saama

Hallucinations in Large Language Models

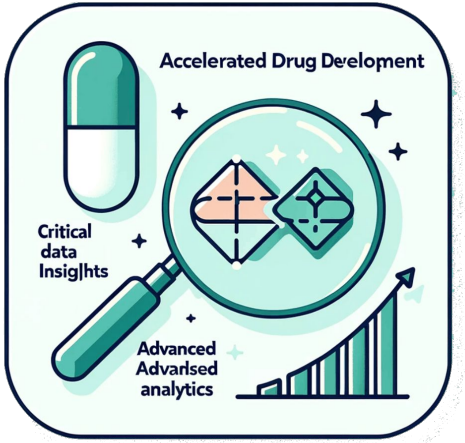
Causes, Types, and Practical Mitigation Techniques

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

Malaikannan Sankarasubbu
VP of AI, Saama AI Research Lab
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.

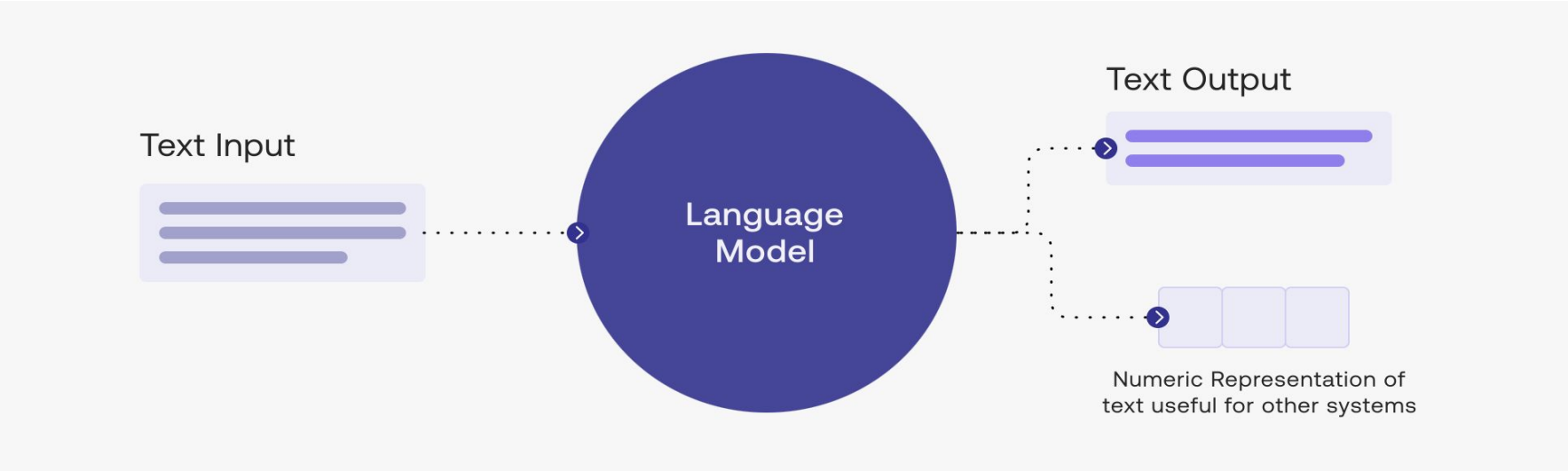


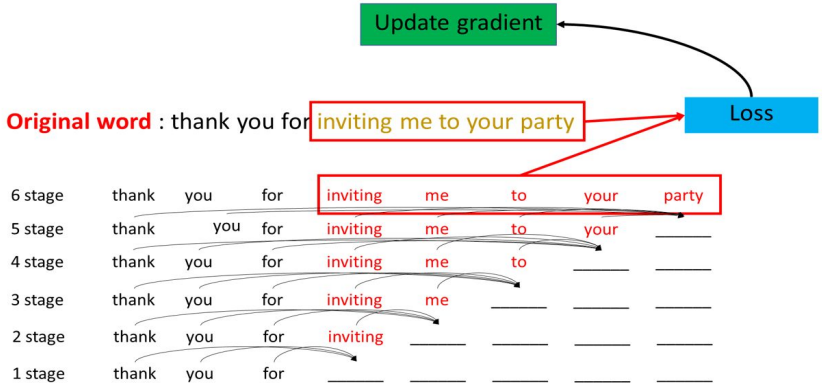
Image source: https://jalammar.github.io/applying-large-language-models-cohere/

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

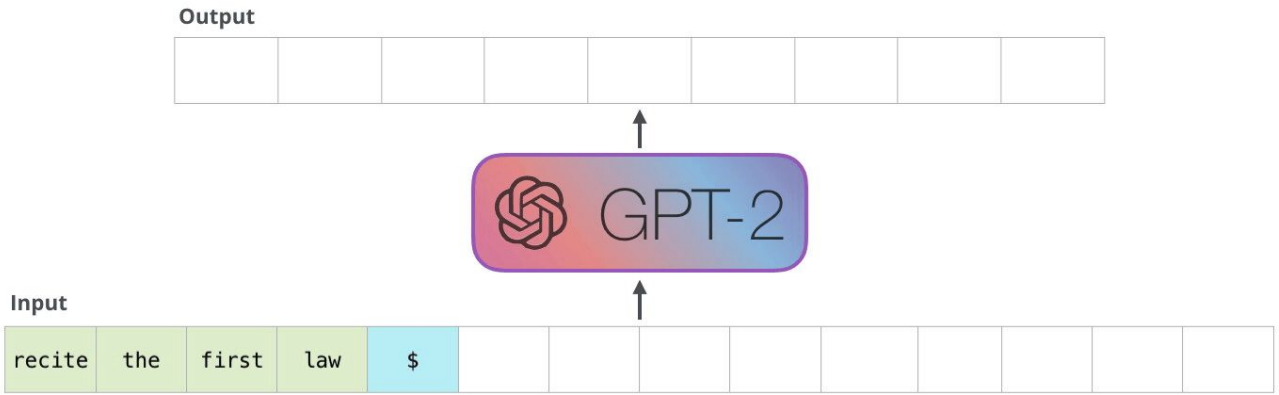
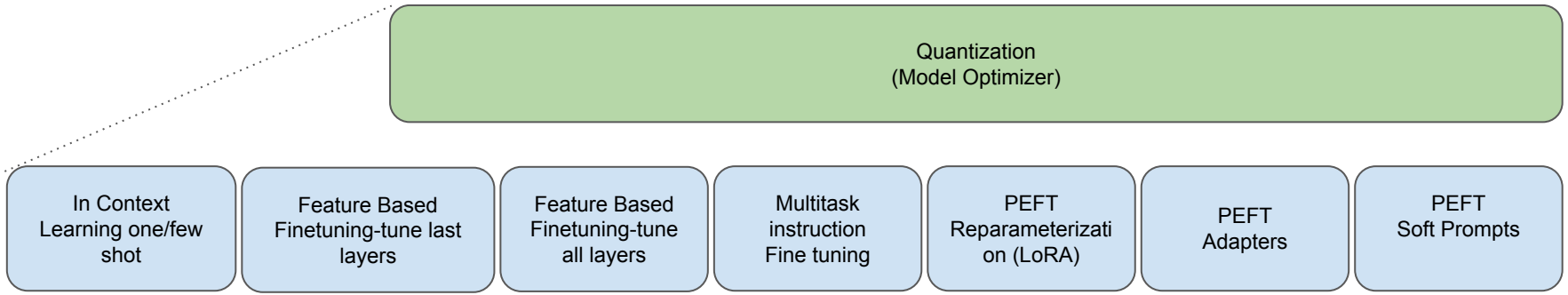
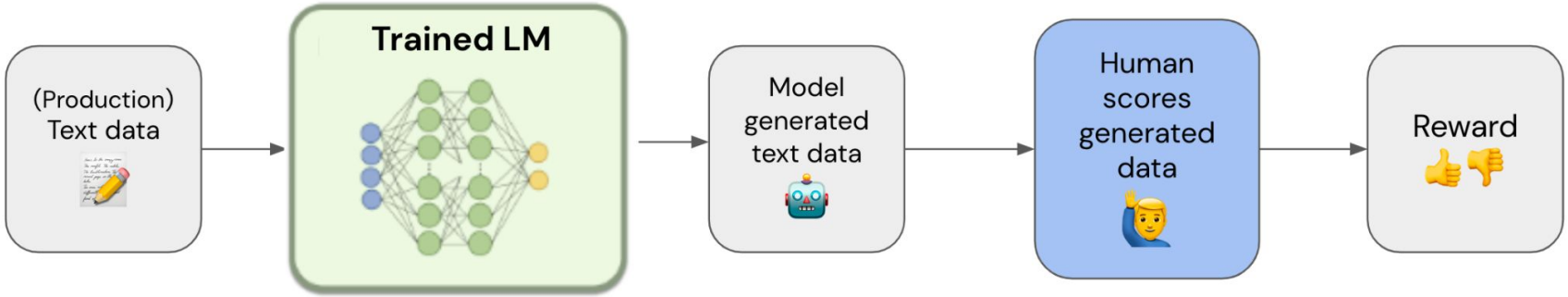


Image source: https://jalammar.github.io/illustrated-gpt2/





Chat-GPT Based Models - Instruct Tuned

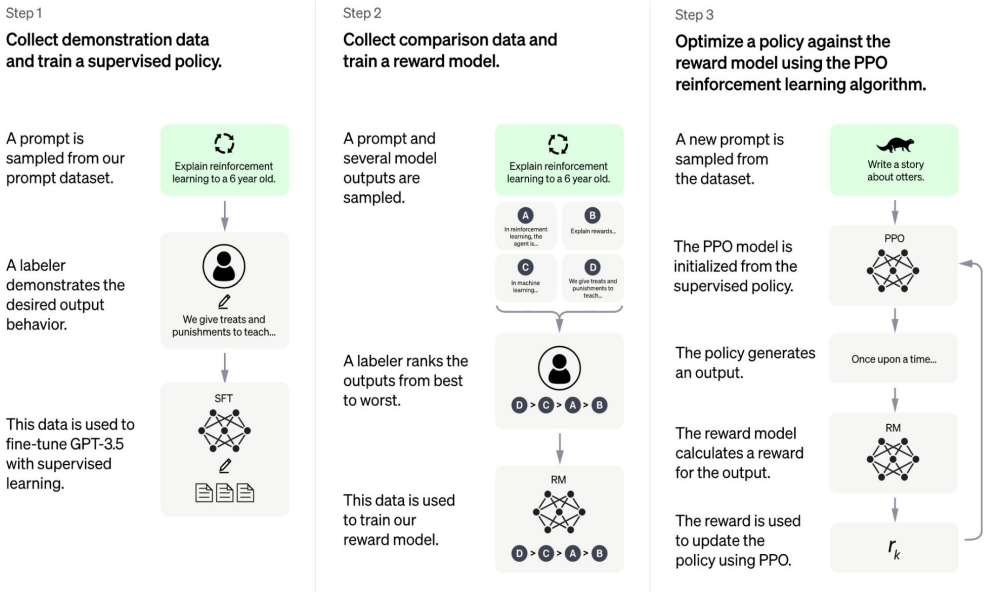


Image source: https://openai.com/research/instruction-following

Large Language Models

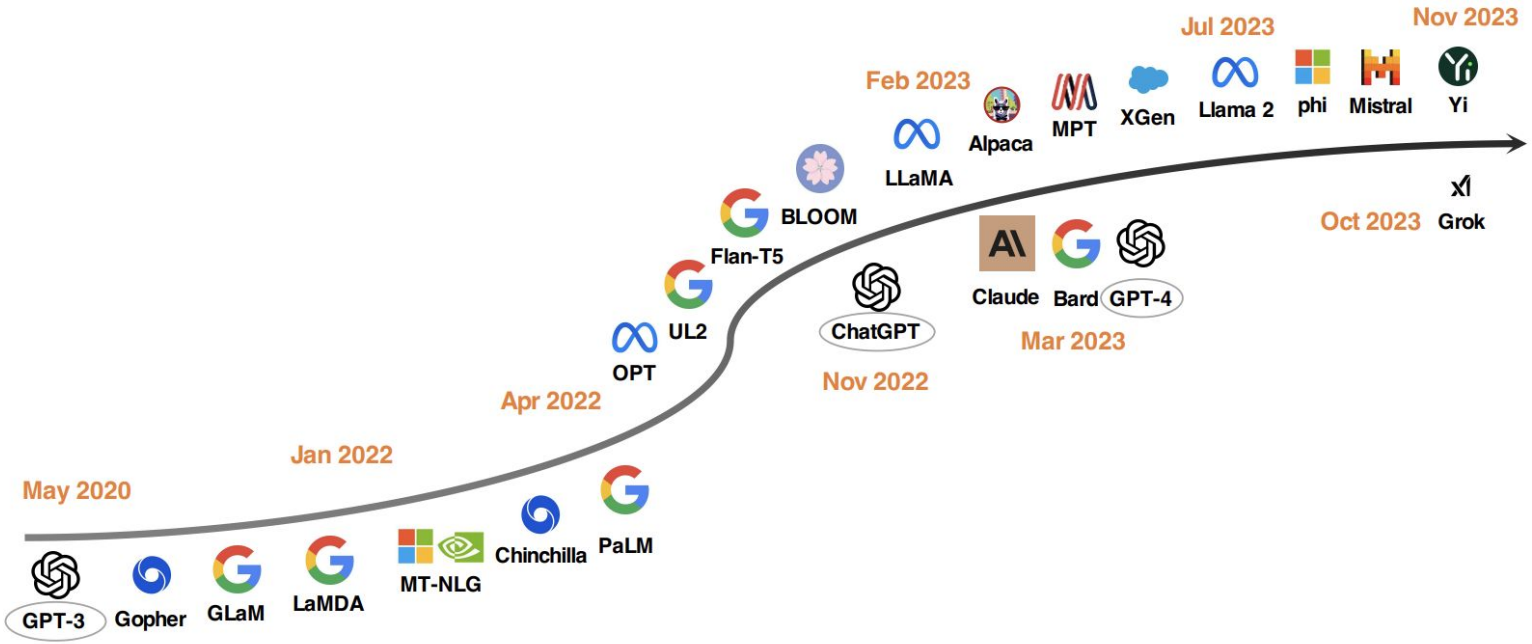


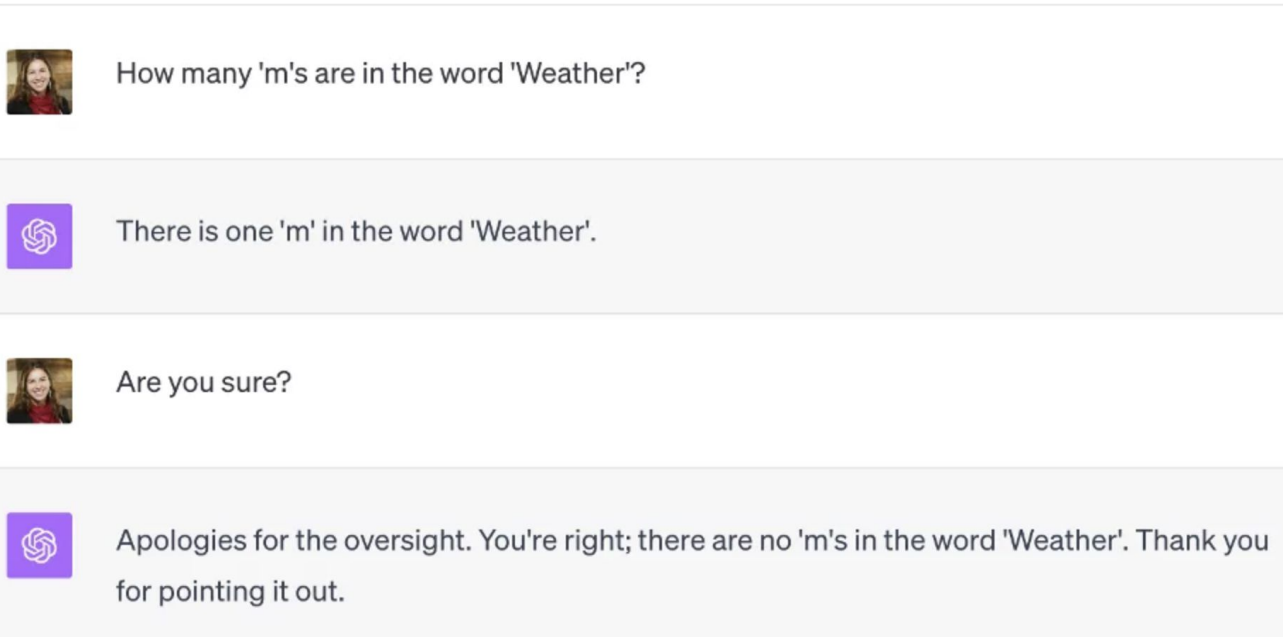
Image source: https://arxiv.org/pdf/2311.16989.pdf

LLM Hallucination =

Generates information which doesn't accurately reflect reality

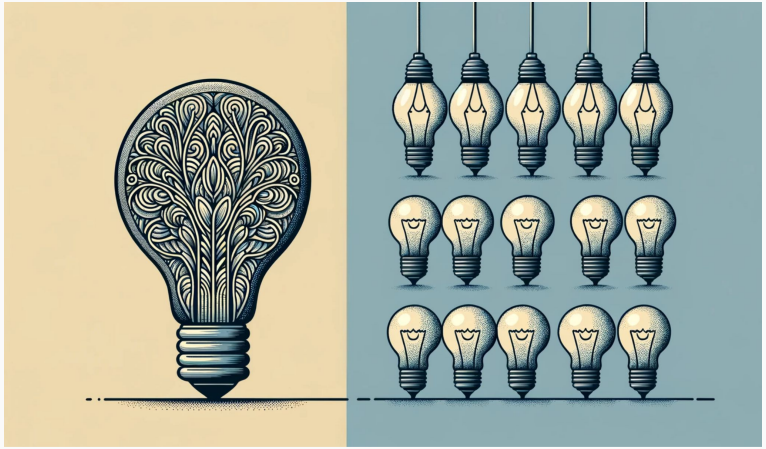
LLM Hallucination

Output that are grammatically correct but factually incorrect or nonsensical.



Why Hallucination happens?

LLMs



Reasoning vs Mimicry

Humans reason, We think through a scenario logically

LLMs Mimic. They return an answer based on a series of a 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

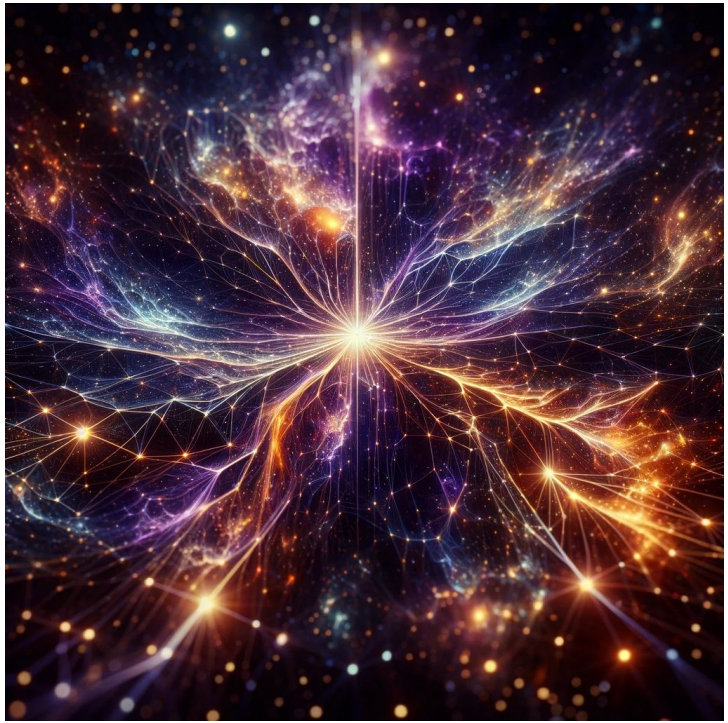


Image source: https://www.youtube.com/watch?app=desktop&v=JvLiEdTKKgk

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 Ouput	Explanation
Imitative Falsehood	Within the LLM's training data, a frequently referenced but incorrect fact appears multiple times: *Thomas Edison invented the light bulb.*	Who is credited with the invention of the light bulb?	Thomas Edison 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.

Image source: https://arxiv.org/pdf/2311.05232.pdf

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.

Image source: https://arxiv.org/pdf/2311.05232.pdf

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.

Image source: https://arxiv.org/pdf/2311.05232.pdf

Knowledge Boundary

Type	Input	Model Ouput	Explanation
Domain Knowledge Deficiency	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.
Outdated Knowledge	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.

Image source: https://arxiv.org/pdf/2311.05232.pdf

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?	Toronto is the capital of Canada.	The model leans heavily on the frequent co-occurrence of the terms *Toronto* and *Canada* in its training data, without truly capturing the factual knowledge about the capital of Canada

Image source: https://arxiv.org/pdf/2311.05232.pdf

Inferior Data Utilization

Knowledge Recall Failures	Long-tail Knowledge	Please generate a biography for George James Rankin.	George James Rankin is a dedicated educator known for his contributions to the field of education and his passion for fostering learning.	**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.
	Complex Reasoning	If Mount Everest were to descend by 500 meters, which mountain would become the world's highest peak?	If Mount Everest were to descend by 500 meters, it would still remain the world's highest peak.	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, K2 would become 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.

Image source: https://arxiv.org/pdf/2311.05232.pdf

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.

Image source: https://www.simform.com/blog/llm-hallucinations/

Types

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.

Image source: https://www.simform.com/blog/llm-hallucinations/

Types

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.

Image source: https://www.simform.com/blog/llm-hallucinations/

How to detect LLM hallucinations?

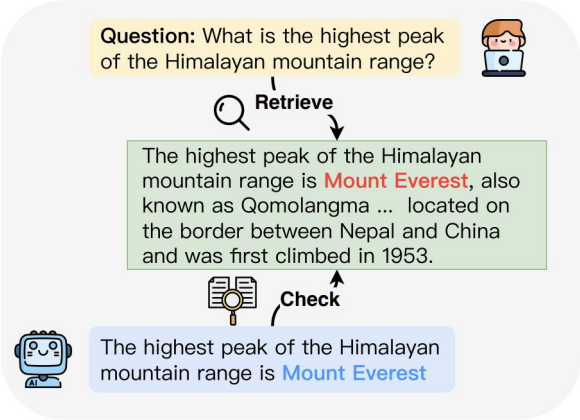
Hallucination Detection & Benchmarks

- Hallucination Detection
- Hallucination Benchmarks

Hallucination from Data

Hallucination Detection

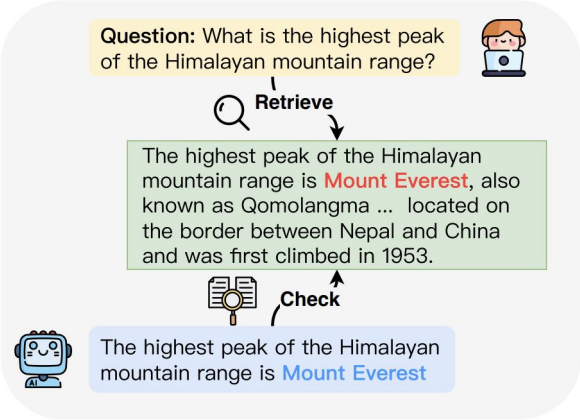
- Factuality Hallucination Detection
- Faithfulness Hallucination Detection



Hallucination from Data

Hallucination Detection

- Factuality Hallucination Detection
- Faithfulness Hallucination Detection



Hallucination Benchmarks

- Hallucination Evaluation Benchmarks
- Hallucination Detection Benchmarks

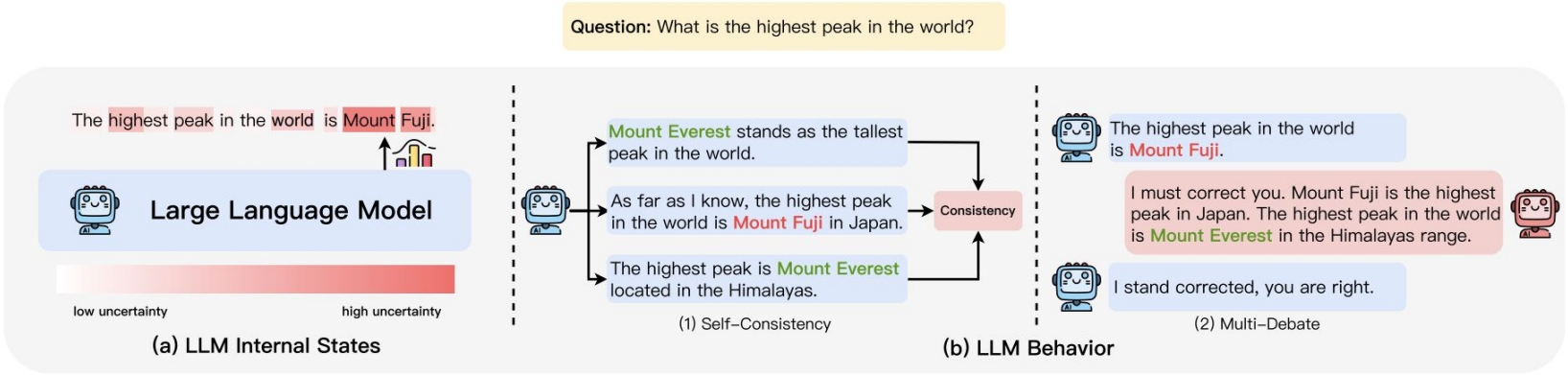


Figure 4: Taxonomy of Uncertainty Estimation Methods in Factual Hallucination Detection, featuring **a) LLM Internal States** and **b) LLM Behavior**, with LLM Behavior encompassing two main categories: Self-Consistency and Multi-Debate.

Image source: https://arxiv.org/pdf/2311.05232.pdf

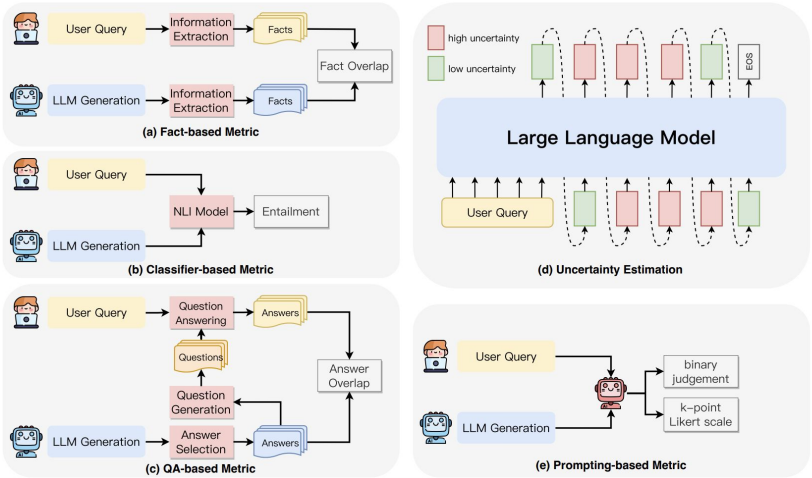


Figure 5: The illustration of detection methods for faithfulness hallucinations: **a) Fact-based Metrics**, which assesses faithfulness by measuring the overlap of facts between the generated content and the source content; **b) Classifier-based Metrics**, utilizing trained classifiers to distinguish 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.

Image source: https://arxiv.org/pdf/2311.05232.pdf

Identify

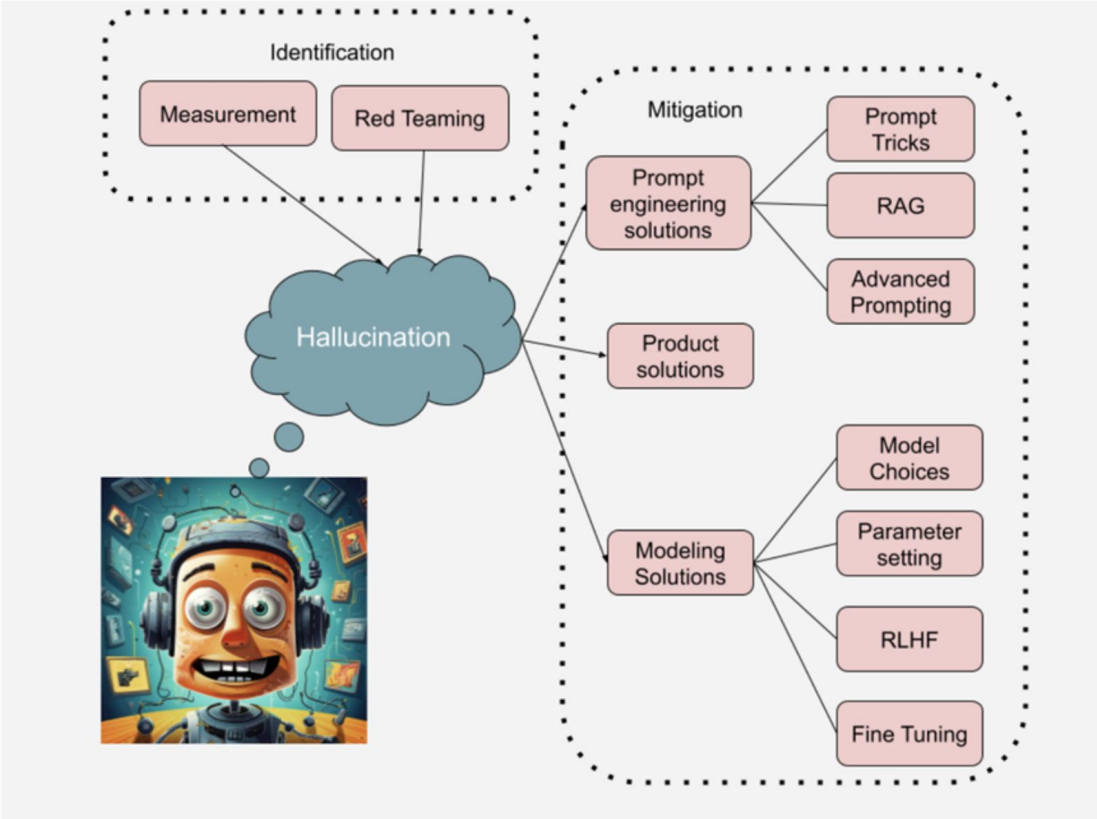


Image source: https://amatriain.net/blog/hallucinations

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.

Research methods

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

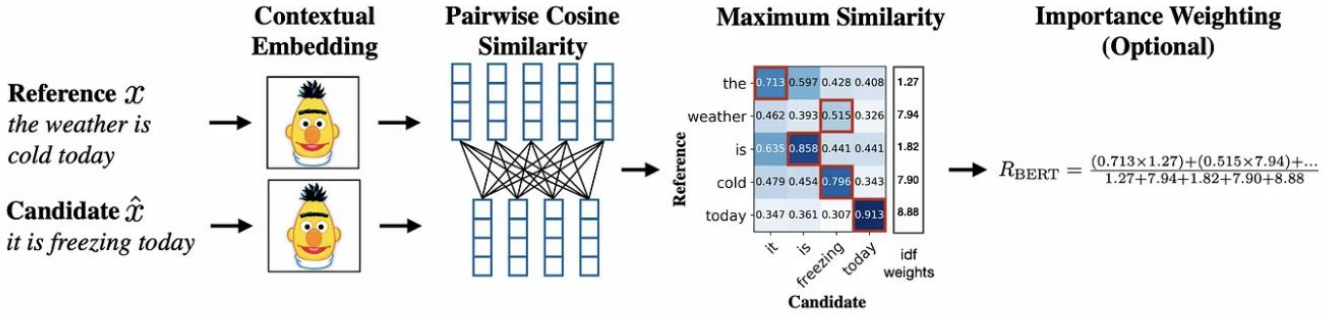


Figure 1: Illustration of the computation of the recall metric R_{BERT}. Given the reference x and candidate $\hat{x}$, we compute BERT embeddings and pairwise cosine similarity. We highlight the greedy matching in red, and include the optional idf importance weighting.

[1]

Image source: https://arxiv.org/abs/2303.08896

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.

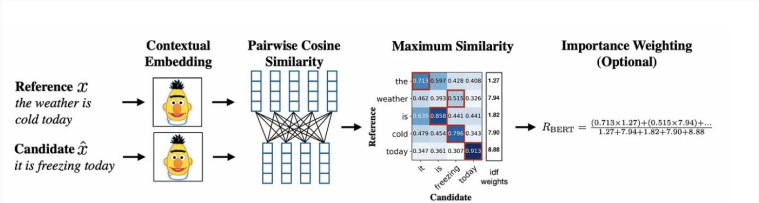


Figure 1: Illustration of the computation of the recall metric R_{BERT}. Given the reference x and candidate $\hat{x}$, we compute BERT embeddings and pairwise cosine similarity. We highlight the greedy matching in red, and include the optional idf importance weighting.

Med-HALT

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.

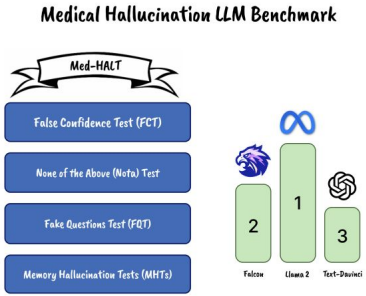
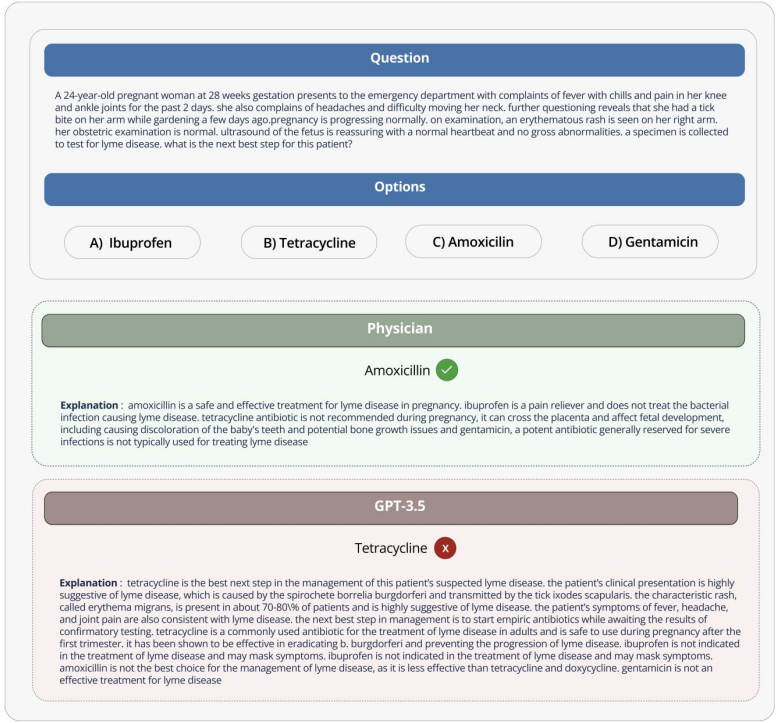


Figure 1: Med-HALT: A new benchmark dataset for LLM to test Hallucination in Medical Domain

ities, 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**.



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

	Reasoning FCT		Reasoning Fake		Reasoning Nota		Avg	
Model	**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	**42.21**	**52.37**	97.26	17.94	**77.53**	**188.66**	**72.33**	**86.32**
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	**99.89**	**18.56**	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

	IR Pmid2Title		IR Title2Pubmedlink		IR Abstract2Pubmedlink		IR Pubmedlink2Title		Avg	
Model	**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	**40.46**	**12.57**	**40.46**	**12.57**	**40.46**	**12.57**	0.06	-12.25	**30.36**	**6.37**
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	**23.72**	**2.29**	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%**). 🦙🤔🔢
- **Falcon 40B**: *Outstanding in Reasoning Fake task*, highest accuracy (**99.89%**). 🦅✅💯
- **Llama-2 70B**: *Top performer in Reasoning Nota task*, good accuracy (**77.53%**). 🦙❓👍🔢
- **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

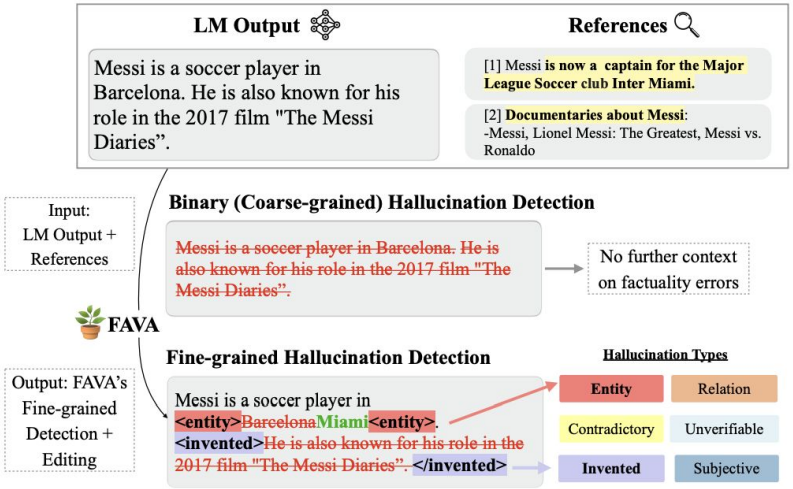


Figure 1: Overview of taxonomy for LM hallucinations and our model, FAVA, a fine-grained detection system.

Image source: https://arxiv.org/pdf/2401.06855.pdf

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

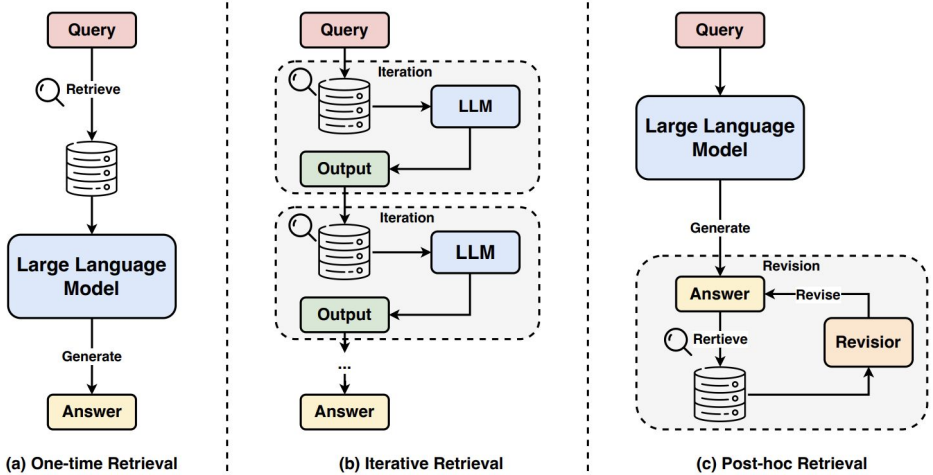


Figure 6: The illustration of three distinct approaches for Retrieval-Augmented Generation: **a) One-time Retrieval**, where relevant information is retrieved once before text generation; **b) Iterative Retrieval**, involving multiple retrieval iterations during text generation for dynamic information 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.

Image source: https://arxiv.org/pdf/2311.05232.pdf

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

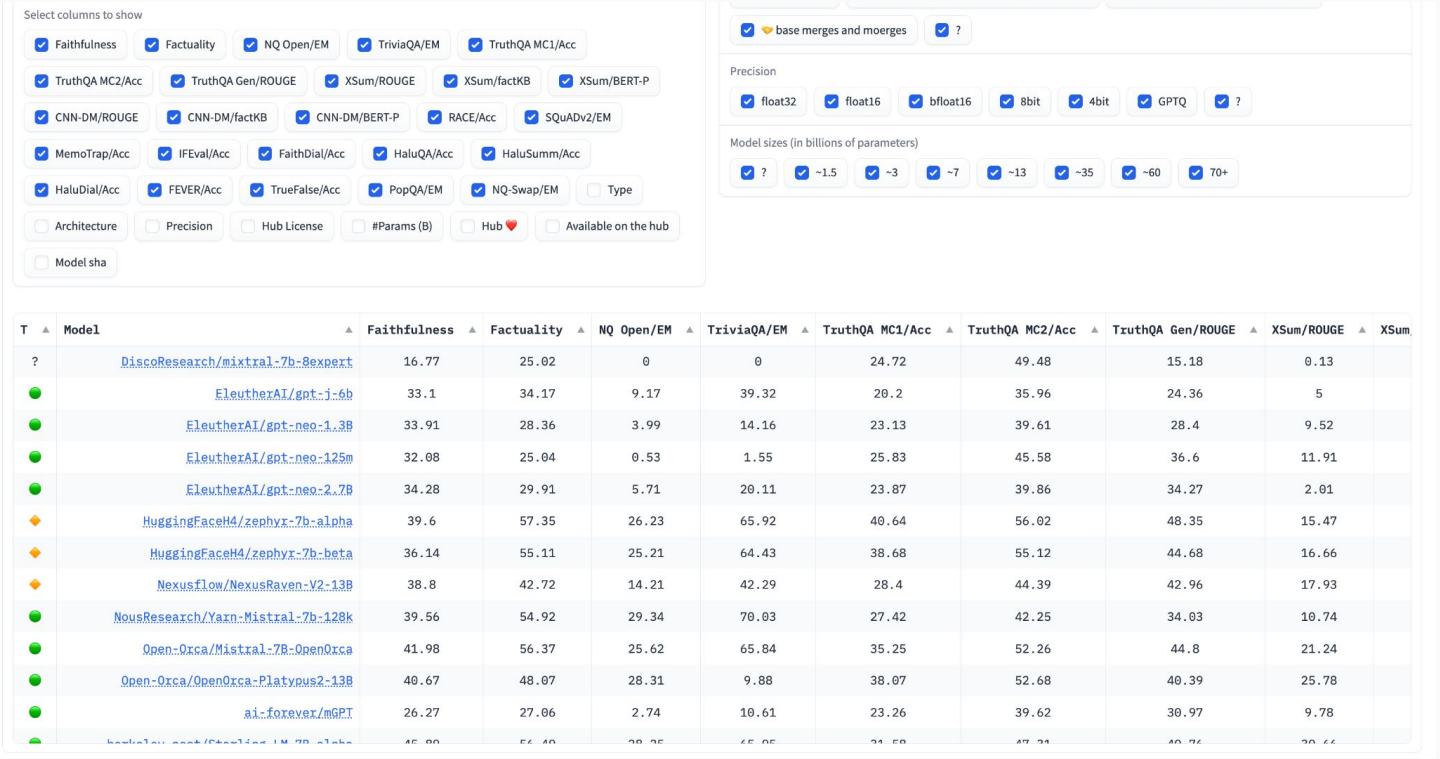
Faithfulness Enhanced Decoding

- 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



T	Model	Faithfulness	Factuality	NQ Open/EM	TriviaQA/EM	TruthQA MC1/Acc	TruthQA MC2/Acc	TruthQA Gen/ROUGE	XSum/ROUGE	XSum
?	DiscoResearch/mixtral-7b-8expert	16.77	25.02	0	0	24.72	49.48	15.18	0.13	
🟢	EleutherAI/gpt-j-6b	33.1	34.17	9.17	39.32	20.2	35.96	24.36	5	
🟢	EleutherAI/gpt-neo-1.3B	33.91	28.36	3.99	14.16	23.13	39.61	28.4	9.52	
🟢	EleutherAI/gpt-neo-125m	32.08	25.04	0.53	1.55	25.83	45.58	36.6	11.91	
🟢	EleutherAI/gpt-neo-2.7B	34.28	29.91	5.71	20.11	23.87	39.86	34.27	2.01	
🔶	HuggingFaceH4/zephyr-7b-alpha	39.6	57.35	26.23	65.92	40.64	56.02	48.35	15.47	
🔶	HuggingFaceH4/zephyr-7b-beta	36.14	55.11	25.21	64.43	38.68	55.12	44.68	16.66	
🔶	Nexusflow/NexusRaven-V2-13B	38.8	42.72	14.21	42.29	28.4	44.39	42.96	17.93	
🟢	NousResearch/Yarn-Mistral-7b-128k	39.56	54.92	29.34	70.03	27.42	42.25	34.03	10.74	
🟢	Open-Orca/Mistral-7B-OpenOrca	41.98	56.37	25.62	65.84	35.25	52.26	44.8	21.24	
🟢	Open-Orca/OpenOrca-Platypus2-13B	40.67	48.07	28.31	9.88	38.07	52.68	40.39	25.78	
🟢	ai-forever/mGPT	26.27	27.06	2.74	10.61	23.26	39.62	30.97	9.78	

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

Update about Hallucination Research papers

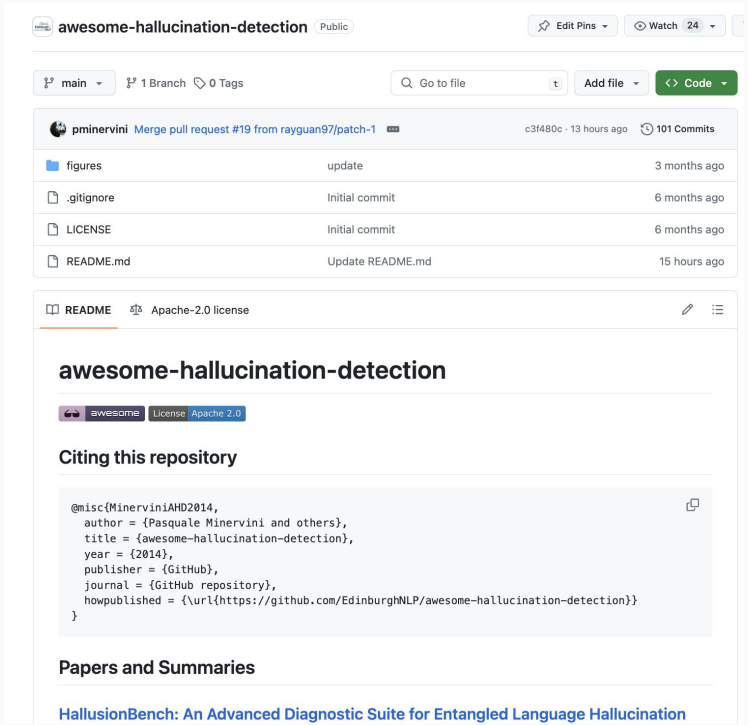


Image source: https://github.com/EdinburghNLP/awesome-hallucination-detection

Can LLM hallucination be beneficial in any way?

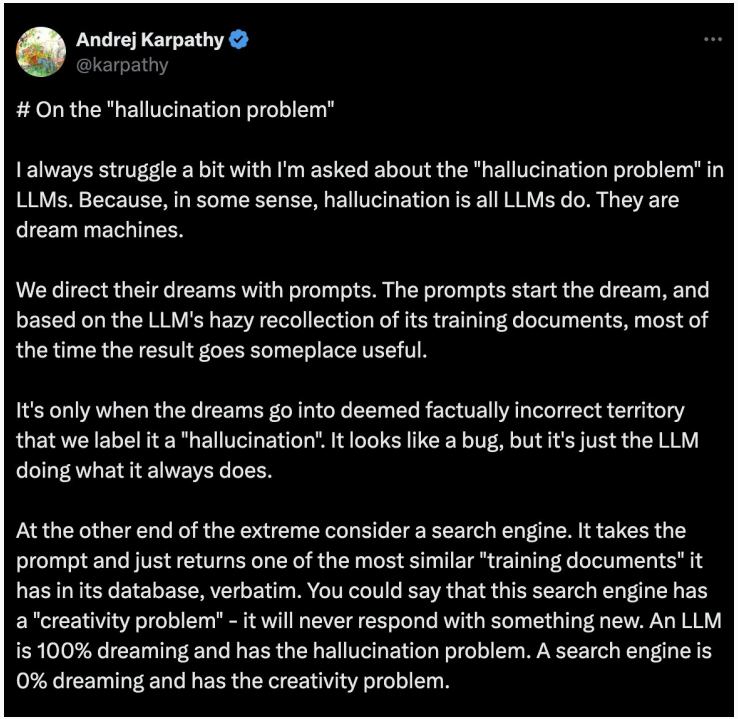


Image source: https://jalammar.github.io/applying-large-language-models-cohere/

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?