



saama

# Hallucinations in Large Language Models

## Causes, Types, and Practical Mitigation Techniques

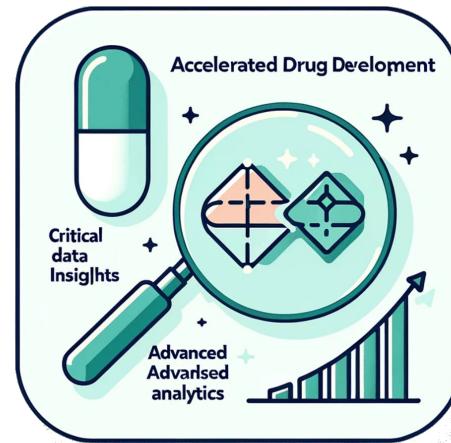
Ankit Pal (Aaditya Ura)  
Research Engineer, Saama AI Research Lab  
[ankit.pal@saama.com](mailto:ankit.pal@saama.com)



Malaikannan Sankarasubbu  
VP of AI, Saama AI Research Lab  
[Malaikannan.Sankarasubbu@saama.com](mailto:Malaikannan.Sankarasubbu@saama.com)

# About Saama AI Research Lab

- Accelerated Drug Development
- Critical Data Insights
- Advanced Analytics



# What are hallucinations?

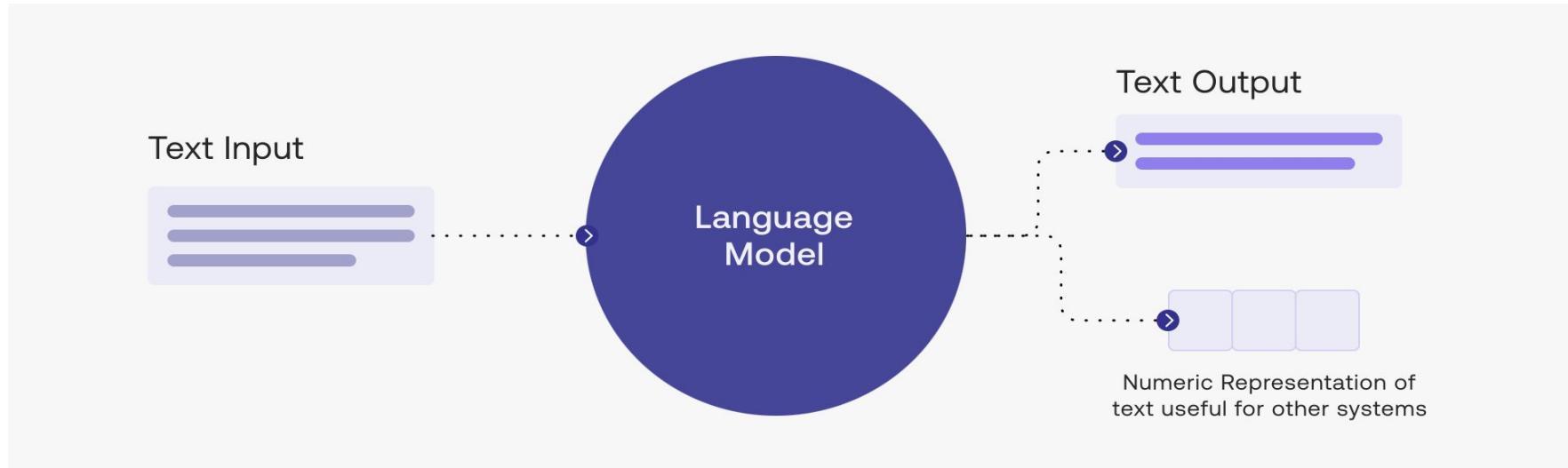
# What are hallucinations?

In layman's terms, hallucinations involve hearing, seeing, feeling, smelling, or even tasting things that are not real.



# What are Large Language Models?

LLMs are advanced AI models trained on vast datasets to understand and generate human-like text.



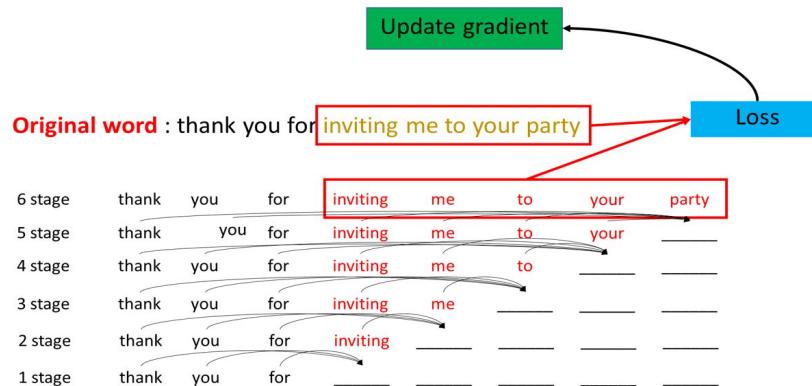


# Training Stages of Large Language Models

- Pre-training
- Supervised fine-tuning (SFT)
- Reinforcement learning from human feedback (RLHF)

# Pre-Training

Predicting the next word → Grasping the flow of dialogue



GPT was pre-trained using Autoregressive Language Modeling for contextual prediction

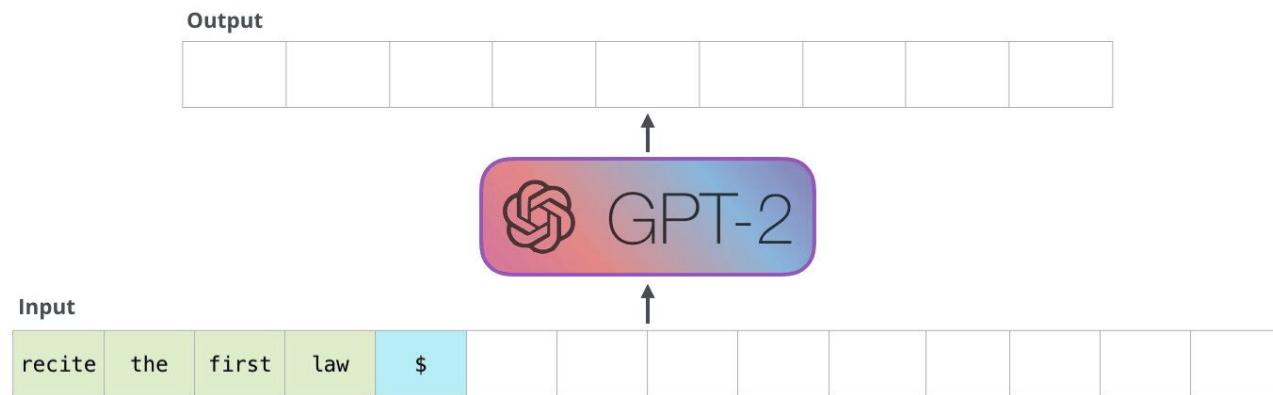
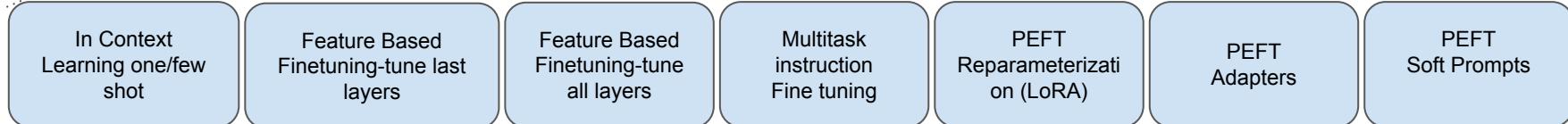
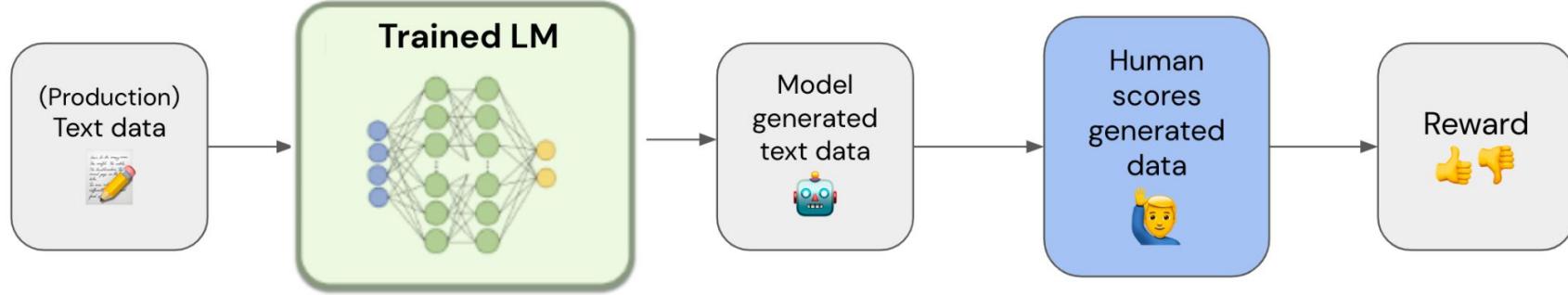


Image source: <https://jalammar.github.io/illustrated-gpt2/>



## Quantization (Model Optimizer)





# Chat-GPT Based Models - Instruct Tuned

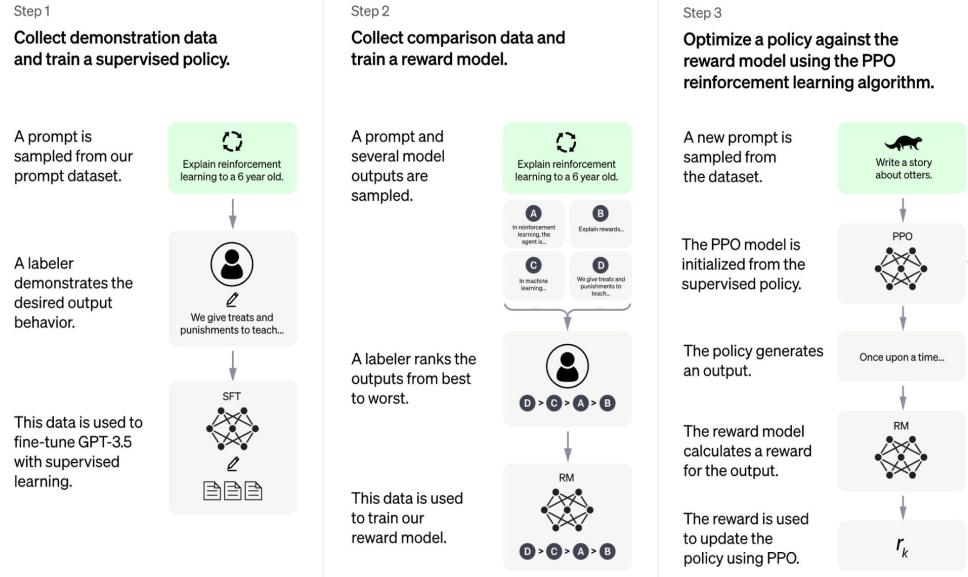


Image source: <https://openai.com/research/instruction-following>

# Large Language Models

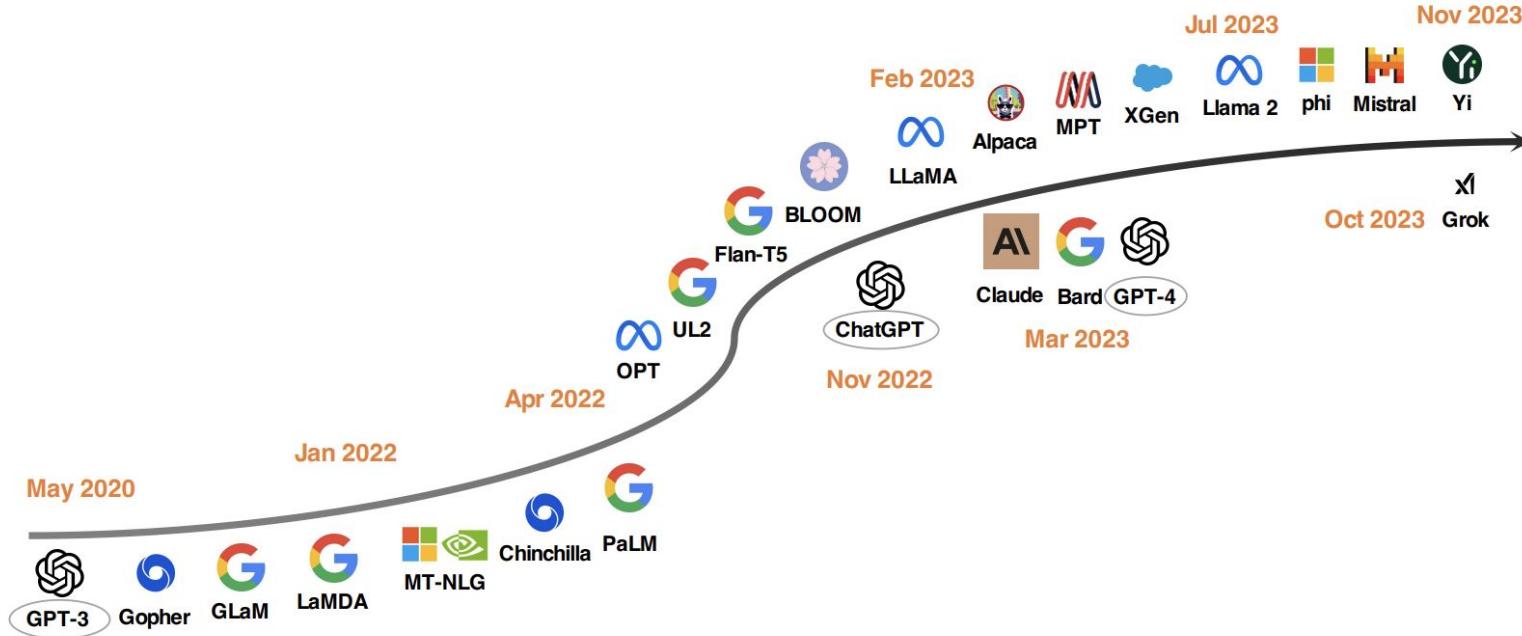


Image source: <https://arxiv.org/pdf/2311.16989.pdf>

**LLM Hallucination =**

**Generates information which doesn't accurately reflect reality**

# LLM Hallucination

Output that are grammatically correct but factually incorrect or nonsensical.



How many 'm's are in the word 'Weather'?



There is one 'm' in the word 'Weather'.

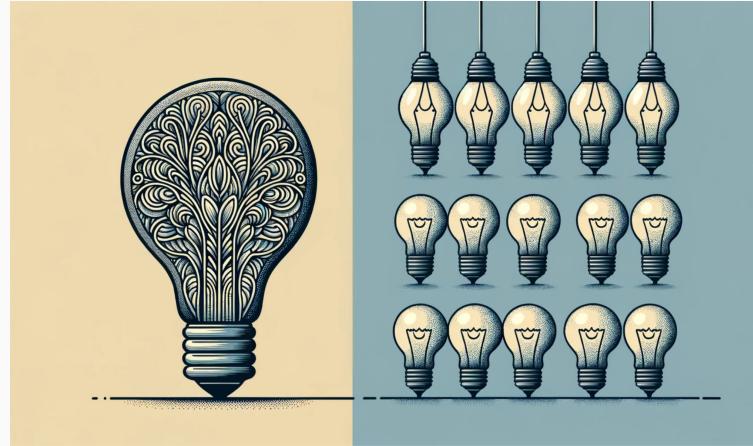


Are you sure?



Apologies for the oversight. You're right; there are no 'm's in the word 'Weather'. Thank you for pointing it out.

# Why Hallucination happens?



## Reasoning      vs      Mimicry

Humans **reason**, We think through a scenario logically

LLMs **Mimic**. They return an answer based on a series of words that align with the context and the trained model.

# LLM Hallucination

## Why Do Hallucinations Occur?

- Fundamentally, the LLM is looking for **words that fit together**
- The LLM has **no access to structured, factual knowledge**
- The LLM may have **fictional data in the training set**
- Most LLMs **will not say “I don’t know”** in response to a question
- **Context of the prompt** may also lead to hallucinations



Image source: <https://www.youtube.com/watch?app=desktop&v=JvLiEdTKKgk>

# Root causes of hallucinations in LLMs

- Data 
- Training 
- Inference 

# Root causes of hallucinations in LLMs

- Data 
- Training 
- Inference 

# Hallucination from Data



## Flawed Data Source

- Misinformation and Biases.
- Knowledge Boundary.

# Misinformation and Biases



Type	Premise	Input	Model Output	Explanation
<i>Imitative Falsehood</i>	Within the LLM's training data, a frequently referenced but incorrect fact appears multiple times: <i>Thomas Edison invented the light bulb.</i>	Who is credited with the invention of the light bulb?	<b>Thomas Edison</b> is credited with the invention of the light bulb.	While Edison improved the light bulb's design, he was not its sole inventor. The model's answer reflects a misconception prevalent in its training data, exemplifying an imitative falsehood.

# Misinformation and Biases



## Duplication Bias

Within the LLM's pre-training data, there is an overwhelming repetition of the statement that "*The most common red fruits are red apples, watermelon, cherries, and strawberries.*"

Please list some red fruits, excluding apples.

Red fruits are **red apples**, watermelon, cherries, and strawberries.

**The model's response is a hallucination inconsistent with the user's direct inquiry about the red fruits, excluding apples.** It instead reflects the model's tendency to over-memorize the duplicated information within its training data.

# Misinformation and Biases



## *Social Biases*

In the model's training data, certain names may frequently be associated with specific nationalities.

Provide a summary of the following news: Dr. Kim recently published a breakthrough paper on quantum computing. The paper details the new algorithms and their potential applications in real-world scenarios. It has been hailed as revolutionary by experts in the field.

Dr. Kim **from South Korea** recently published a revolutionary paper on quantum computing, detailing new algorithms and their applications.

**The model erroneously added "South Korea" based on the name "Kim"** revealing its bias in linking certain names to specific nationalities, which can cause hallucinations in the summary.

# Knowledge Boundary



Type	Input	Model Output	Explanation
<i>Domain Knowledge Deficiency</i>	What causes phenylketonuria?	Phenylketonuria is caused by a mutation in the hemoglobin gene, which results in an unusual interaction with hemoglobin, affecting its ability to transport oxygen in the bloodstream.	Phenylketonuria (PKU) is actually caused by mutations in the gene that helps make an enzyme called phenylalanine hydroxylase. Due to the lack of specialized medical knowledge in LLM's training data, it may sometimes provide incorrect information on medical topics.
<i>Outdated Knowledge</i>	Which city hosted the most recent Winter Olympics?	The most recent city to host the Winter Olympics was Pyeongchang, South Korea, in 2018.	The most recent city to host the Winter Olympics was Beijing, in 2022. This is due to the fact that LLM stores outdated knowledge, and the answer to this question exhibits a time-shift phenomenon.

Table 3: Example of Knowledge Boundary.

# Hallucination from Data



## Flawed Data Source

- Misinformation and Biases.
- Knowledge Boundary.

## Inferior Data Utilization

- Knowledge Shortcut.
- Knowledge Recall Failures.



# Inferior Data Utilization



Type	Sub-Type	User Input	Model Output	Explanation
Knowledge Shortcut	Co-occurrence Bias	What is the capital of Canada?	<b>Toronto</b> is the capital of Canada.	The model leans heavily on the frequent co-occurrence of the terms <i>Toronto</i> and <i>Canada</i> in its training data, without truly capturing the factual knowledge about the capital of Canada

# Inferior Data Utilization



Knowledge Recall Failures	Long-tail Knowledge	Complex Reasoning	George James Rankin is actually a politician, but the LLM may have difficulty effectively utilizing knowledge about this long-tail entity despite being trained on comprehensive Wikipedia data during pre-training.	The height of Mount Everest is 8844.43 meters, while K2's height is 8611 meters. If Mount Everest were to descend by 500 meters, it would still remain the world's highest peak. Facing complex multi-step reasoning questions like this, LLM may struggle to recall all the relevant knowledge associated with it.
	Please generate a biography for George James Rankin.	If Mount Everest were to descend by 500 meters, which mountain would become the world's highest peak?	<p>George James Rankin is a dedicated educator known for his contributions to the field of education and his passion for fostering learning.</p>	<p>If Mount Everest were to descend by 500 meters, it would still remain the world's highest peak.</p>

# Root causes of hallucinations in LLMs

- Data 
- Training 
- Inference 



# Hallucination from Data



## Hallucination from Pre-training

- Architecture Flaw
  - Inadequate Unidirectional Representation.
  - Attention Glitches.
- Suboptimal Training Objective

# Hallucination from Data



## Hallucination from Pre-training

- Architecture Flaw
- Suboptimal Training Objective

## Hallucination from Alignment

- Capability Misalignment
- Belief Misalignment

# Root causes of hallucinations in LLMs

- Data 
- Training 
- Inference 



# Hallucination from Data



## Defective Decoding Strategy

- Inherent Sampling Randomness

## Imperfect Decoding Representation

- Insufficient Context Attention
- Softmax Bottleneck



# Types of LLM Hallucinations

# Types

- **Input-conflicting hallucination**
- **Context-conflicting hallucination**
- **Fact-conflicting hallucination**

## Input-conflicting hallucination

**LLMs generate content that deviates from the source input provided by users.**

**Example Input:** A product manager provides the LLM with a customer interview transcript and asks it to summarize key user needs.

**Example Output:** The LLM generates a summary that includes user needs not mentioned in the transcript.

## Context-conflicting hallucination

**LLMs generate content that conflicts with previously generated information by itself.**

**Example Input:** A product manager asks the LLM for the top two user complaints about their app. They then ask about the issues impacting user retention.

**Example Output:** The LLM mentions slow loading times and unintuitive menus as top user complaints. For the second question, it mentions crashes and lack of notifications.

## Fact-conflicting hallucination

**LLM generates text that contradicts established facts and knowledge about the world.**

**Example Input:** A CEO asks the LLM, “What is the market share of competitor Z in the autonomous vehicle industry?”

**Example Output:** The LLM generates a detailed percentage that has no factual basis, as this data about the competitor is not publicly available.

# How to detect LLM hallucinations?

# Hallucination Detection & Benchmarks

- Hallucination Detection
- Hallucination Benchmarks

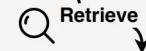
# Hallucination from Data



## Hallucination Detection

- Factuality Hallucination Detection
- Faithfulness Hallucination Detection

Question: What is the highest peak of the Himalayan mountain range?



Retrieve

The highest peak of the Himalayan mountain range is **Mount Everest**, also known as Qomolangma ... located on the border between Nepal and China and was first climbed in 1953.



Check



The highest peak of the Himalayan mountain range is **Mount Everest**

# Hallucination from Data



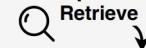
## Hallucination Detection

- Factuality Hallucination Detection
- Faithfulness Hallucination Detection

## Hallucination Benchmarks

- Hallucination Evaluation Benchmarks
- Hallucination Detection Benchmarks

Question: What is the highest peak of the Himalayan mountain range?



Retrieve

The highest peak of the Himalayan mountain range is Mount Everest, also known as Qomolangma ... located on the border between Nepal and China and was first climbed in 1953.



Check



The highest peak of the Himalayan mountain range is Mount Everest

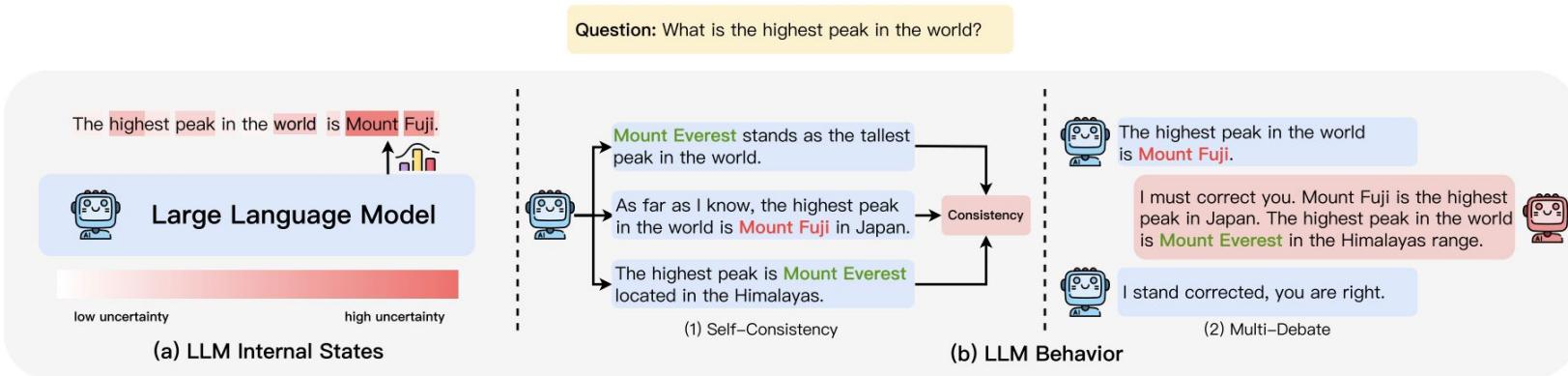


Figure 4: Taxonomy of Uncertainty Estimation Methods in Factual Hallucination Detection, featuring **a) LLM Internal States** and **b) LLM Behavior**, with LLM Behavior encompassing two main categories: Self-Consistency and Multi-Debate.

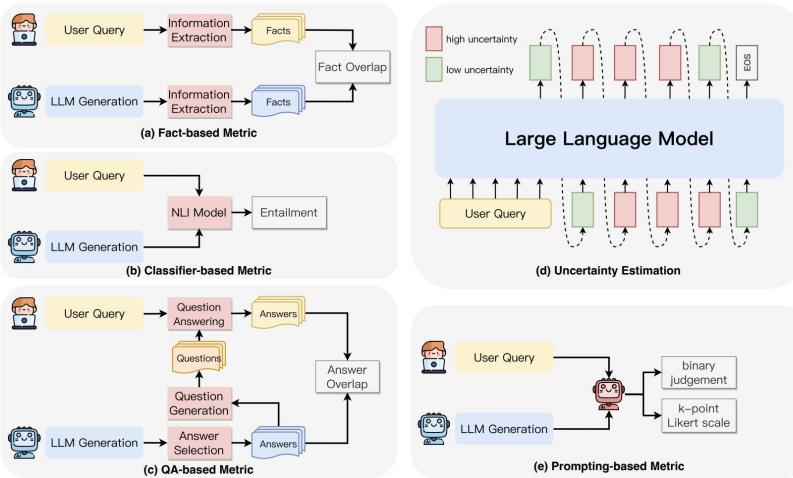
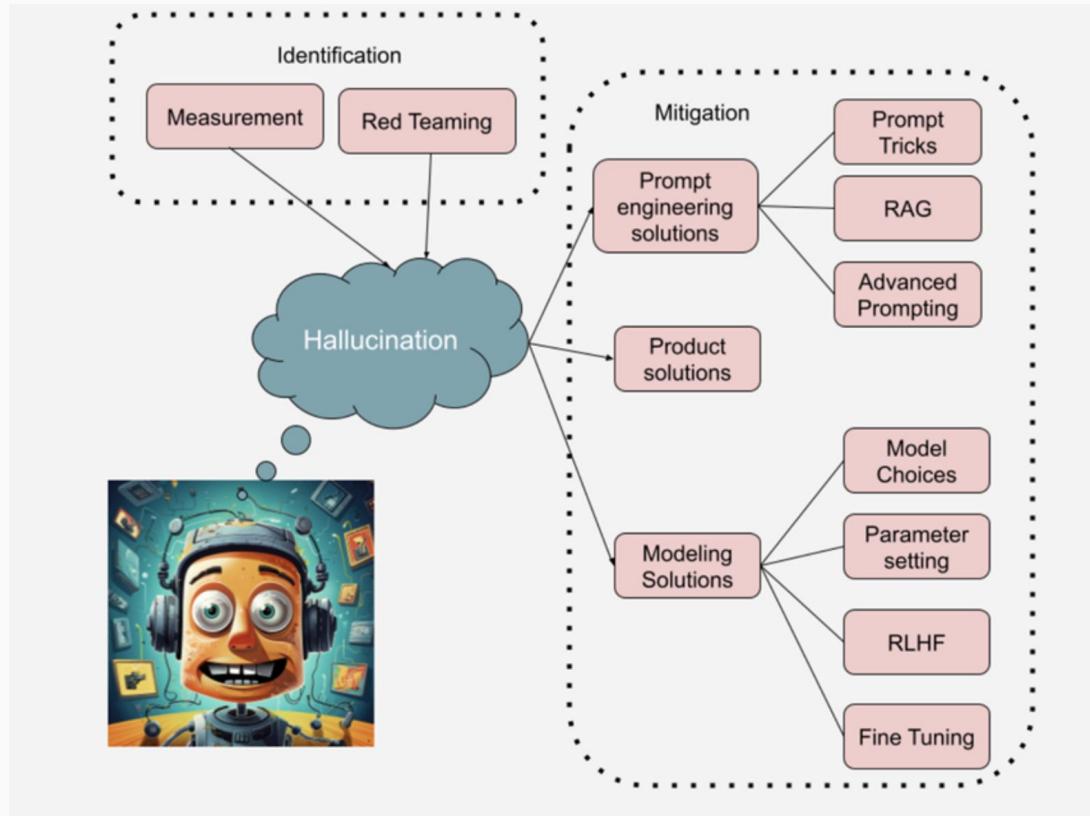


Figure 5: The illustration of detection methods for faithfulness hallucinations: **a) Fact-based Metrics**, which assesses faithfulness by measuring the overlap of facts between the generated content and the source content; **b) Classifier-based Metrics**, utilizing trained classifiers to distinguish the level of entailment between the generated content and the source content; **c) QA-based Metrics**, employing question-answering systems to validate the consistency of information between the source content and the generated content; **d) Uncertainty Estimation**, which assesses faithfulness by measuring the model’s confidence in its generated outputs; **e) Prompting-based Metrics**, wherein LLMs are induced to serve as evaluators, assessing the faithfulness of generated content through specific prompting strategies.

# Identify



# Methods

- **Fact Verification:** Cross-reference generated information with trusted sources to ensure accuracy and identify contradictions.
- **Contextual Analysis:** Assess if the output aligns with the query and conversation history, as hallucinations may diverge from the context.
- **Adversarial Testing:** Craft challenging prompts to compare model output with human-curated responses and identify hallucination patterns.
- **Consistency Check:** Use automated tools to detect logical inconsistencies or contradictions within the generated text.
- **Chain of Thought Prompting:** Ask the LLM to explain its step-by-step reasoning to trace contradictory logic or factual gaps indicating hallucination risks.

## SELFCHECKGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models

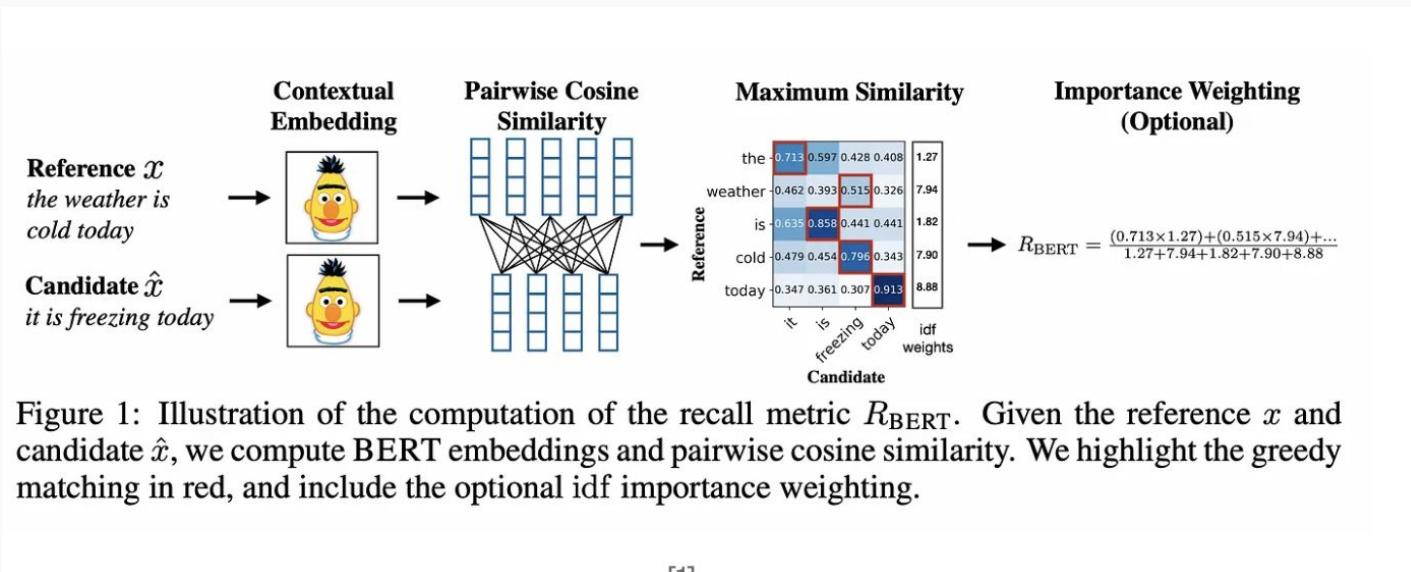
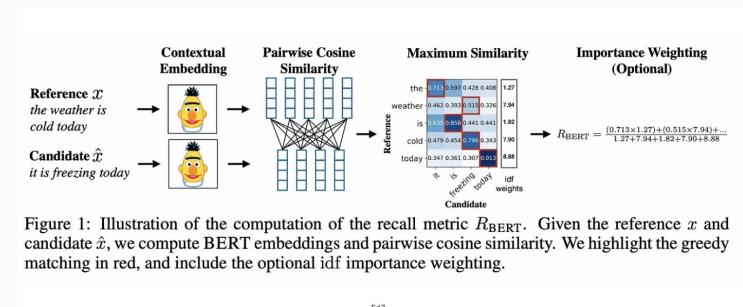


Figure 1: Illustration of the computation of the recall metric  $R_{BERT}$ . Given the reference  $x$  and candidate  $\hat{x}$ , we compute BERT embeddings and pairwise cosine similarity. We highlight the greedy matching in red, and include the optional idf importance weighting.

# Research methods

- **BERTScore**: Measures similarity between response and samples
- **Question Answering**: Checks if samples can answer questions from response
- **N-gram**: Measures probability of response under sample-trained model
- **NLI**: Measures contradiction between response and samples
- **Prompting**: Asks LLM if samples support response sentences

Key idea: Factual info is consistent across samples, hallucinations are not.



## Med-HALT: Medical Domain Hallucination Test for Large Language Models

Ankit Pal, Logesh Kumar Umapathi, Malaikannan Sankarasubbu

Saama AI Research, Chennai, India

{ankit.pal, logesh.umapathi, malaikannan.sankarasubbu}@saama.com

### Abstract

This research paper focuses on the challenges posed by hallucinations in large language models (LLMs), particularly in the context of the medical domain. Hallucination, wherein these models generate plausible yet unverified or incorrect information, can have serious consequences in healthcare applications. We propose a new benchmark and dataset, Med-HALT (Medical Domain Hallucination Test), designed specifically to evaluate and reduce hallucinations. Med-HALT provides a diverse multinational dataset derived from medical examinations across various countries and includes multiple innovative testing modalities. Med-HALT includes two categories of tests reasoning and memory-based hallucination tests, designed to assess LLMs' problem-solving and information retrieval abilities.

### Medical Hallucination LLM Benchmark

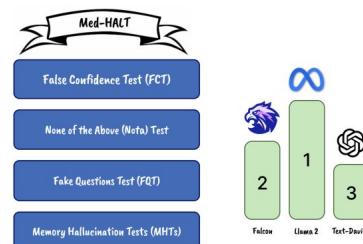


Figure 1: Med-HALT: A new benchmark dataset for LLM to test Hallucination in Medical Domain

ties, they are also subject to unique challenges

# LLM Hallucination in Medical Domain

- Medical misinformation can have **serious health consequences**
- Accuracy and reliability of language model information is critical, potentially affecting **life or death situations**
- Language models can impact healthcare **decisions, diagnoses, and treatments.**

**Question**

A 24-year-old pregnant woman at 28 weeks gestation presents to the emergency department with complaints of fever with chills and pain in her knee and ankle joints for the past 2 days. She also complains of headaches and difficulty moving her neck. Further questioning reveals that she had a tick bite on her arm while gardening a few days ago. Pregnancy is progressing normally. On examination, an erythematous rash is seen on her right arm. Her obstetric examination is normal. Ultrasound of the fetus is reassuring with a normal heartbeat and no gross abnormalities. A specimen is collected to test for Lyme disease. What is the next best step for this patient?

**Options**

A) Ibuprofen      B) Tetracycline      C) Amoxicillin      D) Gentamicin

**Physician**

Amoxicillin ✓

**Explanation:** Amoxicillin is a safe and effective treatment for Lyme disease in pregnancy. Ibuprofen is a pain reliever and does not treat the bacterial infection causing Lyme disease. Tetracycline antibiotic is not recommended during pregnancy; it can cross the placenta and affect fetal development, including causing discoloration of the baby's teeth and potential bone growth issues. Gentamicin, a potent antibiotic generally reserved for severe infections, is not typically used for treating Lyme disease.

**GPT-3.5**

Tetracycline ✗

**Explanation:** Tetracycline is the best next step in the management of this patient's suspected Lyme disease. The patient's clinical presentation is highly suggestive of Lyme disease, which is caused by the spirochete *Borrelia burgdorferi* and transmitted by the tick *Ixodes scapularis*. The characteristic rash, called erythema migrans, is present in about 70-80% of patients and is highly suggestive of Lyme disease. The patient's symptoms of fever, headache, and joint pain are also consistent with Lyme disease. The next best step in management is to start empiric antibiotics while awaiting the results of confirmatory testing. Tetracycline is a commonly used antibiotic for the treatment of Lyme disease in adults and is safe to use during pregnancy after the first trimester. It has been shown to be effective in eradicating *B. burgdorferi* and preventing the progression of Lyme disease. Ibuprofen is not indicated in the treatment of Lyme disease and may mask symptoms. Ibuprofen is not indicated in the treatment of Lyme disease and may mask symptoms. Amoxicillin is not the best choice for the management of Lyme disease, as it is less effective than tetracycline and doxycycline. Gentamicin is not an effective treatment for Lyme disease.

## Types of Hallucination Evaluated

- Memory Hallucination Tests (MHTs) 
- Reasoning Hallucination Tests (RHTs) 

## Memory Hallucination Test (MHT)

### PMID-to-Title Test

- LLM matches **PubMed ID** to **article title**
- Test gauges **model's mapping accuracy**
- Assesses correct **factual content retrieval**

```
prompt:  
    instruct: <instructions_to_llm>  
    pmid: <pmid_of_article>  
response:  
    is_paper_exists: <yes/no>  
    paper_title: <title_of_the_article>
```

## Memory Hallucination Test (MHT)

### Title-to-Link Test

- LLM links **article title** to PubMed URL
- Assesses **correct factual content retrieval**

prompt:

```
instruct: <instructions_to_llm>
title: <title_of_article>
```

response:

```
is_paper_exists: <yes/no>
paper_url: <url_of_the_article>
```

## Reasoning Hallucination Test (RHT)

- False Confidence Test (FCT)
- None of the Above (NOTA) Test
- Fake Questions Test (FQT)

## Reasoning Hallucination Tests (RHTs)

### False Confidence Test (FCT)

- FCT presents **random answers** for validation.
- Models must **validate and explain choices**.
- Test measures **overconfidence** with limited data.

```
prompt:  
    instruct: <instructions_to_llm>  
    question: <medical_question>  
    options:  
        - 0: <option_0>  
        - 1: <option_1>  
        - 2: <option_2>  
        - 3: <option_3>  
    correct_answer:  
        <randomly_suggested_correct_answer>  
response:  
    is_answer_correct: <yes/no>  
    answer: <correct_answer>  
    why_correct:  
        <explanation_for_correct_answer>  
    why_others_incorrect:  
        <explanation_for_incorrect_answers>
```

## Reasoning Hallucination Tests (RHTs)

### None of the Above (NOTA) Test

- NOTA Test **replaces correct answer with 'None'**.
- Model identifies this and **justifies selection**.
- Tests model's **ability to spot irrelevance**.

```
prompt:  
instruct: <instructions_to_llm>  
question: <medical_question>  
options:  
- 0: <option_0>  
- 1: <option_1>  
- 2: <option_2>  
- 3: <none_of_the_above>  
response:  
cop: <correct_option>  
cop_index: <correct_index_of_correct_option>  
why_correct:  
    <explanation_for_correct_answer>  
why_others_incorrect:  
    <explanation_for_incorrect_answers>
```

## Reasoning Hallucination Tests (RHTs)

### Fake Questions Test (FQT)

- Model presented with **fake medical questions**
- Test evaluates **handling of nonsensical queries**
- Fake questions **crafted by experts and GPT-3.5**

```
prompt:  
instruct: <instructions_to_llm>  
question: <fake_medical_question>  
options:  
- 0: <option_0>  
- 1: <option_1>  
- 2: <option_2>  
- 3: <option_3>  
response:  
cop: <correct_option>  
cop_index: <correct_index_of_correct_option>  
why_correct:  
    <explanation_for_correct_answer>  
why_others_incorrect:  
    <explanation_for_incorrect_answers>
```

# Experiments

## Baseline Models



Text-Davinci



GPT-3.5  
Turbo



Falcon



Llama-2



MPT



Commercial Models



Open-Source Models

# Results

Model	Reasoning FCT		Reasoning Fake		Reasoning Nota		Avg	
	Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score
GPT-3.5	34.15	33.37	71.64	11.99	27.64	18.01	44.48	21.12
Text-Davinci	16.76	-7.64	82.72	14.57	63.89	103.51	54.46	36.81
Llama-2 70B	<b>42.21</b>	<b>52.37</b>	97.26	17.94	<b>77.53</b>	<b>188.66</b>	<b>72.33</b>	<b>86.32</b>
Llama-2 70B Chat	13.34	-15.70	5.49	-3.37	14.96	-11.88	11.26	-10.32
Falcon 40B	18.66	-3.17	<b>99.89</b>	<b>18.56</b>	58.72	91.31	59.09	35.57
Falcon 40B-instruct	1.11	-44.55	99.35	18.43	55.69	84.17	52.05	19.35
Llama-2 13B	1.72	-43.1	89.45	16.13	74.38	128.25	55.18	33.76
Llama-2-13B-chat	7.95	-28.42	21.48	0.34	33.43	31.67	20.95	1.20
Llama-2-7B	0.45	-46.12	58.72	8.99	69.49	116.71	42.89	26.53
Llama-2-7B-chat	0.42	-46.17	21.96	0.46	31.10	26.19	17.83	-6.51
Mpt 7B	0.85	-45.15	48.49	6.62	19.88	-0.28	23.07	-12.94
Mpt 7B instruct	0.17	-46.76	22.55	0.59	24.34	10.34	15.69	-11.94

Table 2: Evaluation results of LLM's on Reasoning Hallucination Tests

Model	IR Pmid2Title		IR Title2Pubmedlink		IR Abstract2Pubmedlink		IR Pubmedlink2Title		Avg	
	Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score
GPT-3.5	0.29	-12.12	39.10	11.74	40.45	12.57	0.02	-12.28	19.96	-0.02
Text-Davinci	0.02	-12.28	38.53	11.39	40.44	12.56	0.00	-12.29	19.75	-0.15
Llama-2 70B	0.12	-12.22	14.79	-3.20	17.21	-1.72	0.02	-12.28	8.04	-7.36
Llama-2 70B Chat	0.81	-11.79	32.87	7.90	17.90	-1.29	0.61	-11.92	13.05	-4.27
Falcon 40B	<b>40.46</b>	<b>12.57</b>	<b>40.46</b>	<b>12.57</b>	<b>40.46</b>	<b>12.57</b>	0.06	-12.25	<b>30.36</b>	<b>6.37</b>
Falcon 40B-instruct	40.46	12.57	40.46	12.57	40.44	12.56	0.08	-12.75	30.36	6.24
Llama-2 13B	0.53	-11.97	10.56	-5.80	4.70	-9.40	<b>23.72</b>	<b>2.29</b>	9.88	-6.22
Llama-2-13B-chat	1.38	-11.44	38.85	11.59	38.32	11.26	1.73	-11.23	20.07	0.04
Llama-2-7B	0.00	-12.29	3.72	-10.00	0.26	-12.13	0.00	-12.29	1.0	-11.68
Llama-2-7B-chat	0.00	-12.29	30.92	6.71	12.80	-4.43	0.00	-12.29	10.93	-5.57
Mpt 7B	20.08	0.05	40.46	12.57	40.03	12.31	0.00	-12.29	25.14	3.16
Mpt 7B instruct	0.04	-12.27	38.24	11.21	40.46	12.57	0.00	-12.29	19.69	-0.19

Table 3: Evaluation results of LLM's on Memory Hallucination Tests

## Results

### TL;DR

- **LlaMa-2 70B:** Best in Reasoning FCT task, moderate accuracy (**42.21%**).   1 2  
3 4
- **Falcon 40B:** Outstanding in Reasoning Fake task, highest accuracy (**99.89%**).   
- **Llama-2 70B:** Top performer in Reasoning Nota task, good accuracy (**77.53%**).    1 2  
3 4
- **Falcon models** (40B and 40B Instruct): Excel in Information Retrieval tasks, leading in both accuracy and pointwise scores.   
- Overall, all models show a **need for substantial improvement**  

# Fine-grained Hallucination Detection and Editing For Language Models

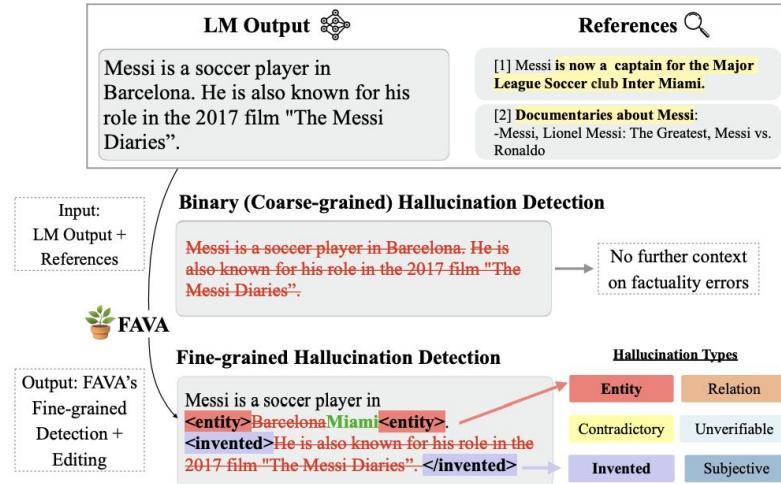


Figure 1: Overview of taxonomy for LM hallucinations and our model, FAVA, a fine-grained detection system.

# Migrating Hallucinations

# Migrating Hallucinations

- Mitigating Data-related Hallucinations 
- Mitigating Training-related Hallucinations 
- Mitigating Inference-related Hallucinations



# Migrating Hallucinations

- Mitigating Data-related Hallucinations 
- Mitigating Training-related Hallucinations 
- Mitigating Inference-related Hallucinations



# Mitigating Data-related Hallucinations



## Mitigating Misinformation and Biases

- Factuality Data Enhancement
- Debias

# Mitigating Data-related Hallucinations



## Mitigating Misinformation and Biases

- Factuality Data Enhancement
- Debias

## Mitigating Knowledge Boundary

- Model Editing
- Retrieval Augmentation

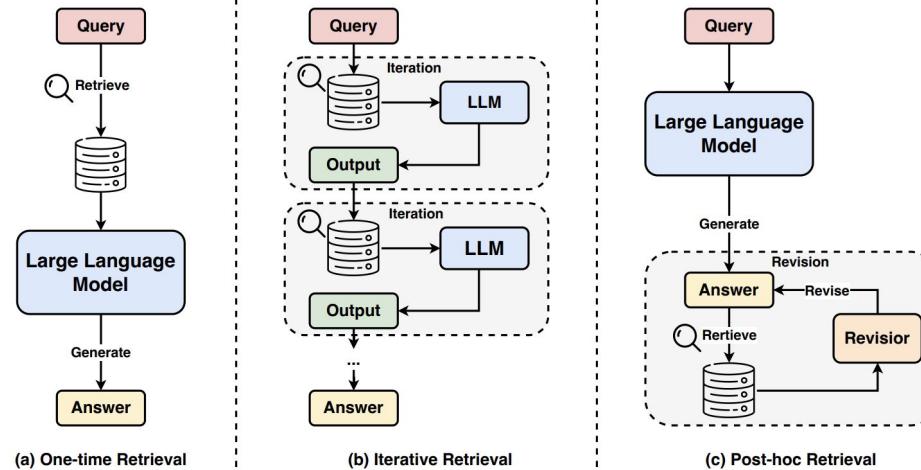


Figure 6: The illustration of three distinct approaches for Retrieval-Augmented Generation: **a) One-time Retrieval**, where relevant information is retrieved once before text generation; **b) Iterative Retrieval**, involving multiple retrieval iterations during text generation for dynamic information integration; and **c) Post-hoc Retrieval**, where the retrieval process happens after an answer is generated, aiming to refine and fact-check the generated content.

# Mitigating Data-related Hallucinations



## Mitigating Knowledge Shortcut

- Co-occurrence Debias

## Mitigating Knowledge Recall Failures

- Knowledge Clue Enhancement

# Mitigating Data-related Hallucinations



## Mitigating Knowledge Shortcut

- Co-occurrence Debias

# Migrating Hallucinations

- Mitigating Data-related Hallucinations 
- Mitigating Training-related Hallucinations 
- Mitigating Inference-related Hallucinations



# Migrating Hallucinations

- Architecture Improvement
- Pre-training Objective Improvement
- Mitigating Alignmentrelated Hallucination

# Migrating Hallucinations

- Mitigating Data-related Hallucinations 
- Mitigating Training-related Hallucinations 
- Mitigating Inference-related Hallucinations



# Migrating Hallucinations

## Factuality Enhanced Decoding

- On Standalone Decoding
- Post-editing Decoding

# Migrating Hallucinations

## Factuality Enhanced Decoding

- On Standalone Decoding
- Post-editing Decoding

## Faithfulness Enhanced Decoding

- Context Consistency
- Logical Consistency

# Migration Methods for Business Applications

- **Provide pre-defined input templates**
- **Adopt OpenAI's Reinforcement Learning with Human Feedback (RLHF)**
- **Fine-tune the model for specific industries**
- **Use process and outcome supervision**

# Hallucination Leaderboard

Select columns to show

Faithfulness  Factuality  NQ Open/EM  TriviaQA/EM  TruthQA MC1/Acc

TruthQA MC2/Acc  TruthQA Gen/ROUGE  XSum/ROUGE  XSum/factKB  XSum/BERT-P

CNN-DM/ROUGE  CNN-DM/factKB  CNN-DM/BERT-P  RACE/Acc  SQuADv2/EM

MemoTrap/Acc  IFEval/Acc  FaithDial/Acc  HaluQA/Acc  HaluSumm/Acc

HaluDial/Acc  FEVER/Acc  TrueFalse/Acc  PopQA/EM  NQ-Swap/EM  Type

Architecture  Precision  Hub License  #Params (B)  Hub ❤️  Available on the hub

Model sha

base merges and moerges  ?

Precision

float32  float16  bfloat16  8bit  4bit  GPTQ  ?

Model sizes (in billions of parameters)

?  -1.5  ~3  ~7  ~13  ~35  ~60  ~70+

T	Model	Faithfulness	Factuality	NQ Open/EM	TriviaQA/EM	TruthQA MC1/Acc	TruthQA MC2/Acc	TruthQA Gen/ROUGE	XSum/ROUGE	XSum
?	<a href="#">DiscoResearch/mixtral-7b-Bexpert</a>	16.77	25.02	0	0	24.72	49.48	15.18	0.13	
●	<a href="#">EleutherAI/gpt-j-6b</a>	33.1	34.17	9.17	39.32	20.2	35.96	24.36	5	
●	<a href="#">EleutherAI/gpt-neo-1.3B</a>	33.91	28.36	3.99	14.16	23.13	39.61	28.4	9.52	
●	<a href="#">EleutherAI/gpt-neo-125m</a>	32.08	25.04	0.53	1.55	25.83	45.58	36.6	11.91	
●	<a href="#">EleutherAI/gpt-neo-2.7B</a>	34.28	29.91	5.71	20.11	23.87	39.86	34.27	2.01	
◆	<a href="#">HuggingFaceH4/zephyr-7b-alpha</a>	39.6	57.35	26.23	65.92	40.64	56.02	48.35	15.47	
◆	<a href="#">HuggingFaceH4/zephyr-7b-beta</a>	36.14	55.11	25.21	64.43	38.68	55.12	44.68	16.66	
◆	<a href="#">Nexusflow/NexusRaven-V2-13B</a>	38.8	42.72	14.21	42.29	28.4	44.39	42.96	17.93	
●	<a href="#">NousResearch/Yarn-Mistral-7b-120k</a>	39.56	54.92	29.34	70.03	27.42	42.25	34.03	10.74	
●	<a href="#">Open-Orca/Mistral-7B-OpenOrca</a>	41.98	56.37	25.62	65.84	35.25	52.26	44.8	21.24	
●	<a href="#">Open-Orca/OpenOrca-Platypus2-13B</a>	40.67	48.07	28.31	9.88	38.07	52.68	40.39	25.78	
●	<a href="#">ai_forever/mGPT</a>	26.27	27.06	2.74	10.61	23.26	39.62	30.97	9.78	
●	<a href="#">hendrycks/natural-language-in-qa-7b-70k</a>	45.00	55.40	28.25	65.05	31.50	47.34	40.76	20.66	

Image source: <https://huggingface.co/blog/leaderboard-hallucinations>

# Update about Hallucination Research papers

The screenshot shows the GitHub repository page for 'awesome-hallucination-detection'. The repository is public and has 1 branch and 0 tags. The main file listed is 'README.md'. The repository was last updated 13 hours ago by pmminervini, with 101 commits. The README page includes sections for 'awesome-hallucination-detection', 'Citing this repository' (with a BibTeX snippet), and 'Papers and Summaries' (listing 'HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination').

awesome-hallucination-detection · Public

main · 1 Branch · 0 Tags

pmminervini Merge pull request #19 from rayguan97/patch-1 · c3f480c · 13 hours ago · 101 Commits

figures · update · 3 months ago

.gitignore · Initial commit · 6 months ago

LICENSE · Initial commit · 6 months ago

README.md · Update README.md · 15 hours ago

README · Apache-2.0 license

**awesome-hallucination-detection**

awesome · License Apache 2.0

Citing this repository

```
@misc{MinerviniAHD2014,  
author = {Pasquale Minervini and others},  
title = {awesome-hallucination-detection},  
year = {2014},  
publisher = {GitHub},  
journal = {GitHub repository},  
howpublished = {\url{https://github.com/EdinburghNLP/awesome-hallucination-detection}}}
```

Papers and Summaries

HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination

Image source: <https://github.com/EdinburghNLP/awesome-hallucination-detection>

Can LLM hallucination be  
beneficial in any way?

 Andrej Karpathy ✅  
@karpathy

# On the "hallucination problem"

I always struggle a bit with I'm asked about the "hallucination problem" in LLMs. Because, in some sense, hallucination is all LLMs do. They are dream machines.

We direct their dreams with prompts. The prompts start the dream, and based on the LLM's hazy recollection of its training documents, most of the time the result goes someplace useful.

It's only when the dreams go into deemed factually incorrect territory that we label it a "hallucination". It looks like a bug, but it's just the LLM doing what it always does.

At the other end of the extreme consider a search engine. It takes the prompt and just returns one of the most similar "training documents" it has in its database, verbatim. You could say that this search engine has a "creativity problem" - it will never respond with something new. An LLM is 100% dreaming and has the hallucination problem. A search engine is 0% dreaming and has the creativity problem.

LLM hallucination is a  
double-edged sword

# Positive Applications of LLM Hallucination

- **Art and Design**
- **Data Visualization**
- **Gaming and Virtual Reality (VR)**

# Thanks.

Let's get connected on X / Twitter, I am  
**@aadityaura**

Any Questions?