Ensuring AI Reliability: Addressing Hallucinations in Healthcare and Life Sciences

Aditya Gadiko
Phuse SDE @ SAS Inc
Cary, NC
November 21, 2024

Aditya Gadiko

Senior Director,
Clinical AI
Product Management

What is a hallucination?

A phenomenon where the model generates text that is incorrect, nonsensical, or not real.

Examples of large language model (LLM) hallucinations include…

- Answering confidently when asked for the weather forecast in a fictional city
- Providing fabricated references in an academic context

Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models

Types of hallucinations

1. User input

Can you recommend a delicious recipe for dinner?

2. LLM response

Yes, here is a delicious recipe for ~~lunch~~. So how about fried chicken with mashed potatoes? In addition, tomatoes are also an excellent pairing for this ~~this as they~~ are rich in **calcium**. Enjoy ~~this~~ **steak**!

Types of hallucinations

Input-conflicting hallucination	The user wants a recipe for **dinner** while LLM provide one for **lunch**
Context-conflicting hallucination	**Steak** has not been mentioned in the preceding context
Fact-conflicting hallucination	**Tomatoes** are not rich in **calcium** in fact

Source: Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models

Examples of hallucinations

Disrupted

New York lawyers sanctioned for using fake ChatGPT cases in legal brief

By **Sara Merken**

June 26, 2023 1:58 PM GMT+5:30 · Updated 4 months ago

A response by ChatGPT, an AI chatbot developed by OpenAI, is seen on its website in this illustration picture taken February 9, 2023. REUTERS/Florence Lo/Illustration Acquire Licensing Rights

NEW YORK, June 22 (Reuters) - A U.S. judge on Thursday imposed sanctions on two New York lawyers who submitted a legal brief that included six fictitious case citations generated by an artificial intelligence chatbot, ChatGPT.

U.S. District Judge P. Kevin Castel in Manhattan ordered lawyers Steven Schwartz, Peter LoDuca and their law firm Levidow, Levidow & Oberman to pay a $5,000 fine in total.

The judge found the lawyers acted in bad faith and made "acts of conscious avoidance and false and misleading statements to the court."

Examples of hallucinations

Type	Turn	Example
Input-conflicting Hallucination	*User*	Please summarize the following content: My friend Hill and I love playing basketball together. We spend hours on the court, dribbling, shooting, and competing against each other. Our passion for this game brings us closer as good friends.
	Model	Lucas and I love playing basketball and this makes us good friends.

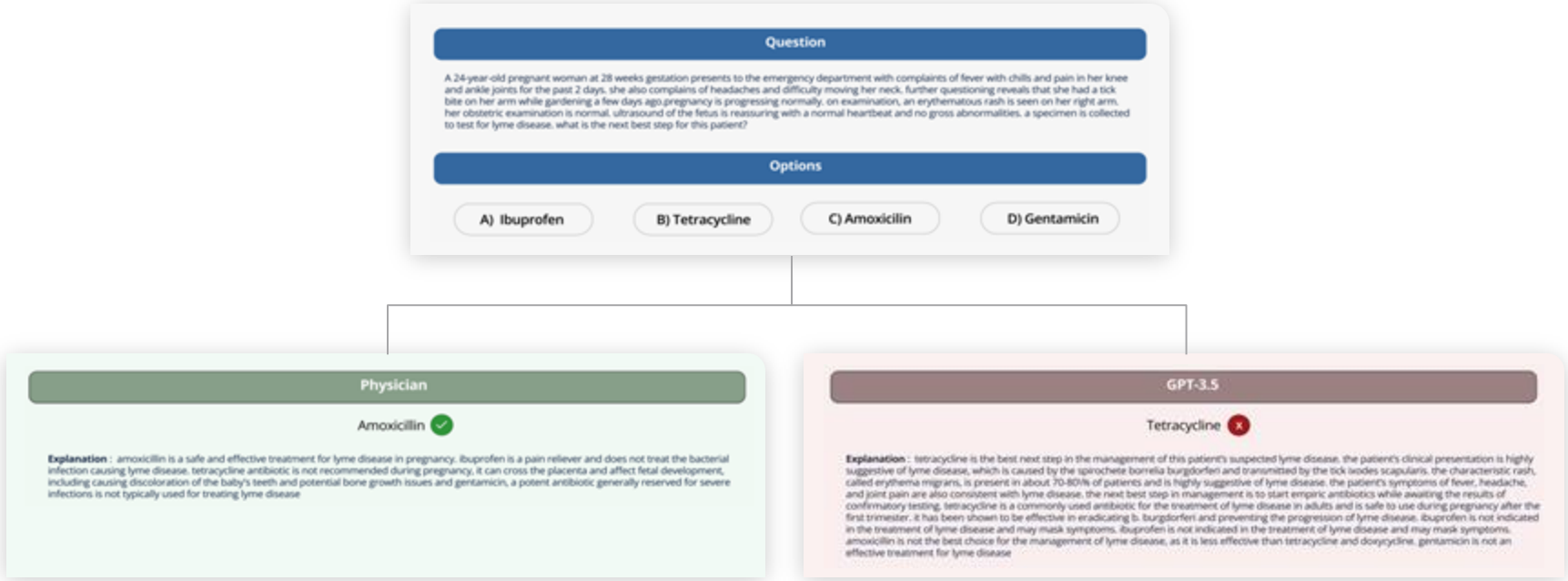
Examples of hallucinations

Context-conflicting Hallucination	*User*	Tell me something about the NBA Commissioner.
	Model	In the world's most popular basketball league, the NBA, Commissioner Adam Silver has long been regarded as a leader and decision-maker. Firstly, Silver has claimed his dedication to maintaining fairness and competitiveness in the game. Secondly, Silver has repeatedly emphasized his concern for player health and well-being. In this case, Stern's words and actions clash with his stated goal of caring for player welfare.

Examples of hallucinations

Fact-conflicting Hallucination	*User*	Who was the mother of Afonso II, the third king of Portugal?
	Model	The mother of Afonso II was Queen Urraca of Castile.

Examples of hallucinations



Measuring Hallucinations

Fact score

Give me a bio of Bridget Moynahan

ChatGPT

Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, I, Robot, and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and…

- ✓ Bridget Moynahan is American
- ✓ Bridget Moynahan is an actress
- ✓ Bridget Moynahan is a model
- ✗ Bridget Moynahan is a producer
- ✗ She is best known for her roles in Grey's Anatomy
- ✓ She is best known for her roles in I, Robot
- ✓ She is best known for her roles in Blue Bloods.
- ✓ She studied acting
- ✗ She studied at the American Academy of Dramatic Arts

66.7%

Stable LM

Bridget Moynahan is an American filmmaker and writer. She is best known for her work on the soap opera General Hospital, which she co-created with husband Charles Kelly. Moynahan was raised in a middle-class family in Los Angeles.

- ✓ Bridget Moynahan is American
- ✗ Bridget Moynahan is a filmmaker
- ✗ Bridget Moynahan is a writer
- ✗ She is best known for her work on General Hospital
- ✗ General Hospital is the soap opera
- ✗ She co-created General Hospital
- ✗ She co-created General Hospital with her husband
- ✗ Her husband is Charles Kelly
- ✗ Moynahan was raised in a middle-class family
- ✗ Moynahan was raised in Los Angeles

10.0%

FACTSCORE: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation

Med-HALT: medical domain hallucination test for large language models (LLMs)

The **Med-HALT** framework proposes a two-tiered approach to evaluate the presence and impact of hallucinations in biomedical generated outputs of LLMs:

1

Reasoning Hallucination Tests (RHTs)

False confidence test (FCT)
Involves presenting a multiple-choice medical question and a randomly suggested correct answer to the LLM, tasking it with evaluating the validity of the proposed answer, and providing detailed explanations for its correctness or incorrectness, in addition to explaining why the other options are wrong.

None of the above (Nota) test
The model is presented with a multiple-choice medical question correct answer is replaced by 'None of the above', requiring the model to identify this and justify its selection.

Fake question test
This test involves presenting the model with fake or nonsensical medical questions to examine whether it can correctly identify and handle such queries.

2

Memory hallucination tests

Tests the memory of LLMs to retrieve factual information from a biomedical corpus like Pubmed

Med-HALT: medical domain hallucination test for large language models (LLMs)

	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

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

Controlling Hallucinations

Retrieval augmented generation (RAG)

Some of the reasons for hallucinations:

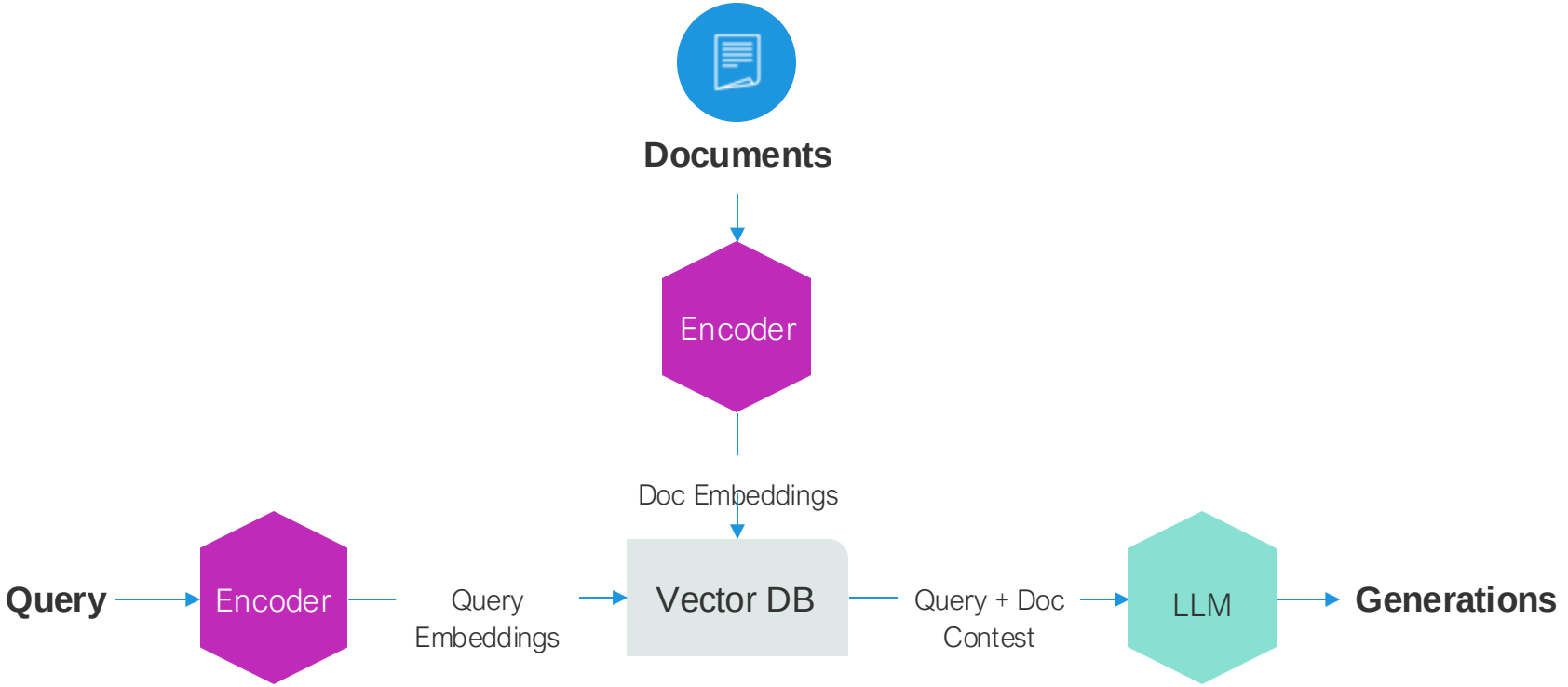
1

LLMs are static
LLMs are “frozen in time” and lack up-to-date information.

2

LLMs lack domain-specific knowledge
LLMs are trained for generalized tasks, meaning they do not know a company’s private data.

Retrieval augmented generation (RAG)



Retrieval augmented generation (RAG)

These features of RAG help reduce the hallucination:

1

RAG makes LLMs dynamic
Ability to augment LLMs with updated and unavailable information (incl. private data) for generation

2

Offloads factual information from LLM
RAG offloads the factual information to the retrieval system and LLMs only have to curate the response from the contexts provided

Chain-of-Verification (CoVe)

Query
Name some politicians who were born in NY, New York

1 Baseline response

Here are some politicians who were born in NY, New York:

1. Hillary Clinton — former secretary of state and former Democratic presidential nominee
2. Donald Trump — former president of the United States
3. Michael Bloomberg — former Mayor of New York City and former Democratic presidential candidate

…<list continues>

2 Plan verifications

- Where was Hillary Clinton born?
- Where was Donald Trump born?
- Where was Michael Bloomberg born?

…<questions continue>

3 Execute verifications

Hillary Clinton was born in **Chicago, Illinois,** United States on October 26, 1947.

~~Donald Trump was born on June 14, 1946, in~~ **Queens, New York City, New York,** United States.

~~Michael Bloomberg was born on February~~ 14, 1942, in **Boston, Massachusetts,** United States.

4 Final verified response

Here are some politicians who were born in NY, New York:

1. Donald Trump — former president of the United States
2. Alexandria Ocasio-Cortez — Democratic member of the U.S. House of Representatives

…<list continues>

Source: CHAIN-OF-VERIFICATION REDUCES HALLUCINATION IN LARGE LANGUAGE MODELS

Chain-of-Verification (CoVe)

1

Generate baseline response: Given a query, generate the response using the LLM

2

Plan verifications: Given both query and baseline response, generate a list of verification questions that could help to self-analyze if there are any mistakes in the original response

3

Execute verifications: Answer each verification question in turn, and hence check the answer against the original response to check for inconsistencies or mistakes

4

Generate final verified response: Given the discovered inconsistencies (if any), generate a revised response incorporating the verification results

Source: CHAIN-OF-VERIFICATION REDUCES HALLUCINATION IN LARGE LANGUAGE MODELS

Chain-of-Verification (CoVe)

		Wikidata (Easier)			Wiki-Category list (Harder)		
LLM	Method	Prec. (↑)	Pos.	Neg.	Prec. (↑)	Pos.	Neg.
Llama 2 70B Chat	Zero-shot	0.12	0.55	3.93	0.05	0.35	6.85
Llama 2 70B Chat	CoT	0.08	0.75	8.92	0.03	0.30	11.1
Llama 65B	Few-shot	0.17	0.59	2.95	0.12	0.55	4.05
Llama 65B	CoVe (joint)	0.29	0.41	0.98	0.15	0.30	1.69
Llama 65B	CoVe (two-step)	**0.36**	0.38	0.68	0.21	0.50	0.52
Llama 65B	CoVe (factored)	0.32	0.38	0.79	**0.22**	0.52	1.52

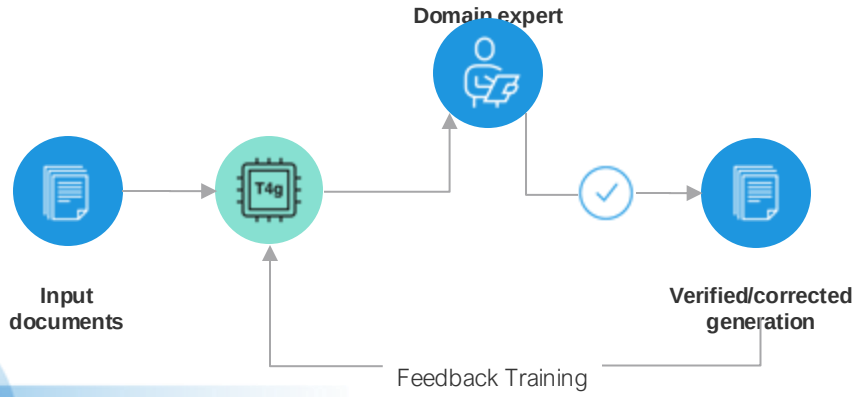
Table 1: Test Precision and average number of positive and negative (hallucination) entities for list-based questions on the Wikidata and Wiki-Category list tasks.

Source: CHAIN-OF-VERIFICATION REDUCES HALLUCINATION IN LARGE LANGUAGE MODELS

Other methods:

1

Human in the loop: Human reviewers add domain expertise to the process, evaluate generated content, and ensure the outputs meet the desired criteria. Eventually this could be used to improve the LLM.



2

Limit the outcomes: Wherever possible, if the outcomes can be limited to a certain fixed outcomes. For example, In a classification problem - the LLM can be prompted only to generate the class label instead of a free form text.

3

Specificity of prompts: Being very specific about the expected generation and what is not expected from the model. Example, We could prompt the model to say “I don’t know”, if the answer is not present in the retrieved context (in a RAG system).

Conclusions

- There are many types of AI hallucinations, as well as multiple techniques to **measure and control hallucinations**
- AI systems must be **designed in a responsible manner**
- AI is not a replacement for humans, rather, it **augments human intelligence**

Want to learn more?

Read our paper on limiting hallucinations

Follow our AI blog

Get a personalized demonstration

saama

Aditya Gadiko

Thank you!