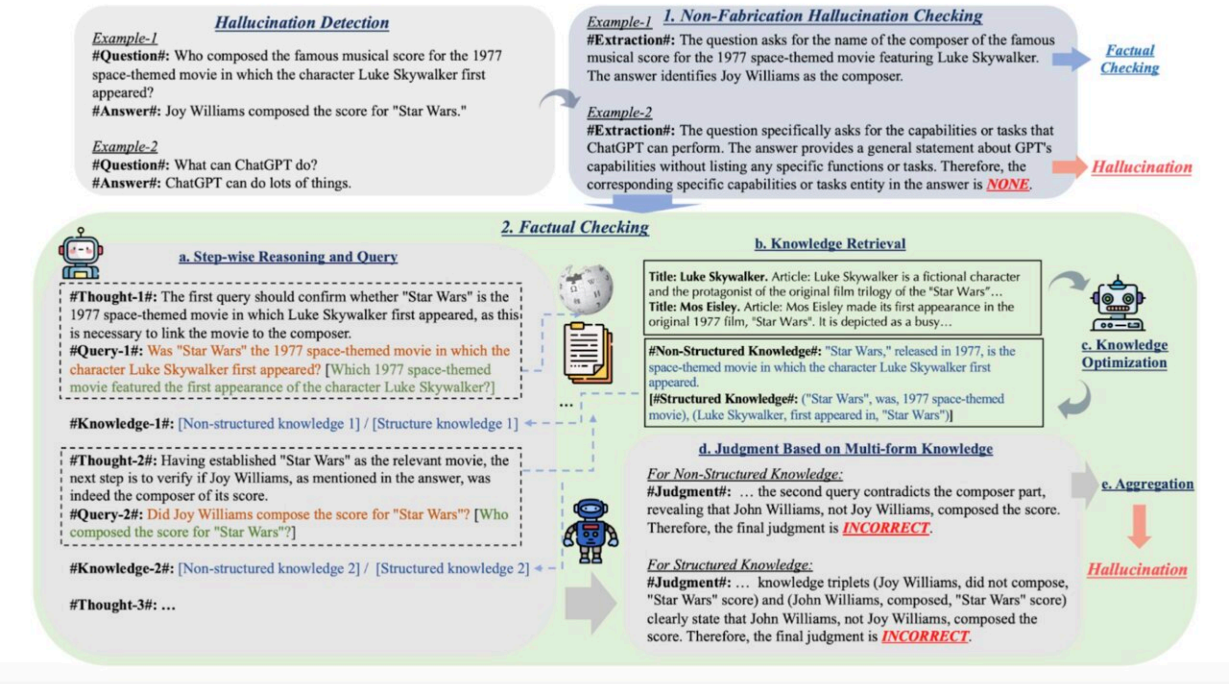
TechXplore

A framework to detect hallucinations in the text generated by LLMs

May 7 2024, by Ingrid Fadelli


Hallucination Detection
Example-1
#Question#: Who composed the famous musical score for the 1977 space-themed movie in which the character Luke Skywalker first appeared?
#Answer#: Joy Williams composed the score for "Star Wars."
Example-2
#Question#: What can ChatGPT do?
#Answer#: ChatGPT can do lots of things.
1. Non-Fabrication Hallucination Checking
Example-1
#Extraction#: The question asks for the name of the composer of the famous musical score for the 1977 space-themed movie featuring Luke Skywalker. The answer identifies Joy Williams as the composer.
Factual Checking
Example-2
#Extraction#: The question specifically asks for the capabilities or tasks that ChatGPT can perform. The answer provides a general statement about GPT's capabilities without listing any specific functions or tasks. Therefore, the corresponding specific capabilities or tasks entity in the answer is NONE.
Hallucination
2. Factual Checking
a. Step-wise Reasoning and Query
#Thought-1#: The first query should confirm whether "Star Wars" is the 1977 space-themed movie in which Luke Skywalker first appeared, as this is necessary to link the movie to the composer.
#Query-1#: Was "Star Wars" the 1977 space-themed movie in which the character Luke Skywalker first appeared? [Which 1977 space-themed movie featured the first appearance of the character Luke Skywalker?]
#Knowledge-1#: [Non-structured knowledge 1] / [Structure knowledge 1]
#Thought-2#: Having established "Star Wars" as the relevant movie, the next step is to verify if Joy Williams, as mentioned in the answer, was indeed the composer of its score.
#Query-2#: Did Joy Williams compose the score for "Star Wars"? [Who composed the score for "Star Wars"?]
#Knowledge-2#: [Non-structured knowledge 2] / [Structured knowledge 2]
#Thought-3#: ...
...
b. Knowledge Retrieval
Title: Luke Skywalker. Article: Luke Skywalker is a fictional character and the protagonist of the original film trilogy of the "Star Wars"...
Title: Mos Eisley. Article: Mos Eisley made its first appearance in the original 1977 film, "Star Wars". It is depicted as a busy...
#Non-Structured Knowledge#: "Star Wars," released in 1977, is the space-themed movie in which the character Luke Skywalker first appeared.
[#Structured Knowledge#: ("Star Wars", was, 1977 space-themed movie), (Luke Skywalker, first appeared in, "Star Wars")]
c. Knowledge Optimization
d. Judgment Based on Multi-form Knowledge
For Non-Structured Knowledge:
#Judgment#: ... the second query contradicts the composer part, revealing that John Williams, not Joy Williams, composed the score. Therefore, the final judgment is INCORRECT.
For Structured Knowledge:
#Judgment#: ... knowledge triplets (Joy Williams, did not compose, "Star Wars" score) and (John Williams, composed, "Star Wars" score) clearly state that John Williams, not Joy Williams, composed the score. Therefore, the final judgment is INCORRECT.
e. Aggregation
Hallucination

Overview of KnowHalu. The hallucination detection process starts with "Non-Fabrication Hallucination Checking," a phase focusing on the early identification of non-fabrication hallucinations by scrutinizing the specificity of the answers. For potential fabrication hallucinations, KnowHalu then provides a comprehensive "Factual Checking," which consists of five steps: (a) "Step-wise Reasoning and Query" breaks down the original query into step-wise reasoning and sub-queries for detailed factual checking; (b) "Knowledge Retrieval" retrieves unstructured knowledge via RAG and structured knowledge in the form of triplets for each sub-query; (c) "Knowledge Optimization" leverages LLMs to summarize and refine the retrieved knowledge into different forms; (d) "Judgment Based on Multi-form Knowledge" employs LLMs to critically assesses the answer to sub-queries, based on each form of knowledge; (e) "Aggregation" provides a further refined judgment by aggregating predictions based on different forms of knowledge Credit: Zhang et al, *arXiv* (2024). DOI: 10.48550/arxiv.2404.02935

Large language models (LLMs) are advanced AI-based dialogue systems that can answer user queries and generate convincing texts following human instructions. After the advent of ChatGPT, the highly performing model developed by OpenAI, these models have become increasingly popular, and more companies are now investing in their development.

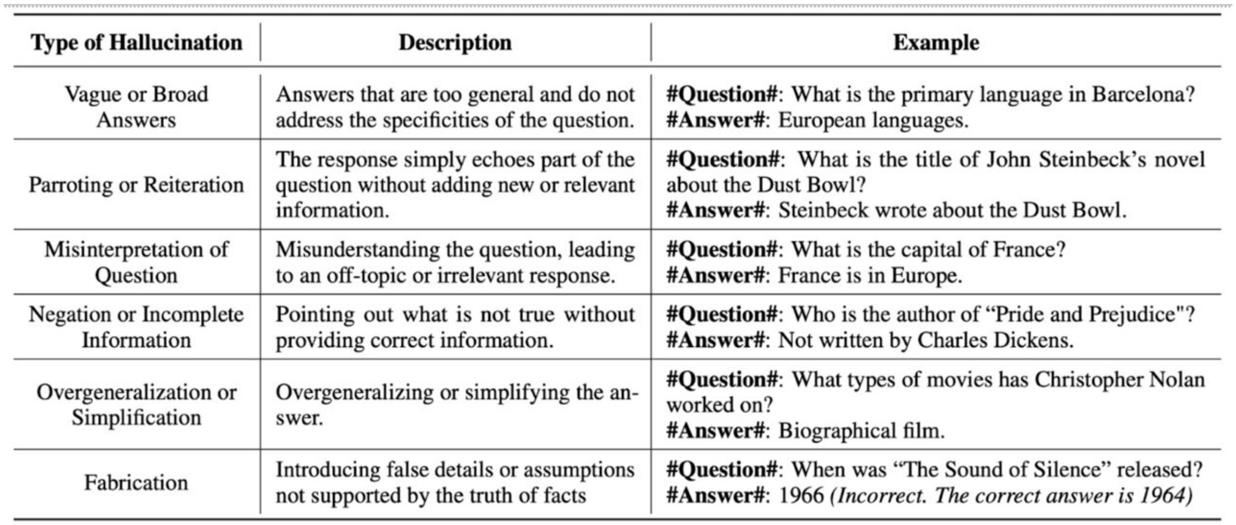
Despite their promise for answering human questions in real-time and creating texts for specific purposes, LLMs can sometimes generate nonsensical, inaccurate or irrelevant texts that diverge from the prompts that were fed to them by human users. This phenomenon, which is often linked to the limitations of the data used to train the models or mistakes in their underlying reasoning, is referred to as LLM "hallucinations."

Researchers at University of Illinois Urbana-Champaign recently introduced KnowHalu, a framework to detect hallucinations in the text generated by LLMs. This framework, introduced in a paper posted to the preprint server *arXiv*, could help to improve the reliability of these models and simplify their use for completing various text generation tasks.

"As advancements in LLMs continue, hallucinations emerge as a critical obstacle impeding their broader real-world application," Bo Li, advisor of the project, told Tech Xplore. "Although numerous studies have addressed LLM hallucinations, existing methods often fail to effectively leverage real-world knowledge or utilize it inefficiently.

"Motivated by this gap, we developed a novel multi-form knowledge-based hallucination detection framework for LLMs. Furthermore, we identified a gap in current research concerning non-fabrication hallucinations: responses that are factually correct but irrelevant or not specific to the query."

When they reviewed past literature, Li and her collaborators found that many past approaches aimed at detecting LLM hallucinations focused on the generation of nonsensical texts, rather than factually accurate texts that are not aligned with user prompts. The new framework they developed thus also features a dedicated component designed to detect these types of accurate but irrelevant hallucinations.



Type of Hallucination	Description	Example
Vague or Broad Answers	Answers that are too general and do not address the specificities of the question.	**#Question#**: What is the primary language in Barcelona? **#Answer#**: European languages.
Parroting or Reiteration	The response simply echoes part of the question without adding new or relevant information.	**#Question#**: What is the title of John Steinbeck's novel about the Dust Bowl? **#Answer#**: Steinbeck wrote about the Dust Bowl.
Misinterpretation of Question	Misunderstanding the question, leading to an off-topic or irrelevant response.	**#Question#**: What is the capital of France? **#Answer#**: France is in Europe.
Negation or Incomplete Information	Pointing out what is not true without providing correct information.	**#Question#**: Who is the author of "Pride and Prejudice"? **#Answer#**: Not written by Charles Dickens.
Overgeneralization or Simplification	Overgeneralizing or simplifying the answer.	**#Question#**: What types of movies has Christopher Nolan worked on? **#Answer#**: Biographical film.
Fabrication	Introducing false details or assumptions not supported by the truth of facts	**#Question#**: When was "The Sound of Silence" released? **#Answer#**: 1966 *(Incorrect. The correct answer is 1964)*

Different types of hallucinations for question-answer (QA) tasks. Credit: Zhang et al, *arXiv* (2024). DOI: 10.48550/arxiv.2404.02935

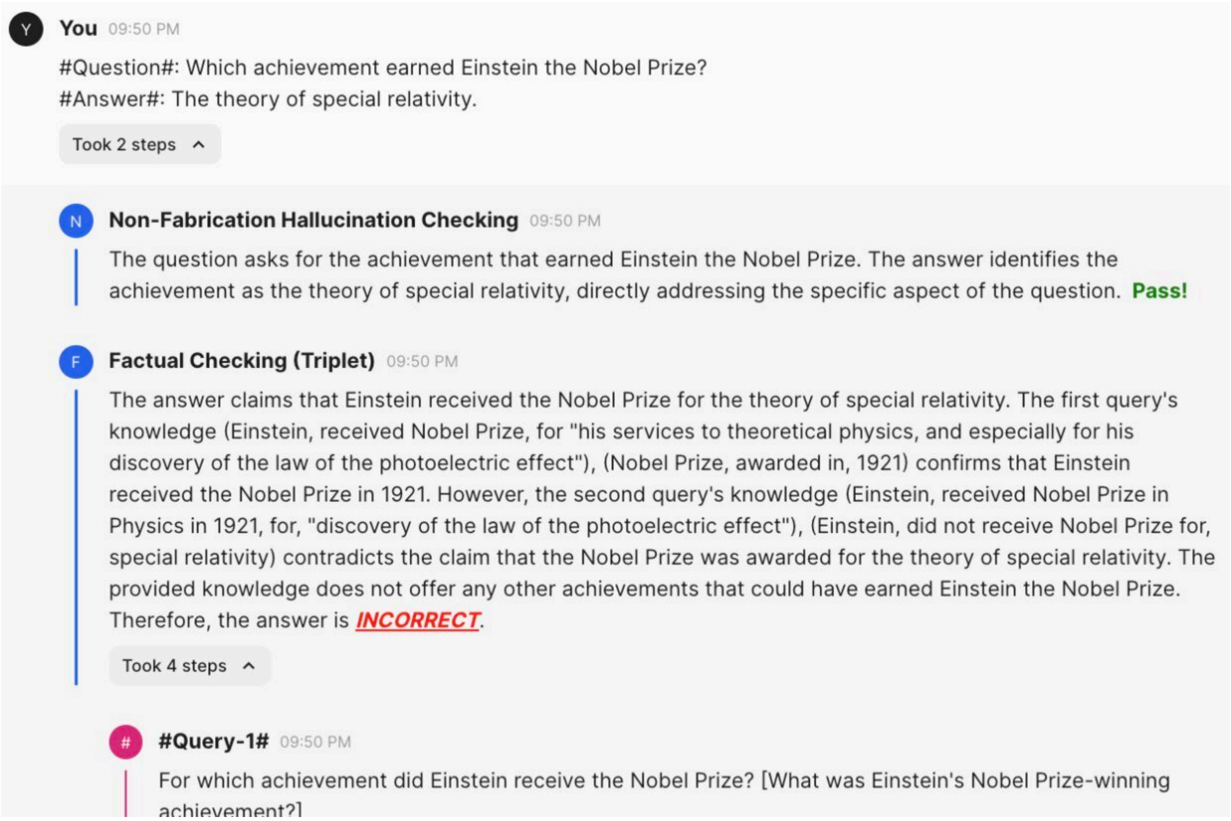
"KnowHalu is a novel framework designed to detect hallucinations in responses generated by LLMs," Li explained. "It operates by using a twophase process that involves multiple components to ensure the accuracy and relevance of LLM outputs. The first phase focuses on detecting non-fabrication hallucinations, which are responses that may be factually correct but are irrelevant or not specific to the query at hand, and such detection is largely missing in current literature."

During the second phase of its operation, KnowHalu employs a multi-form knowledge-based fact checking process that spans across five steps. These steps are: step-wise reasoning and query, knowledge retrieval, knowledge optimization, judgment based on multi-form knowledge and judgment aggregation.

"This comprehensive process helps in identifying ungrounded or irrelevant information provided by LLMs, making KnowHalu particularly effective across different applications, such as QA and summarization tasks," Li said.

KnowHalu has several unique characteristics and advantages over other LLM hallucination detection approaches. Most notably, it can also detect non-fabricated hallucinations, can assess different types of queries, and utilizes a newly developed multi-form knowledge-enabled fact-checking process.

Li and her students tested their framework in a series of tests and found that outperformed various other baseline methods and LLM hallucination detection tools. Using KnowHalu, the researchers also gathered interesting insights about hallucination in LLM models.



You 09:50 PM

#Question#: Which achievement earned Einstein the Nobel Prize?
#Answer#: The theory of special relativity.

Non-Fabrication Hallucination Checking 09:50 PM

The question asks for the achievement that earned Einstein the Nobel Prize. The answer identifies the achievement as the theory of special relativity, directly addressing the specific aspect of the question. **Pass!**

Factual Checking (Triplet) 09:50 PM

The answer claims that Einstein received the Nobel Prize for the theory of special relativity. The first query's knowledge (Einstein, received Nobel Prize, for "his services to theoretical physics, and especially for his discovery of the law of the photoelectric effect"), (Nobel Prize, awarded in, 1921) confirms that Einstein received the Nobel Prize in 1921. However, the second query's knowledge (Einstein, received Nobel Prize in Physics in 1921, for, "discovery of the law of the photoelectric effect"), (Einstein, did not receive Nobel Prize for, special relativity) contradicts the claim that the Nobel Prize was awarded for the theory of special relativity. The provided knowledge does not offer any other achievements that could have earned Einstein the Nobel Prize. Therefore, the answer is ***INCORRECT***.

#Query-1# 09:50 PM

For which achievement did Einstein receive the Nobel Prize? [What was Einstein's Nobel Prize-winning achievement?]

An example of detecting hallucinations with KnowHalu for QA tasks. Credit: Zhang et al.

First, they found that different prompts and different models attain better results on some types of knowledge. For instance, the Starling-7B model excels when given unstructured knowledge, whereas GPT-3.5 is more efficient with structured knowledge.

"Our multi-form knowledge-based RAG significantly outperforms the standard RAG, which is proposed for the first time," Li said. "Moreover, we found that models released later have a higher capability of utilizing structured data, highlighting the importance of our multi-form knowledge algorithm.

"KnowHalu significantly outperforms different SOTA baselines, and even performs much better than directly prompting GPT-4 to perform hallucination detection, which demonstrates its effectiveness and the possibility of hallucination detection and mitigation."

The findings gathered by Li and her collaborators also demonstrate that the formulation of user queries aimed at information retrieval significantly impact the quality of responses produced by LLMs.

Specifically, if users are seeking speculative or vague responses, it would be advisable to formulate general questions, but if they are seeking more specific answers, they should offer more detailed prompts highlighting the type of information they are seeking for using so-called "identifiers." These identifiers are generally also present in the database that models rely on, thus it will be easier for them to retrieve accurate information.

In the future, KnowHalu could inform the development of better performing LLMs that do not hallucinate as often and generate more reliable responses. In addition, the new framework could inspire other research teams to devise approaches that tackle a wider range of LLM hallucinations.

"We now plan to further automatically parse different documents and extract knowledge to help mitigate hallucinations for LLMs and explore diverse forms of knowledge and map the retrieved knowledge to other forms such as higher-order logic forms to help ground the model generation," Li added.

"Moreover, we will try to provide theoretical guarantees for LLM hallucination based on given knowledge bases and adapt our framework to diverse application domains such as autonomous driving agents and health care agents."

More information: Jiawei Zhang et al, KnowHalu: Hallucination Detection via Multi-Form Knowledge Based Factual Checking, *arXiv* (2024). [DOI: 10.48550/arxiv.2404.02935](https://dx.doi.org/10.48550/arxiv.2404.02935)

Citation: A framework to detect hallucinations in the text generated by LLMs (2024, May 7) retrieved 29 June 2024 from https://techxplore.com/news/2024-05-framework-hallucinations-text-generated-llms.html