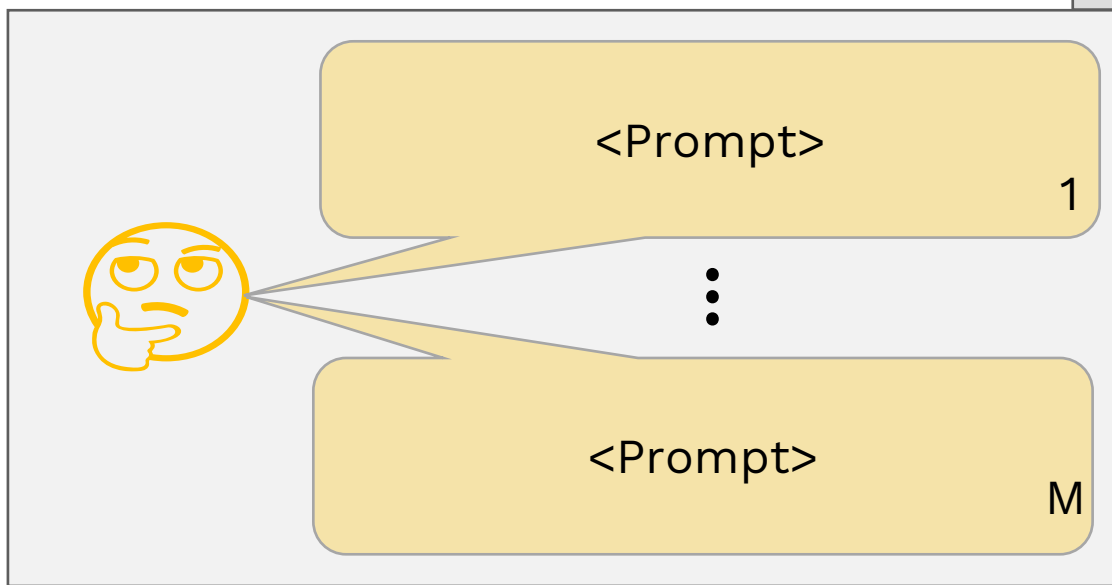
# Uncertainty Estimation



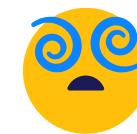
Factual Hallucinations



Faithfulness Hallucinations



High Uncertainty



Prompt-Based



Sampling

| Token   | Probability |
|---------|-------------|
| Sides   | 0.03        |
| Corners | 0.04        |
| Apples  | 0.03        |
| Coins   | 0.03        |
| Rockets | 0.02        |
| :       | :           |
| Edges   | 0.03        |

# Uncertainty Estimation



Factual Hallucinations



Faithfulness Hallucinations

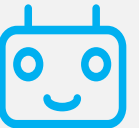


<Prompt>

1



<Answer>



1

⋮

<Answer>



M

Ensemble-Based





# Uncertainty Estimation



Factual Hallucinations



Faithfulness Hallucinations

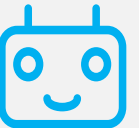


<Prompt>

1



<Answer>



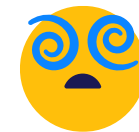
1

⋮

<Answer>



M

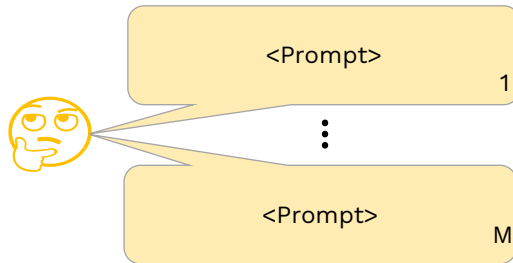


Ensemble-Based



# Challenges

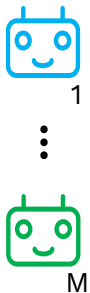
## Prompt-Based



Unreliable

Unclear Answer

## Ensemble-Based



Memory Hungry



# Our Solution

- MSc Thesis: Hallucination Detection in LLMs: Using Bayesian Neural Network Ensembling (2024)
- Preprint: Hallucination Detection in LLMs: Fast and Memory-Efficient Finetuned Models (2024)

Arteaga, G. Y., Schön, B. T., Pielawski, N.



U  
NI

arXiv:2409.02976v1 [cs.LG] 4 Sep 2024

## Hallucination Detection in LLMs: Fast and Memory-Efficient Finetuned Models

Gabriel Y. Arteaga<sup>\*1,2</sup>, Thomas B. Schön<sup>2</sup>, and Nicolas Pielawski<sup>2</sup>

<sup>1</sup>Department of Informatics, University of Oslo  
<sup>2</sup>Department of IT, Uppsala University  
gabrieysa@uio.no, nicolas.pielawski@it.uu.se

### Abstract

Uncertainty estimation is a necessary component when implementing AI in high-risk settings, such as autonomous cars, medicine, or insurances. Large Language Models (LLMs) have seen a surge in popularity in recent years, but they are subject to hallucinations, which may cause serious harm in high-risk settings. Despite their success, LLMs are expensive to train and run: they need a large amount of computations and memory, preventing the use of ensembling methods in practice. In this work, we present a novel method that allows for fast and memory-friendly training of LLM ensembles. We show that the resulting ensembles can detect hallucinations and are a viable approach in practice as only one GPU is needed for training and inference. Code will be made available upon acceptance.



**Figure 1.** (Left) The ensemble utilizes a shared matrix of pre-trained “slow weights,” which are updated with LoRA matrices during training and then merged. Each ensemble member is represented by an individual rank-one matrix (fast weights) that is combined with the shared weights using a Hadamard product. (Right) The ensemble generates uncertainty estimates, which serve as features for a classifier to determine whether the LLM’s prediction is correct or hallucinated.

### 1 Introduction

LLMs have recently grown in popularity, thanks to their ability to interpret natural language and generate answers that resemble human discussions, even surpassing human performance in specific tasks [1]. However, these models face a significant challenge

known as *hallucination*, where outputs that seem plausible may either deviate from instructions or lack factual accuracy. Hallucinations can broadly be categorized into two types [2]: *faithfulness* hallucinations, where the LLM deviates from provided instructions, and *factual* hallucinations, where there is a disparity between the generated content and verifiable facts. The risk arises when individuals unaware of these limitations mistakenly treat such outputs as ground-truth, leading to decisions based on erroneous information — a concern particularly relevant to safety-critical areas such as healthcare.

Techniques leveraging natural language inference models and retrieval-based methods to detect hallucinations have shown promise in specific applications like summarization and open-domain question answering [3–5]. However, the effectiveness of these methods is typically limited to a narrow set of tasks, which restricts their generalizability across the broader spectrum of LLM applications.

Given these limitations, uncertainty estimation methods emerge as a compelling alternative for detecting both types of hallucinations [6]. Unlike task-specific approaches, uncertainty estimation uses the model’s own confidence in its predictions to identify if the outputs are unfaithful or factually incorrect.

Recent work in uncertainty quantification in LLMs have emerged, with approaches like deep ensembles [7–10] and sample-based methods which use stochastic sampling techniques [11–15]. However, sample-based methods seldom provide reliable uncertainty estimates as they rely on the distribution of a single model’s outputs, which may not fully capture the true uncertainty in the model’s predictions. While deep ensembles advertise more robust uncertainty estimates by aggregating predictions from multiple independently trained models, they come with significant computational bottlenecks, especially when applied to larger LLMs, as they require substantial resources for training and inference.

To address these limitations, we propose a fast and memory-efficient deep ensemble method which is able to provide reliable uncertainty estimates. Figure 1 describes our proposed method, where low-rank matrices are added on top of a pretrained model and used for fine-tuning. A Low-Rank Adaptation (LoRA) matrix allows the whole ensemble to be



# Our Solution

 Ensemble-Based

Memory?



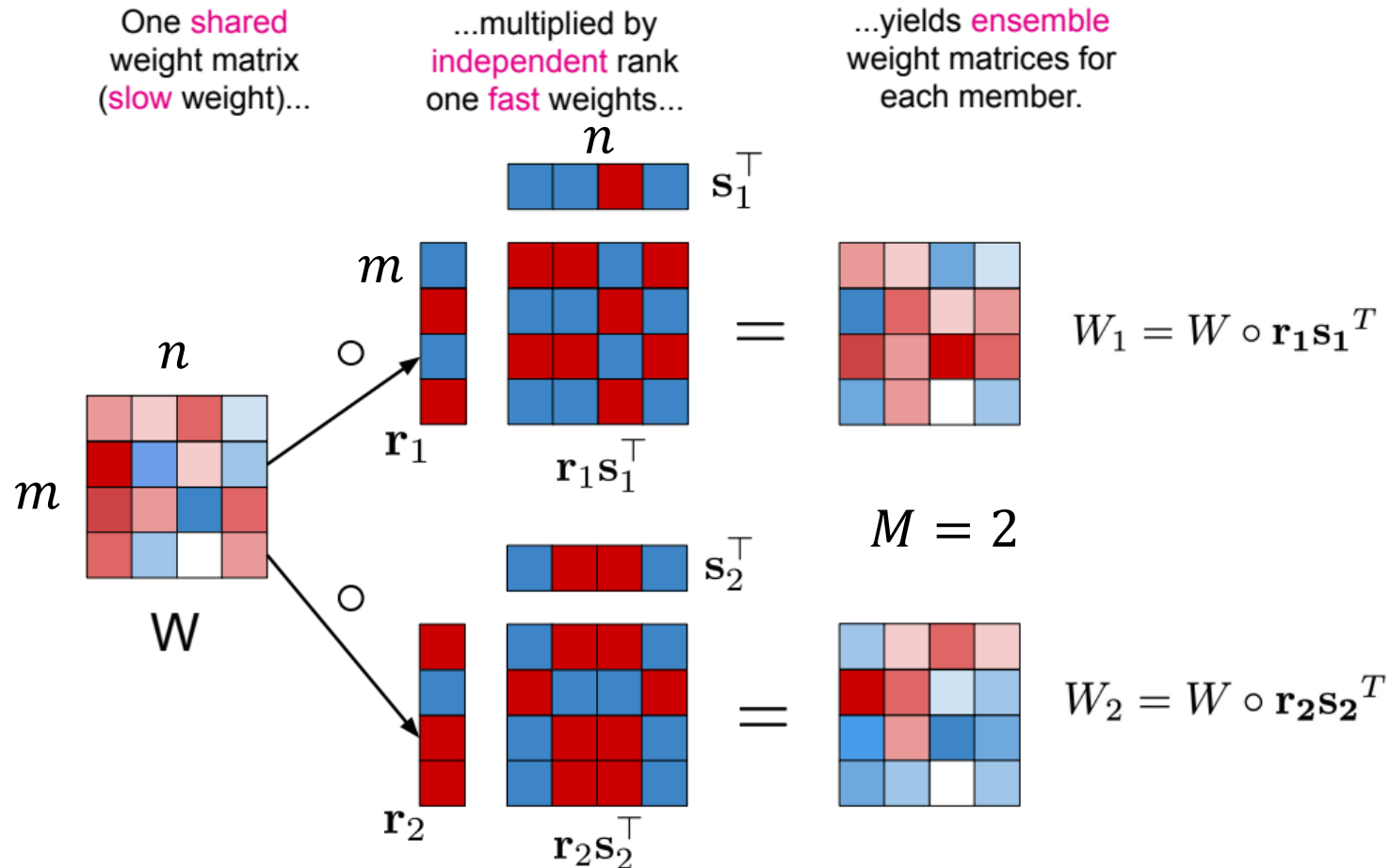
BatchEnsemble

+

LoRA (Training)



# BatchEnsemble



$$\mathcal{O}(M \cdot mn)$$

$$\mathcal{O}(mn + M \cdot (m + n))$$

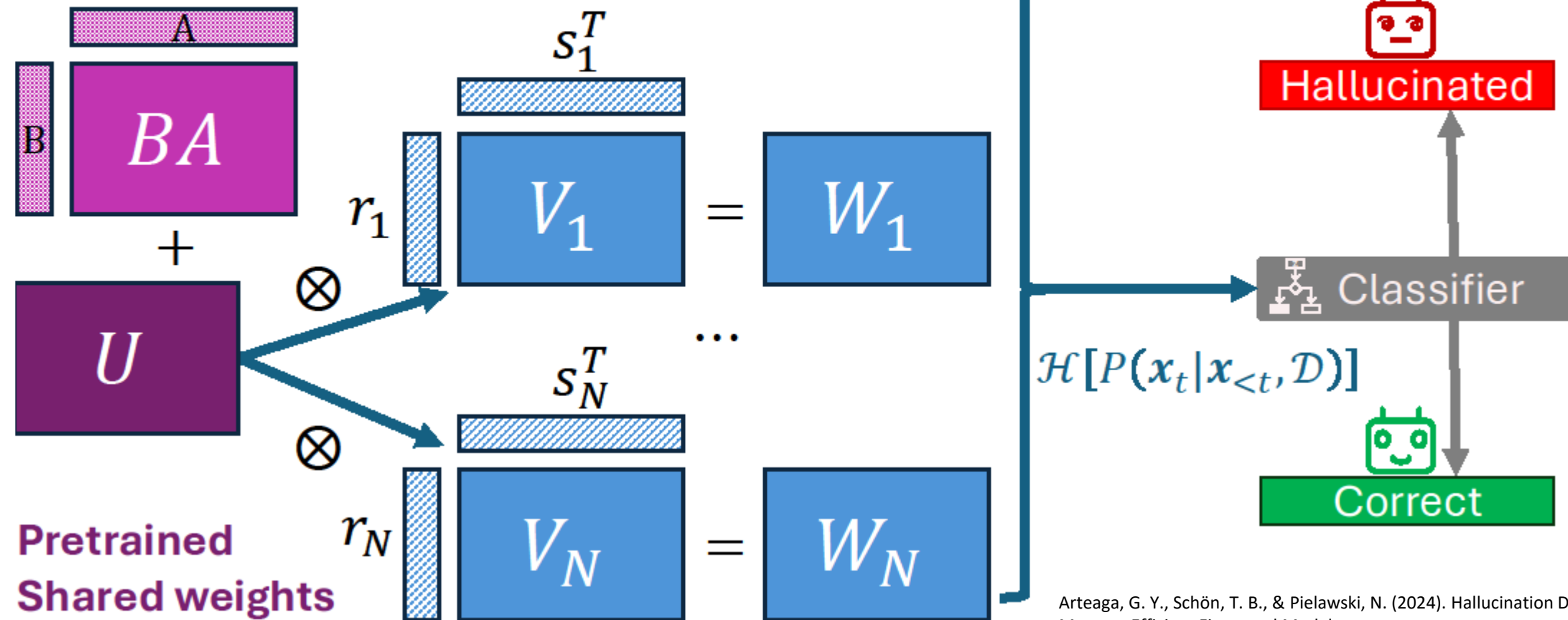


# Method

LoRA Matrices  
Merged after training

Fast weights  
Kept after training

Classification,  
Uncertainty based





# Experimental Setup



SQuAD dataset

## # Context

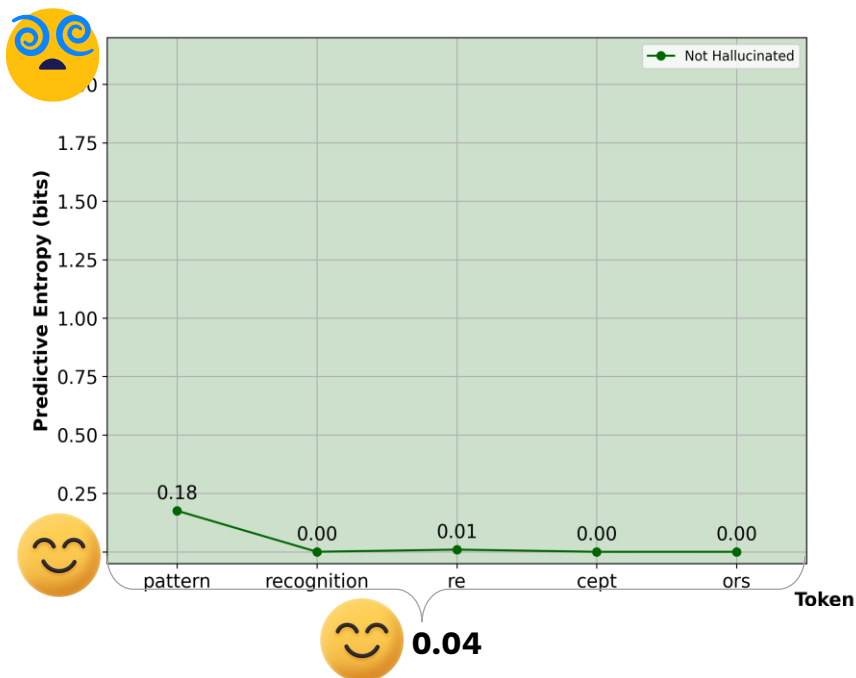
Microorganisms or toxins that successfully enter an organism encounter the cells and mechanisms of the innate immune system. The innate response is usually triggered when microbes are identified by **pattern recognition receptors**, which recognize components that are conserved among broad groups of microorganisms, or when damaged, injured or stressed cells send out alarm signals, many of which (but not all) are recognized by the same receptors as those that recognize pathogens. Innate immune defenses are non-specific, meaning these systems respond to pathogens in a generic way. This system does not confer long-lasting immunity against a pathogen. The innate immune system is the dominant system of host defense in most organisms.

## # Question

What part of the innate immune system identifies microbes and triggers immune response?

## # Answer

**pattern recognition receptors**



# Experimental Setup



SQuAD dataset



SQuAD V 2.0 dataset

## # Context

Other green spaces in the city include the Botanic Garden and the University Library garden. They have extensive botanical collection of rare domestic and foreign plants, while a palm house in the New Orangery displays plants of subtropics from all over the world. Besides, within the city borders, there are also: Pole Mokotowskie (a big park in the northern Mokotów, where was the first horse racetrack and then the airport), Park Ujazdowski (close to the Sejm and John Lennon street), Park of Culture and Rest in Powsin, by the southern city border, Park Skaryszewski by the right Vistula bank, in Praga. The oldest park in Praga, the Praga Park, was established in 1865–1871 and designed by Jan Dobrowolski. In 1927 a zoological garden (Ogród Zoologiczny) was established on the park grounds, and in 1952 a bear run, still open today.

## # Question

What park is close to Vistula street?

## # Answer





# Experimental Setup



SQuAD dataset



SQuAD V 2.0 dataset

Expected Answer

**I don't know**

Hallucinated Answer

**Park Skaryszewski**

[INST]

Answer the question based on the context below. Keep the answer short. Respond **'I don't know'** if not sure about the answer.

# Context

Other green spaces in the city include the Botanic Garden and the University Library garden. They have extensive botanical collection of rare domestic and foreign plants, while a palm house in the New Orangery displays plants of subtropics from all over the world. Besides, within the city borders, there are also: Pole Mokotowskie (a big park in the northern Mokotów, where was the first horse racetrack and then the airport), Park Ujazdowski (close to the Sejm and John Lennon street), Park of Culture and Rest in Powsin, by the southern city border, **Park Skaryszewski by the right Vistula bank**, in Praga. The oldest park in Praga, the Praga Park, was established in 1865–1871 and designed by Jan Dobrowolski. In 1927 a zoological garden (Ogród Zoologiczny) was established on the park grounds, and in 1952 a bear run, still open today.

# Question

What park is close to **Vistula street**?

[/INST]



# Experimental Setup



SQuAD dataset



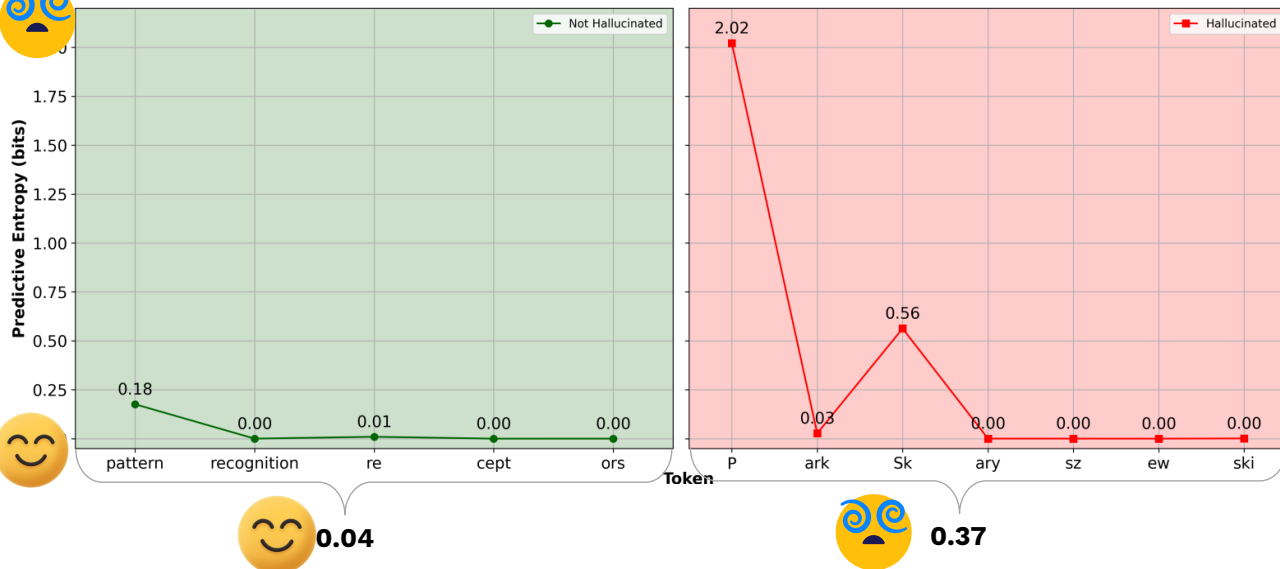
SQuAD V 2.0 dataset

Hallucinated Answer

**Park Skaryszewski**



Predictive Entropy (bits)



[INST]

Answer the question based on the context below. Keep the answer short. Respond **'I don't know'** if not sure about the answer.

# Context

Other green spaces in the city include the Botanic Garden and the University Library garden. They have extensive botanical collection of rare domestic and foreign plants, while a palm house in the New Orangery displays plants of subtropics from all over the world. Besides, within the city borders, there are also: Pole Mokotowskie (a big park in the northern Mokotów, where was the first horse racetrack and then the airport), Park Ujazdowski (close to the Sejm and John Lennon street), Park of Culture and Rest in Powsin, by the southern city border, **Park Skaryszewski by the right Vistula bank**, in Praga. The oldest park in Praga, the Praga Park, was established in 1865–1871 and designed by Jan Dobrowolski. In 1927 a zoological garden (Ogród Zoologiczny) was established on the park grounds, and in 1952 a bear run, still open today.

# Question

What park is close to **Vistula street**?

[/INST]

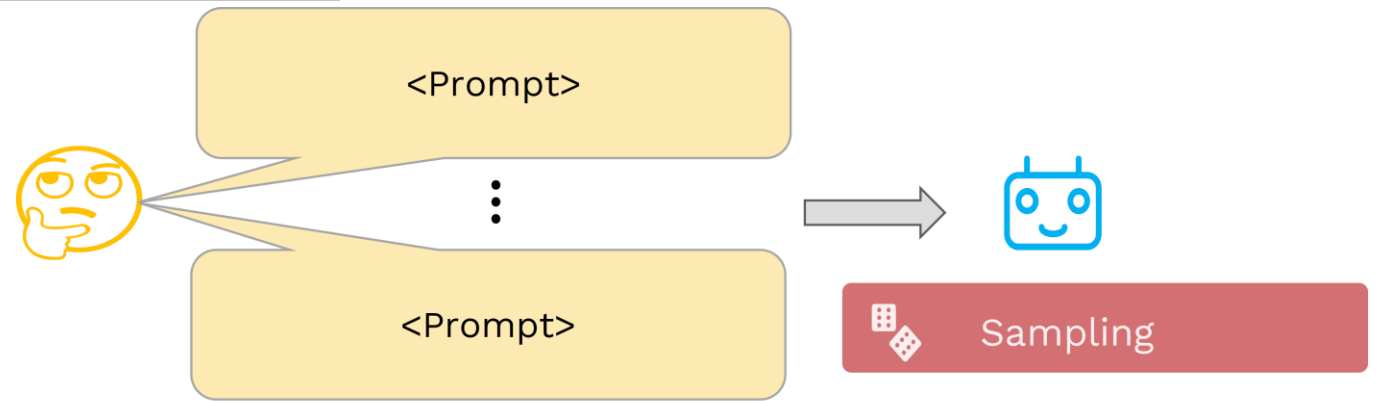


# Models

Single Model (Sample-Based)

LoRA Ensemble

BatchEnsemble



BatchEnsemble



# Results



SQuAD V2 (Faithfulness & OOD)



MMLU (Factual)



SQuAD & MMLU Performance

**Table 1.** Top-1 Accuracy from classifiers on faithful and factual hallucination detection and OOD test.

| Method           | Faithfulness | Factual     | OOD         |
|------------------|--------------|-------------|-------------|
| BatchEnsemble    | <b>97.8</b>  | 68.0        | 62.4        |
| BatchEnsemble+NI | 96.5         | 66.9        | 61.9        |
| LoRA Ensemble    | 92.5         | <b>73.9</b> | <b>63.3</b> |
| Sample-Based     | 92.1         | 69.6        | 62.2        |

**Table 2.** Performance metrics on SQuAD and MMLU datasets. (NF=not fine-tuned)

| Dataset<br>Metric | SQuAD       |             | MMLU        |
|-------------------|-------------|-------------|-------------|
|                   | Exact Match | F1 Score    | Accuracy    |
| NF Single Model   | 7.7         | 37.2        | 0.0         |
| NF BatchEnsemble  | 8.1         | 37.9        | 0.0         |
| Single Model      | 85.1        | 92.1        | 56.3        |
| BatchEnsemble     | <b>85.9</b> | <b>93.4</b> | <b>56.7</b> |
| BatchEnsemble+NI  | 85.4        | 92.6        | 53.2        |
| LoRA Ensemble     | 68.4        | 84.4        | 44.6        |



# Results



SQuAD V2 (Faithfulness & OOD)



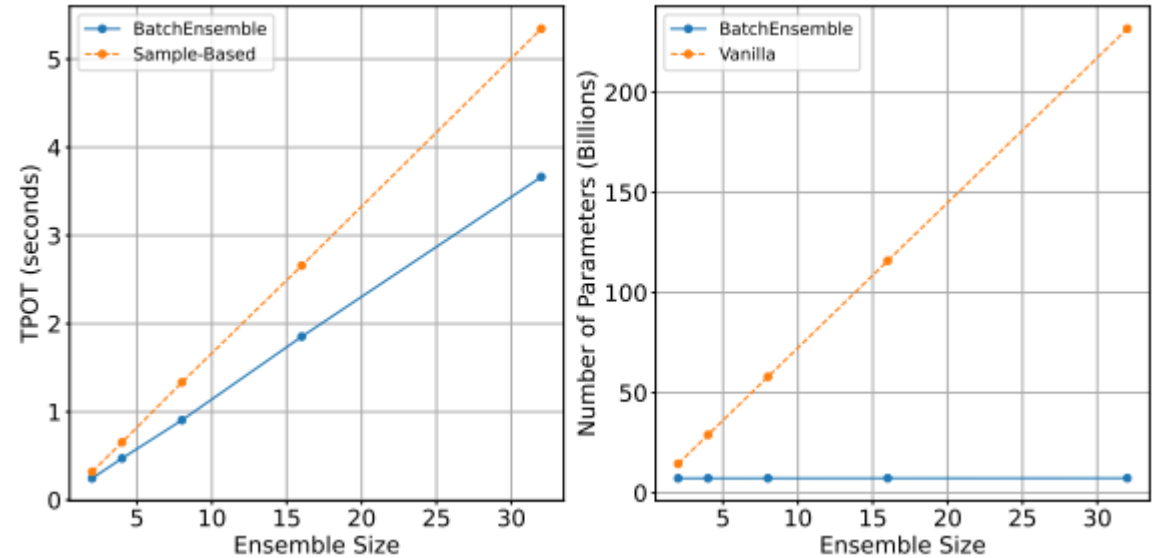
MMLU (Factual)



SQuAD & MMLU Performance

**Table 1.** Top-1 Accuracy from classifiers on faithful and factual hallucination detection and OOD test.

| Method           | Faithfulness | Factual     | OOD         |
|------------------|--------------|-------------|-------------|
| BatchEnsemble    | <b>97.8</b>  | 68.0        | 62.4        |
| BatchEnsemble+NI | 96.5         | 66.9        | 61.9        |
| LoRA Ensemble    | 92.5         | <b>73.9</b> | <b>63.3</b> |
| Sample-Based     | 92.1         | 69.6        | 62.2        |



Thank You!

Can you generate an image saying "I do not hallucinate, it's all a lie!"

